Statystyczne metody uczenia maszynowego – redukcja wymiarowości i klasyfikacja

Przemysław Głomb

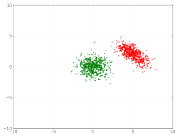
Instytut Informatyki Teoretycznej i Stosowanej Polskiej Akademii Nauk

redaktor pomocniczy: Bartłomiej Gardas

Metody liniowe

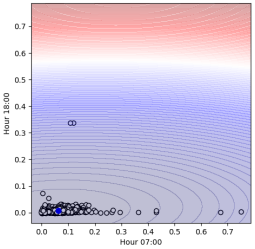
Analiza składowych głównych

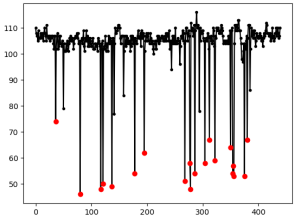
Macierz kowariancji, odległość Mahalanobisa, RX, ...



Parametry rozkładu normalnego $\mathcal{N}(\mu, \Sigma)$

$$\mu = \begin{bmatrix} 5 \\ 2 \end{bmatrix} \qquad \Sigma = \begin{bmatrix} 1 & -0.69 \\ -0.69 & 1 \end{bmatrix}$$





Zbiory danych są wielowymiarowe

1 cecha	przykład z wykładu o perceptronie - dojrzałość owocu
2 cechy	właściwości przeciwrakowe peptydów (sekwencja dna, aktywność)
5 cech	predykcja awarii silników wentylatora na podstawie danych z akcelerometrów
10 cech	predykcja właściwości rozbłysków słonecznych
24 cechy	klasyfikacja choroby nerek na podstawie parametrów biochemicznych
127 cech	predykcja parametrów przestępstw i wykroczeń na podstawie danych socjoekonomicznych
1203 cechy	Toksyczność cząsteczek mających wpływ na zegar biologiczny
65 tys. cech	obrazek 256×256
150 tys. cech	Zespół czujników gazowych wystawiony na działanie turbulentnych mieszanek gazowych
3+ mln cech	reputacja adresów URL

Zbiory danych są wielowymiarowe

1 cecha	przykład z wykładu o perceptronie - dojrzałość owocu
2 cechy	właściwości przeciwrakowe peptydów (sekwencja dna, aktywność)[1]
5 cech	predykcja awarii silników wentylatora na podstawie danych z akcelerometrów
10 cech	predykcja właściwości rozbłysków słonecznych
24 cechy	klasyfikacja choroby nerek na podstawie parametrów biochemicznych
127 cech	predykcja parametrów przestępstw i wykroczeń na podstawie danych socjoekonomicznych
1203 cechy	Toksyczność cząsteczek mających wpływ na zegar biologiczny
65 tys. cech	obrazek 256×256
150 tys. cech	Zespół czujników gazowych wystawiony na działanie turbulentnych mieszanek gazowych
3+ mln cech	reputacja adresów URL

[1] Grisoni, F., Neuhaus, C.S., Hishinuma, M., Gabernet, G., Hiss, J.A., Kotera, M. and Schneider, G., 2019. De novo design of anticancer peptides by ensemble artificial neural networks. Journal of Molecular Modeling, 25(5), 112

Zbiory danych są wielowymiarowe

1 cecha	przykład z wykładu o perceptronie - dojrzałość owocu
2 cechy	właściwości przeciwrakowe peptydów (sekwencja dna, aktywność)
5 cech	predykcja awarii silników wentylatora na podstawie danych z akcelerometrów[1]
10 cech	predykcja właściwości rozbłysków słonecznych
24 cechy	klasyfikacja choroby nerek na podstawie parametrów biochemicznych
127 cech	predykcja parametrów przestępstw i wykroczeń na podstawie danych socjoekonomicznych
1203 cechy	Toksyczność cząsteczek mających wpływ na zegar biologiczny
65 tys. cech	obrazek 256×256
150 tys. cech	Zespół czujników gazowych wystawiony na działanie turbulentnych mieszanek gazowych
3+ mln cech	reputacja adresów URL

[1] Scalabrini Sampaio, G., Vallim Filho, A. R. d. A., Santos da Silva, L., & Augusto da Silva, L. (2019). Prediction of Motor Failure Time Using An Artificial Neural Network. Sensors, 19(19), 4342

Zbiory danych są wielowymiarowe

1 cecha	przykład z wykładu o perceptronie - dojrzałość owocu
2 cechy	właściwości przeciwrakowe peptydów (sekwencja dna, aktywność)
5 cech	predykcja awarii silników wentylatora na podstawie danych z akcelerometrów
10 cech	predykcja właściwości rozbłysków słonecznych[1]
24 cechy	klasyfikacja choroby nerek na podstawie parametrów biochemicznych
127 cech	predykcja parametrów przestępstw i wykroczeń na podstawie danych socjoekonomicznych
1203 cechy	Toksyczność cząsteczek mających wpływ na zegar biologiczny
65 tys. cech	obrazek 256×256
150 tys. cech	Zespół czujników gazowych wystawiony na działanie turbulentnych mieszanek gazowych
3+ mln cech	reputacja adresów URL

[1] https://archive.ics.uci.edu/dataset/89/solar+flare

Zbiory danych są wielowymiarowe

1 cecha	przykład z wykładu o perceptronie - dojrzałość owocu
2 cechy	właściwości przeciwrakowe peptydów (sekwencja dna, aktywność)
5 cech	predykcja awarii silników wentylatora na podstawie danych z akcelerometrów
10 cech	predykcja właściwości rozbłysków słonecznych
24 cechy	klasyfikacja choroby nerek na podstawie parametrów biochemicznych[1]
127 cech	predykcja parametrów przestępstw i wykroczeń na podstawie danych socjoekonomicznych
1203 cechy	Toksyczność cząsteczek mających wpływ na zegar biologiczny
65 tys. cech	obrazek 256×256
150 tys. cech	Zespół czujników gazowych wystawiony na działanie turbulentnych mieszanek gazowych
3+ mln cech	reputacja adresów URL

[1] https://archive.ics.uci.edu/dataset/336/chronic+kidney+disease

Zbiory danych są wielowymiarowe

1 cecha	przykład z wykładu o perceptronie - dojrzałość owocu
2 cechy	właściwości przeciwrakowe peptydów (sekwencja dna, aktywność)
5 cech	predykcja awarii silników wentylatora na podstawie danych z akcelerometrów
10 cech	predykcja właściwości rozbłysków słonecznych
24 cechy	klasyfikacja choroby nerek na podstawie parametrów biochemicznych
127 cech	predykcja parametrów przestępstw i wykroczeń na podstawie danych socjoekonomicznych[1]
1203 cechy	Toksyczność cząsteczek mających wpływ na zegar biologiczny
65 tys. cech	obrazek 256×256
150 tys. cech	Zespół czujników gazowych wystawiony na działanie turbulentnych mieszanek gazowych
3+ mln cech	reputacja adresów URL

[1] Redmond, Michael and Alok Baveja. "A data-driven software tool for enabling cooperative information sharing among police departments." Eur. J. Oper. Res. 141 (2002): 660-678.

Zbiory danych są wielowymiarowe

1 cecha	przykład z wykładu o perceptronie - dojrzałość owocu
2 cechy	właściwości przeciwrakowe peptydów (sekwencja dna, aktywność)
5 cech	predykcja awarii silników wentylatora na podstawie danych z akcelerometrów
10 cech	predykcja właściwości rozbłysków słonecznych
24 cechy	klasyfikacja choroby nerek na podstawie parametrów biochemicznych
127 cech	predykcja parametrów przestępstw i wykroczeń na podstawie danych socjoekonomicznych
1203 cechy	Toksyczność cząsteczek mających wpływ na zegar biologiczny[1]
65 tys. cech	obrazek 256×256
150 tys. cech	Zespół czujników gazowych wystawiony na działanie turbulentnych mieszanek gazowych
3+ mln cech	reputacja adresów URL

[1] Gul, S., Rahim, F., Isin, S. et al. Structure-based design and classifications of small molecules regulating the circadian rhythm period. Sci Rep 11, 18510 (2021)

Zbiory danych są wielowymiarowe

1 cecha	przykład z wykładu o perceptronie - dojrzałość owocu
2 cechy	właściwości przeciwrakowe peptydów (sekwencja dna, aktywność)
5 cech	predykcja awarii silników wentylatora na podstawie danych z akcelerometrów
10 cech	predykcja właściwości rozbłysków słonecznych
24 cechy	klasyfikacja choroby nerek na podstawie parametrów biochemicznych
127 cech	predykcja parametrów przestępstw i wykroczeń na podstawie danych socjoekonomicznych
1203 cechy	Toksyczność cząsteczek mających wpływ na zegar biologiczny
65 tys. cech	obrazek 256×256
150 tys. cech	Zespół czujników gazowych wystawiony na działanie turbulentnych mieszanek gazowych
3+ mln cech	reputacja adresów URL

Zbiory danych są wielowymiarowe

1 cecha	przykład z wykładu o perceptronie - dojrzałość owocu
2 cechy	właściwości przeciwrakowe peptydów (sekwencja dna, aktywność)
5 cech	predykcja awarii silników wentylatora na podstawie danych z akcelerometrów
10 cech	predykcja właściwości rozbłysków słonecznych
24 cechy	klasyfikacja choroby nerek na podstawie parametrów biochemicznych
127 cech	predykcja parametrów przestępstw i wykroczeń na podstawie danych socjoekonomicznych
1203 cechy	Toksyczność cząsteczek mających wpływ na zegar biologiczny
65 tys. cech	obrazek 256×256
150 tys. cech	Zespół czujników gazowych wystawiony na działanie turbulentnych mieszanek gazowych[1]
3+ mln cech	reputacja adresów URL

[1] https://archive.ics.uci.edu/dataset/309/gas+sensor+array+exposed+to+turbulent+gas+mixtures

Zbiory danych są wielowymiarowe

1 cecha	przykład z wykładu o perceptronie - dojrzałość owocu
2 cechy	właściwości przeciwrakowe peptydów (sekwencja dna, aktywność)
5 cech	predykcja awarii silników wentylatora na podstawie danych z akcelerometrów
10 cech	predykcja właściwości rozbłysków słonecznych
24 cechy	klasyfikacja choroby nerek na podstawie parametrów biochemicznych
127 cech	predykcja parametrów przestępstw i wykroczeń na podstawie danych socjoekonomicznych
1203 cechy	Toksyczność cząsteczek mających wpływ na zegar biologiczny
65 tys. cech	obrazek 256×256
150 tys. cech	Zespół czujników gazowych wystawiony na działanie turbulentnych mieszanek gazowych
3+ mln cech	reputacja adresów URL[1]

[1] `https://archive.ics.uci.edu/dataset/187/url+reputation`

Zbiory danych są wielowymiarowe

1 cecha	przykład z wykładu o perceptronie - dojrzałość owocu
2 cechy	właściwości przeciwrakowe peptydów (sekwencja dna, aktywność)
5 cech	predykcja awarii silników wentylatora na podstawie danych z akcelerometrów
10 cech	predykcja właściwości rozbłysków słonecznych
24 cechy	klasyfikacja choroby nerek na podstawie parametrów biochemicznych
127 cech	predykcja parametrów przestępstw i wykroczeń na podstawie danych socjoekonomicznych
1203 cechy	Toksyczność cząsteczek mających wpływ na zegar biologiczny
65 tys. cech	obrazek 256×256
150 tys. cech	Zespół czujników gazowych wystawiony na działanie turbulentnych mieszanek gazowych
3+ mln cech	reputacja adresów URL

1203 cechy? 150 tys. cech? Miliony cech?

- Trudność inspekcji (tysiące → miliony przykładów)
- Złożoność – wymagana „pojemność" modelu (zdolność do reprezentacji)
  - Trudność doboru modelu, architektury
  - Czas trwania przygotowania modelu
  - Zasoby pamięciowe i obliczeniowe na przygotowanie i testowanie modelu
- Właściwości przestrzeni wielowymiarowych
  - Przekleństwo wymiarowości (**curse of dimensionality**)

1203 cechy? 150 tys. cech? Miliony cech?

- Trudność inspekcji (tysiące → miliony przykładów)
- Złożoność – wymagana „pojemność" modelu (zdolność do reprezentacji)
  - Trudność doboru modelu, architektury
  - Czas trwania przygotowania modelu
  - Zasoby pamięciowe i obliczeniowe na przygotowanie i testowanie modelu
- Właściwości przestrzeni wielowymiarowych
  - Przekleństwo wymiarowości (**curse of dimensionality**)

1203 cechy? 150 tys. cech? Miliony cech?

- Trudność inspekcji (tysiące → miliony przykładów)
- Złożoność – wymagana „pojemność" modelu (zdolność do reprezentacji)
  - Trudność doboru modelu, architektury
  - Czas trwania przygotowania modelu
  - Zasoby pamięciowe i obliczeniowe na przygotowanie i testowanie modelu
- Właściwości przestrzeni wielowymiarowych
  - Przekleństwo wymiarowości (**curse of dimensionality**)

1203 cechy? 150 tys. cech? Miliony cech?

- Trudność inspekcji (tysiące → miliony przykładów)
- Złożoność – wymagana „pojemność" modelu (zdolność do reprezentacji)
  - Trudność doboru modelu, architektury
  - Czas trwania przygotowania modelu
  - Zasoby pamięciowe i obliczeniowe na przygotowanie i testowanie modelu
- Właściwości przestrzeni wielowymiarowych
  - Przekleństwo wymiarowości (**curse of dimensionality**)

1203 cechy? 150 tys. cech? Miliony cech?

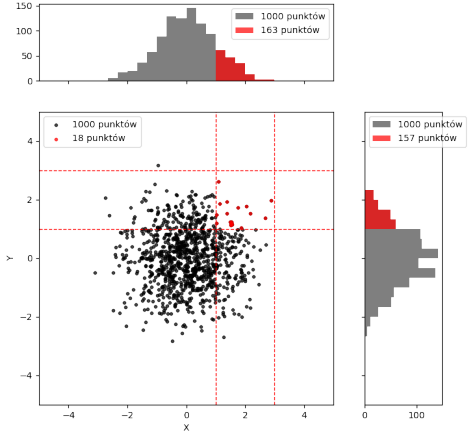
- Trudność inspekcji (tysiące → miliony przykładów)
- Złożoność – wymagana „pojemność" modelu (zdolność do reprezentacji)
  - Trudność doboru modelu, architektury
  - Czas trwania przygotowania modelu
  - Zasoby pamięciowe i obliczeniowe na przygotowanie i testowanie modelu
- Właściwości przestrzeni wielowymiarowych
  - Przekleństwo wymiarowości (**curse of dimensionality**)

Przekleństwo wymiarowości

- Rzadkość danych w przestrzeniach wielowymiarowych
- „Ucentrowienie" (hubness), np. przy szukaniu podobnych przykładów
- Wymagania liczby punktów danych $n > d, n \geqslant 10d, n \gg d$
- Wymagania obliczeniowe i pamięciowe
- Trudność oszacowania parametrów statystycznych (korelacje, outliery)

Przekleństwo wymiarowości

- Rzadkość danych w przestrzeniach wielowymiarowych



- „Ucentrowienie” (hubness), np. przy szukaniu podobnych przykładów

Przekleństwo wymiarowości

- Rzadkość danych w przestrzeniach wielowymiarowych
- „Ucentrowienie" (hubness), np. przy szukaniu podobnych przykładów
- Wymagania liczby punktów danych $n > d, n \geqslant 10d, n \gg d$
- Wymagania obliczeniowe i pamięciowe
- Trudność oszacowania parametrów statystycznych (korelacje, outliery)

Korelacje między cechami

Intuicja:

- Dane z urządzeń IoT mierzących temperaturę, wilgotność itp. w mieście
- Film z kamery monitoringu parkingu
- Dane mieszkań: cena, rozmiar, liczba pokojów, liczba łazienek, miejsce dodatkowe (piwnica/garaż)
- Zużycie wody w okresie godzinowym w ciągu tygodnia

Reprezentacja korelacji:

- Nazwane cechy (np. cena mieszkania i rozmiar w m^2)
- Nie nazwane cechy (np. piksele obrazka)

Korelacje między cechami

Intuicja:

- Dane z urządzeń IoT mierzących temperaturę, wilgotność itp. w mieście
- Film z kamery monitoringu parkingu
- Dane mieszkań: cena, rozmiar, liczba pokojów, liczba łazienek, miejsce dodatkowe (piwnica/garaż)
- Zużycie wody w okresie godzinowym w ciągu tygodnia

Reprezentacja korelacji:

- Nazwane cechy (np. cena mieszkania i rozmiar w m^2)
- Nie nazwane cechy (np. piksele obrazka)

Korelacje między cechami

Intuicja:

- Dane z urządzeń IoT mierzących temperaturę, wilgotność itp. w mieście
- Film z kamery monitoringu parkingu
- Dane mieszkań: cena, rozmiar, liczba pokojów, liczba łazienek, miejsce dodatkowe (piwnica/garaż)
- Zużycie wody w okresie godzinowym w ciągu tygodnia

Reprezentacja korelacji:

- Nazwane cechy (np. cena mieszkania i rozmiar w m^2)
- Nie nazwane cechy (np. piksele obrazka)

Korelacje między cechami

Intuicja:

- Dane z urządzeń IoT mierzących temperaturę, wilgotność itp. w mieście
- Film z kamery monitoringu parkingu
- Dane mieszkań: cena, rozmiar, liczba pokojów, liczba łazienek, miejsce dodatkowe (piwnica/garaż)
- Zużycie wody w okresie godzinowym w ciągu tygodnia

Reprezentacja korelacji:

- Nazwane cechy (np. cena mieszkania i rozmiar w m^2)
- Nie nazwane cechy (np. piksele obrazka)

Korelacje między cechami

Intuicja:

- Dane z urządzeń IoT mierzących temperaturę, wilgotność itp. w mieście
- Film z kamery monitoringu parkingu
- Dane mieszkań: cena, rozmiar, liczba pokojów, liczba łazienek, miejsce dodatkowe (piwnica/garaż)
- Zużycie wody w okresie godzinowym w ciągu tygodnia

Reprezentacja korelacji:

- Nazwane cechy (np. cena mieszkania i rozmiar w m^2)
- Nie nazwane cechy (np. piksele obrazka)

Korelacje między cechami

Intuicja:

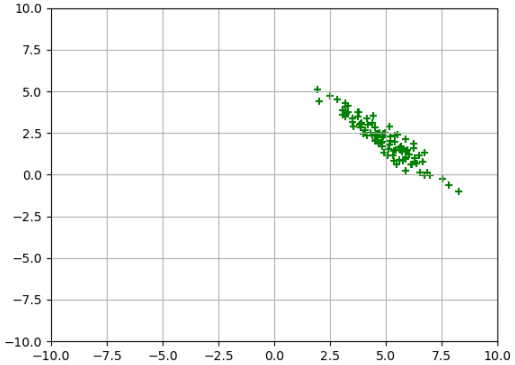
- Dane z urządzeń IoT mierzących temperaturę, wilgotność itp. w mieście
- Film z kamery monitoringu parkingu
- Dane mieszkań: cena, rozmiar, liczba pokojów, liczba łazienek, miejsce dodatkowe (piwnica/garaż)
- Zużycie wody w okresie godzinowym w ciągu tygodnia

Reprezentacja korelacji:

- Nazwane cechy (np. cena mieszkania i rozmiar w m^2)
- Nie nazwane cechy (np. piksele obrazka)

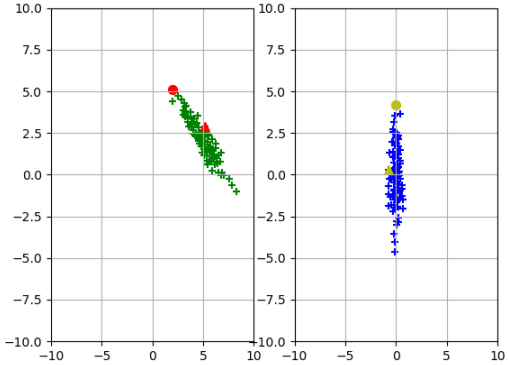
Dekorelacja

Intuicja:



Dekorelacja

Intuicja:



Analiza składowych głównych – wyznaczanie

1. Zbiór przykładów $\mathcal{X}$
   $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m\}$
   $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{in})$, $x_{ij} \in \mathbb{R}$
2. Zbiór po odjęciu średniej
   $\mathbf{x}'_i = \mathbf{x}_i - \frac{1}{m}\sum_{i=1}^{m} \mathbf{x}_i = \mathbf{x}_i - \bar{\mathbf{x}}$
   zapisujemy w postaci macierzy
   $$X = \begin{bmatrix} x'_{11} & x'_{12} & \ldots \\ x'_{21} & x'_{22} & \ldots \\ \ldots & \ldots & x'_{mn} \end{bmatrix}$$
3. Wyznaczamy macierz kowariancji
   $S = \frac{1}{m-1} X X^\top$
   i jej wektory własne (posortowane po wartościach własnych) układamy w macierz A
4. Dla dowolnego (w ramach rodziny danych) wektora x uzyskujemy składowe z równania
   $\mathbf{c} = A(\mathbf{x} - \bar{\mathbf{x}})$

Analiza składowych głównych – wyznaczanie

1. Zbiór przykładów $\mathcal{X}$
   $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m\}$
   $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{in})$, $x_{ij} \in \mathbb{R}$
2. Zbiór po odjęciu średniej
   $\mathbf{x}'_i = \mathbf{x}_i - \frac{1}{m}\sum_{i=1}^{m} \mathbf{x}_i = \mathbf{x}_i - \bar{\mathbf{x}}$
   zapisujemy w postaci macierzy
   $$X = \begin{bmatrix} x'_{11} & x'_{12} & \ldots \\ x'_{21} & x'_{22} & \ldots \\ \ldots & \ldots & x'_{mn} \end{bmatrix}$$
3. Wyznaczamy macierz kowariancji
   $S = \frac{1}{m-1} X X^\top$
   i jej wektory własne (posortowane po wartościach własnych) układamy w macierz A
4. Dla dowolnego (w ramach rodziny danych) wektora x uzyskujemy składowe z równania
   $\mathbf{c} = A(\mathbf{x} - \bar{\mathbf{x}})$

Analiza składowych głównych – wyznaczanie

1. Zbiór przykładów $\mathcal{X}$
   $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m\}$
   $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{in})$, $x_{ij} \in \mathbb{R}$
2. Zbiór po odjęciu średniej
   $\mathbf{x}'_i = \mathbf{x}_i - \frac{1}{m}\sum_{i=1}^{m} \mathbf{x}_i = \mathbf{x}_i - \bar{\mathbf{x}}$
   zapisujemy w postaci macierzy
   $$X = \begin{bmatrix} x'_{11} & x'_{12} & \ldots \\ x'_{21} & x'_{22} & \ldots \\ \ldots & \ldots & x'_{mn} \end{bmatrix}$$
3. Wyznaczamy macierz kowariancji
   $S = \frac{1}{m-1} X X^\top$
   i jej wektory własne (posortowane po wartościach własnych) układamy w macierz A
4. Dla dowolnego (w ramach rodziny danych) wektora x uzyskujemy składowe z równania
   $\mathbf{c} = A(\mathbf{x} - \bar{\mathbf{x}})$

Analiza składowych głównych – wyznaczanie

1. Zbiór przykładów $\mathcal{X}$
   $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m\}$
   $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{in})$, $x_{ij} \in \mathbb{R}$
2. Zbiór po odjęciu średniej
   $\mathbf{x}'_i = \mathbf{x}_i - \frac{1}{m}\sum_{i=1}^{m} \mathbf{x}_i = \mathbf{x}_i - \bar{\mathbf{x}}$
   zapisujemy w postaci macierzy
   $$X = \begin{bmatrix} x'_{11} & x'_{12} & \ldots \\ x'_{21} & x'_{22} & \ldots \\ \ldots & \ldots & x'_{mn} \end{bmatrix}$$
3. Wyznaczamy macierz kowariancji
   $S = \frac{1}{m-1} X X^\top$
   i jej wektory własne (posortowane po wartościach własnych) układamy w macierz A
4. Dla dowolnego (w ramach rodziny danych) wektora x uzyskujemy składowe z równania
   $\mathbf{c} = A(\mathbf{x} - \bar{\mathbf{x}})$

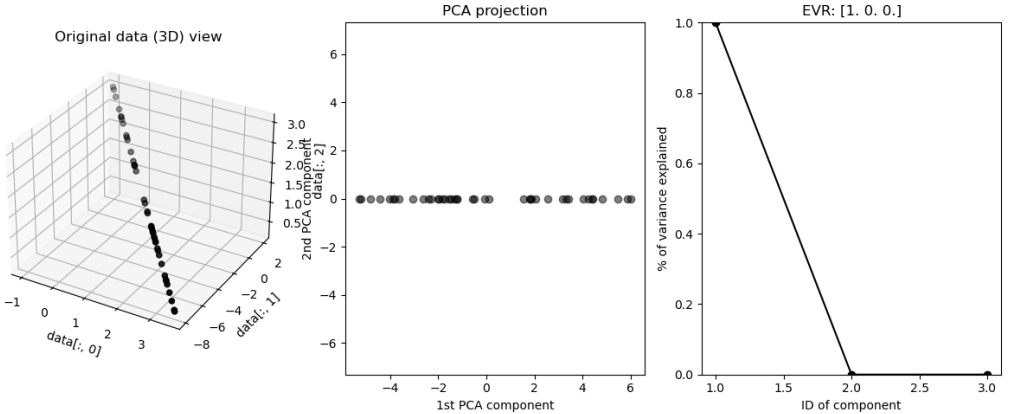
Analiza składowych głównych – wyznaczanie

1. Zbiór przykładów $\mathcal{X}$
   $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m\}$
   $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{in})$, $x_{ij} \in \mathbb{R}$
2. Zbiór po odjęciu średniej
   $\mathbf{x}'_i = \mathbf{x}_i - \frac{1}{m}\sum_{i=1}^{m} \mathbf{x}_i = \mathbf{x}_i - \bar{\mathbf{x}}$
   zapisujemy w postaci macierzy
   $$X = \begin{bmatrix} x'_{11} & x'_{12} & \ldots \\ x'_{21} & x'_{22} & \ldots \\ \ldots & \ldots & x'_{mn} \end{bmatrix}$$
3. Wyznaczamy macierz kowariancji
   $S = \frac{1}{m-1} X X^\top$
   i jej wektory własne (posortowane po wartościach własnych) układamy w macierz A
4. Dla dowolnego (w ramach rodziny danych) wektora x uzyskujemy składowe z równania
   $\mathbf{c} = A(\mathbf{x} - \bar{\mathbf{x}})$

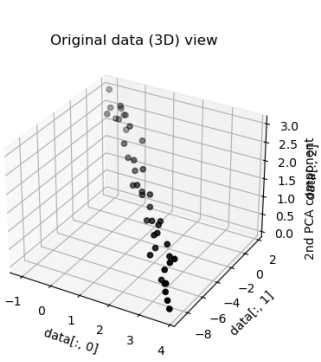
Analiza składowych głównych – wyznaczanie

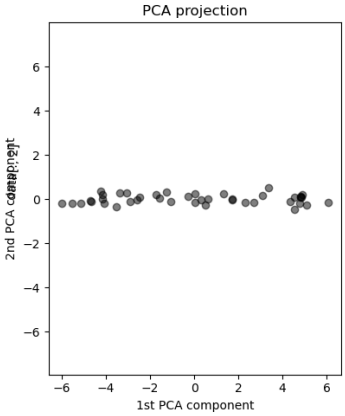
1. Zbiór przykładów $\mathcal{X}$
   $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m\}$
   $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{in})$, $x_{ij} \in \mathbb{R}$
2. Zbiór po odjęciu średniej
   $\mathbf{x}'_i = \mathbf{x}_i - \frac{1}{m}\sum_{i=1}^{m} \mathbf{x}_i = \mathbf{x}_i - \bar{\mathbf{x}}$
   zapisujemy w postaci macierzy
   $$X = \begin{bmatrix} x'_{11} & x'_{12} & \ldots \\ x'_{21} & x'_{22} & \ldots \\ \ldots & \ldots & x'_{mn} \end{bmatrix}$$
3. Wyznaczamy macierz kowariancji
   $S = \frac{1}{m-1} X X^\top$
   i jej wektory własne (posortowane po wartościach własnych) układamy w macierz A
4. Dla dowolnego (w ramach rodziny danych) wektora x uzyskujemy składowe z równania
   $\mathbf{c} = A(\mathbf{x} - \bar{\mathbf{x}})$

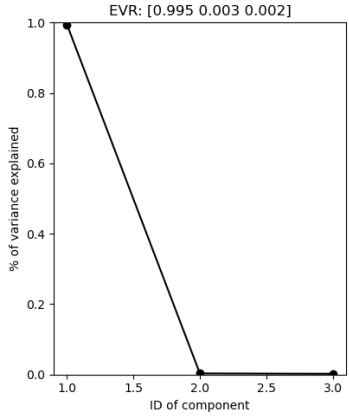
Analiza składowych głównych – przykład/idea działania



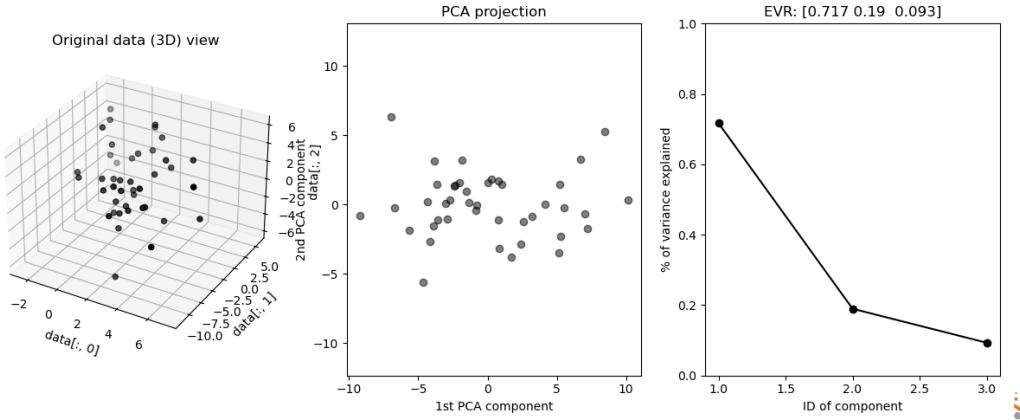
Analiza składowych głównych – przykład/idea działania



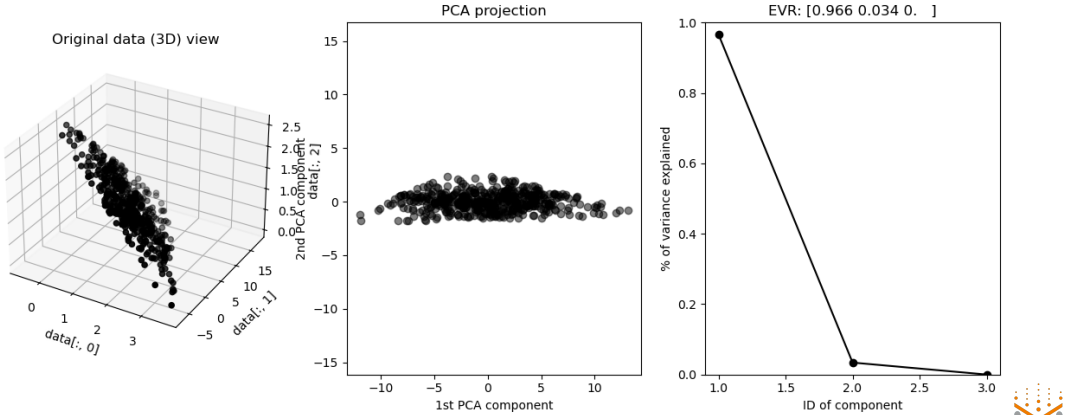




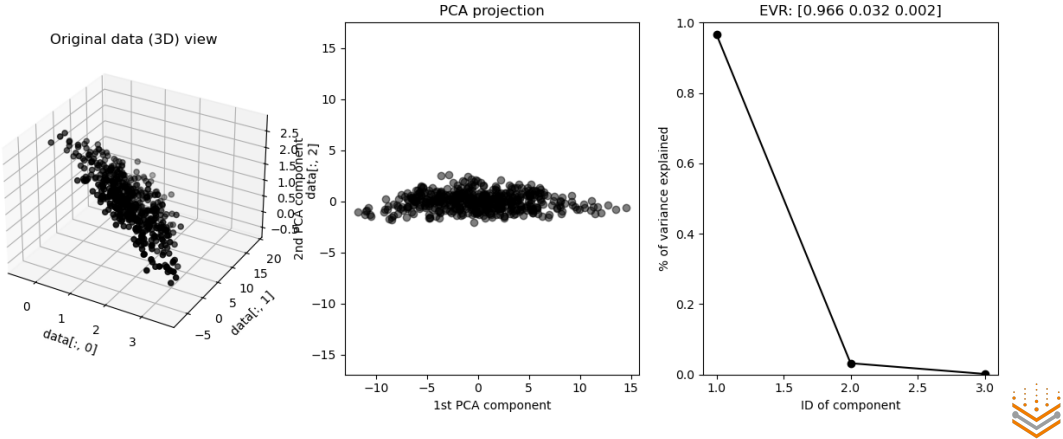
Analiza składowych głównych – przykład/idea działania



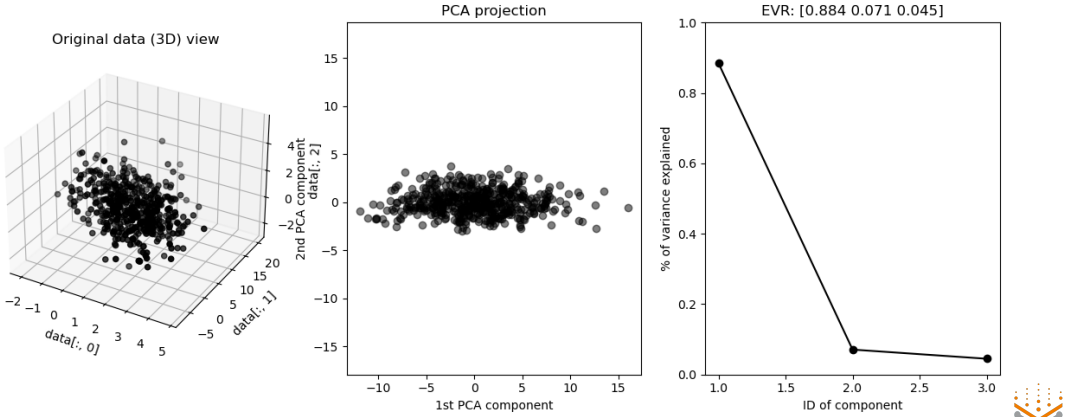
Analiza składowych głównych – przykład/idea działania



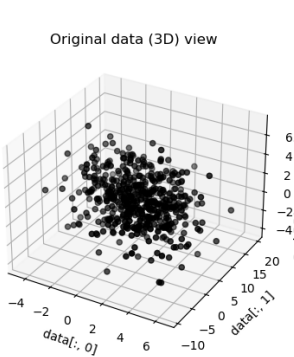
Analiza składowych głównych – przykład/idea działania

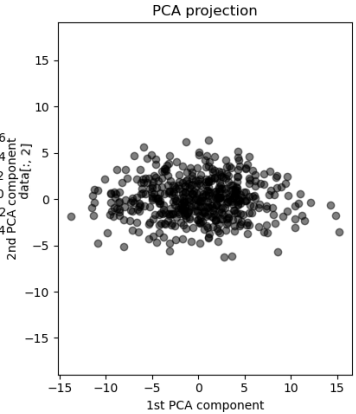


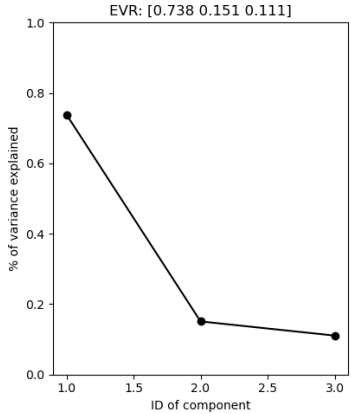
Analiza składowych głównych – przykład/idea działania



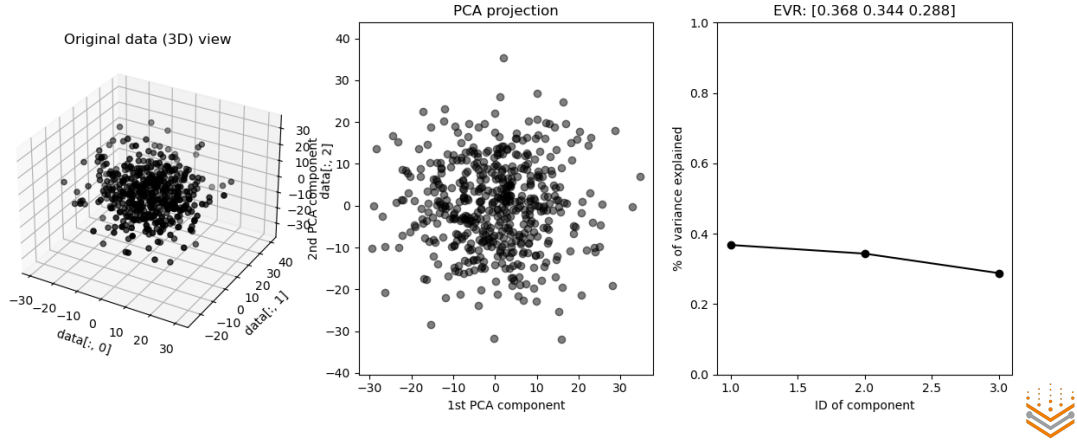
Analiza składowych głównych – przykład/idea działania



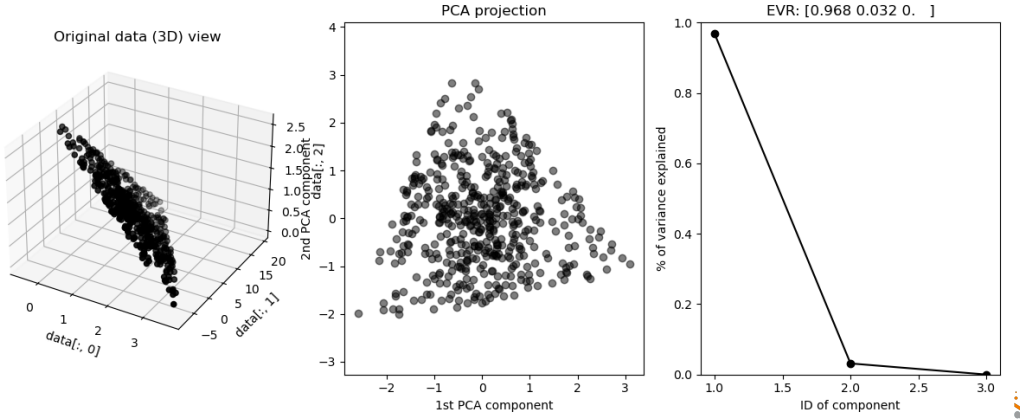




Analiza składowych głównych – przykład/idea działania



Analiza składowych głównych – przykład (wybielenie – whitening)



Analiza składowych głównych – przykład zastosowania (reprezentacja)

- x to zlinearyzowany obraz ($x \in \mathbb{R}^{4096}$):

$\mathbf{x}_1$	$\mathbf{x}_2$
$\mathbf{x}_3$	$\mathbf{x}_4$

- Składowe (wektory z macierzy A) tzw. „eigenfaces”

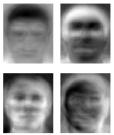
. . . cechy twarzy

Analiza składowych głównych – przykład zastosowania (reprezentacja)

- x to zlinearyzowany obraz ($x \in \mathbb{R}^{4096}$):

$\mathbf{x}_1$	$\mathbf{x}_2$
$\mathbf{x}_3$	$\mathbf{x}_4$

- Składowe (wektory z macierzy A) tzw. „eigenfaces”



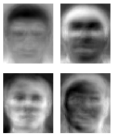
. . . cechy twarzy

Analiza składowych głównych – przykład zastosowania (reprezentacja)

- x to zlinearyzowany obraz ($x \in \mathbb{R}^{4096}$):

$\mathbf{x}_1$	$\mathbf{x}_2$
$\mathbf{x}_3$	$\mathbf{x}_4$

- Składowe (wektory z macierzy A) tzw. „eigenfaces”



. . . cechy twarzy

Analiza składowych głównych – przykład zastosowania (reprezentacja)

- $\mathbf{x} \in \mathbb{R}^{4096} \rightarrow \mathbf{c} \in \mathbb{R}^{4096}$?
  - Macierz kowariancji $S = \frac{1}{m-1} X X^\top$
  - Wektory własne ($\rightarrow A$) oraz wartości własne (wariancja, posortowane)
  - Można wykorzystać pierwszych n wektorów – $\mathbf{x} \in \mathbb{R}^{4096} \rightarrow \mathbf{c} \in \mathbb{R}^{n}$; np. $n = 5, 10, 50$
- Rekonstrukcja z 1, 9, 17, ... (+8 składowych):

Analiza składowych głównych – przykład zastosowania (reprezentacja)

- $\mathbf{x} \in \mathbb{R}^{4096} \rightarrow \mathbf{c} \in \mathbb{R}^{4096}$?
  - Macierz kowariancji $S = \frac{1}{m-1} X X^\top$
  - Wektory własne ($\rightarrow A$) oraz wartości własne (wariancja, posortowane)
  - Można wykorzystać pierwszych n wektorów – $\mathbf{x} \in \mathbb{R}^{4096} \rightarrow \mathbf{c} \in \mathbb{R}^n$; np. $n = 5, 10, 50$
- Rekonstrukcja z 1, 9, 17, ... (+8 składowych):

Analiza składowych głównych – przykład zastosowania (reprezentacja)

- $\mathbf{x} \in \mathbb{R}^{4096} \rightarrow \mathbf{c} \in \mathbb{R}^{4096}$?
  - Macierz kowariancji $S = \frac{1}{m-1} X X^\top$
  - Wektory własne ($\rightarrow A$) oraz wartości własne (wariancja, posortowane)
  - Można wykorzystać pierwszych n wektorów – $\mathbf{x} \in \mathbb{R}^{4096} \rightarrow \mathbf{c} \in \mathbb{R}^{n}$; np. $n = 5, 10, 50$
- Rekonstrukcja z 1, 9, 17, ... (+8 składowych):

Analiza składowych głównych – przykład zastosowania (reprezentacja)

- $\mathbf{x} \in \mathbb{R}^{4096} \rightarrow \mathbf{c} \in \mathbb{R}^{4096}$?
  - Macierz kowariancji $S = \frac{1}{m-1} X X^\top$
  - Wektory własne ($\rightarrow A$) oraz wartości własne (wariancja, posortowane)
  - Można wykorzystać pierwszych n wektorów – $\mathbf{x} \in \mathbb{R}^{4096} \rightarrow \mathbf{c} \in \mathbb{R}^{n}$; np. $n = 5, 10, 50$
- Rekonstrukcja z 1, 9, 17, ... (+8 składowych):

Analiza składowych głównych – przykład zastosowania (reprezentacja)

- $\mathbf{x} \in \mathbb{R}^{4096} \rightarrow \mathbf{c} \in \mathbb{R}^{4096}$?
  - Macierz kowariancji $S = \frac{1}{m-1} X X^\top$
  - Wektory własne ($\rightarrow A$) oraz wartości własne (wariancja, posortowane)
  - Można wykorzystać pierwszych n wektorów – $\mathbf{x} \in \mathbb{R}^{4096} \rightarrow \mathbf{c} \in \mathbb{R}^{n}$; np. $n = 5, 10, 50$
- Rekonstrukcja z 1, 9, 17, ... (+8 składowych):

Analiza składowych głównych – przykład zastosowania (analiza)

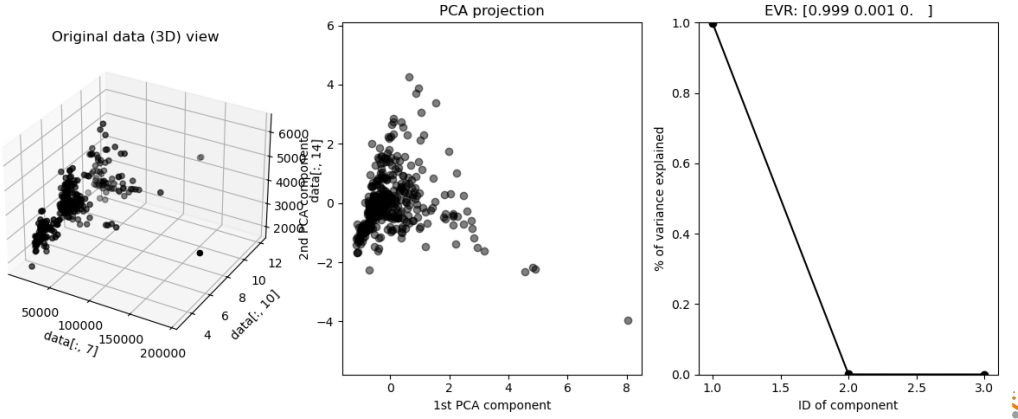
```
Chevrolet Aveo 4dr                  0 0 0 0 0 0 0 0  11690  10965 1.6  4 103 28 34 2370  98 167 66
Chevrolet Aveo LS 4dr hatch         0 0 0 0 0 0 0 0  12585  11802 1.6  4 103 28 34 2348  98 153 66
Chevrolet Cavalier 2dr              0 0 0 0 0 0 0 0  14610  13697 2.2  4 140 26 37 2617 104 183 69
Chevrolet Cavalier 4dr              0 0 0 0 0 0 0 0  14810  13884 2.2  4 140 26 37 2676 104 183 68
Chevrolet Cavalier LS 2dr           0 0 0 0 0 0 0 0  16385  15357 2.2  4 140 26 37 2617 104 183 69
Dodge Neon SE 4dr                   0 0 0 0 0 0 0 0  13670  12849 2.0  4 132 29 36 2581 105 174 67
Dodge Neon SXT 4dr                  0 0 0 0 0 0 0 0  15040  14086 2.0  4 132 29 36 2626 105 174 67
Ford Focus ZX3 2dr hatch
Ford Focus LX 4dr
Ford Focus SE 4dr
Ford Focus ZX5 5dr
Honda Civic DX 2dr
Honda Civic HX 2dr
Honda Civic LX 4dr
Hyundai Accent 2dr hatch
Hyundai Accent GL 4dr
Hyundai Accent GT 2dr hatch
Hyundai Elantra GLS 4dr
Hyundai Elantra GT 4dr
Hyundai Elantra GT 4dr hatch
Kia Optima LX 4dr
Kia Rio 4dr manual
Kia Rio 4dr auto
Kia Spectra 4dr
Kia Spectra GS 4dr hatch
Kia Spectra GSX 4dr hatch
Mazda3 i 4dr
Mini Cooper
Mitsubishi Lancer ES 4dr
Mitsubishi Lancer LS 4dr
Nissan Sentra 1.8 4dr
Nissan Sentra 1.8 S 4dr
Pontiac Sunfire 1SA 2dr
Saturn ion1 4dr
Saturn ion2 4dr
Saturn ion3 4dr
Saturn ion2 quad coupe 2dr
Saturn ion3 quad coupe 2dr
Scion xA 4dr hatch
Suzuki Aeno S 4dr
Suzuki Aerio LX 4dr
Suzuki Forenza S 4dr
Suzuki Forenza EX 4dr
Toyota Corolla CE 4dr
Toyota Corolla S 4dr
```

04cars.txt - Notepad

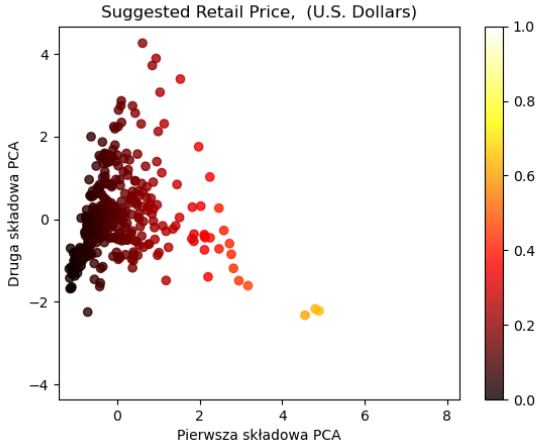
File Edit Format View Help

```
SOURCE:
_Kiplinger's Personal Finance_, December 2003, vol. 57, no. 12, pp. 104-123,
http:/www.kiplinger.com (permission to post on
the JSE Web site kindly granted by PARS International Corporation, 102 West 38th
Street, New York, NY 10018)

VARIABLE DESCRIPTIONS:

Columns Variables
  1- 45 Vehicle Name
 47             Sports Car? (1=yes, 0=no)
 49             Sport Utility Vehicle? (1=yes, 0=no)
 51             Wagon? (1=yes, 0=no)
 53             Minivan? (1=yes, 0=no)
 55             Pickup? (1=yes, 0=no)
 57             All-Wheel Drive? (1=yes, 0=no)
 59             Rear-Wheel Drive? (1=yes, 0=no)
 61- 66 Suggested Retail Price, what the manufacturer thinks the
                vehicle is worth, including adequate profit for the
                automaker and the dealer (U.S. Dollars)
 68- 73 Dealer Cost (or "invoice price"), what the dealership pays
                the manufacturer (U.S. Dollars)
 75- 77 Engine Size (liters)
 79- 80 Number of Cylinders (--1 if rotary engine)
 82- 84 Horsepower
 86- 87 City Miles Per Gallon
 89- 90 Highway Miles Per Gallon
 92- 95 Weight (Pounds)
 97- 99 Wheel Base (inches)
101-103 Length (inches)
105-106 Width (inches)
```

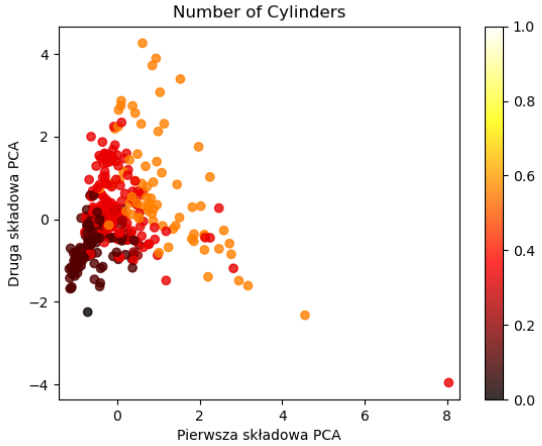
Analiza składowych głównych – przykład zastosowania (analiza)



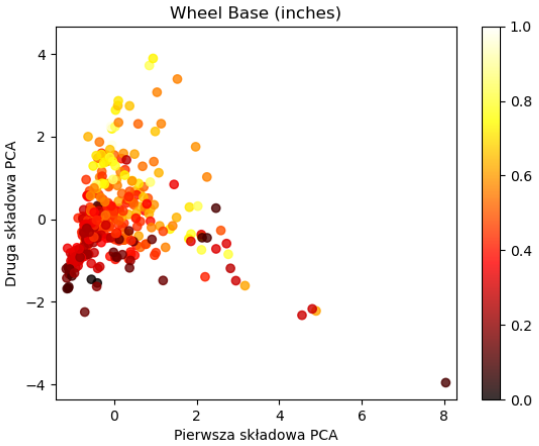
Analiza składowych głównych – przykład zastosowania (analiza)



Analiza składowych głównych – przykład zastosowania (analiza)



Analiza składowych głównych – przykład zastosowania (analiza)



Analiza składowych głównych – przykład zastosowania (analiza)

```
Component 0 ---------------------------

Examples
Lowest
['Kia Rio 4dr manual' 'Hyundai Accent 2dr hatch' 'Toyota Echo 2dr manual']
Highest
['Mercedes-Benz SL600 convertible 2dr' 'Mercedes-Benz CL600 2dr'
'Porsche 911 GT2 2dr']

Column correlations
07 1.00  61- 66 Suggested Retail Price, what the manufacturer thinks the vehicle
is worth, including adequate profit for the automaker and the dealer (U.S. Dollars)
08 1.00  68- 73 Dealer Cost (or "invoice price"), what the dealership pays
```

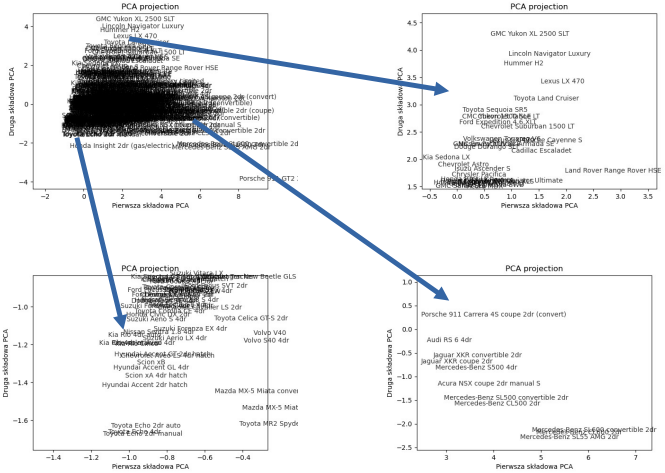
Analiza składowych głównych – przykład zastosowania (analiza)

```
Component 1 ----------------------------

Examples
Lowest
['Porsche 911 GT2 2dr' 'Mercedes-Benz SL55 AMG 2dr'
'Honda Insight 2dr (gas/electric)']
Highest
['Hummer H2' 'Lincoln Navigator Luxury' 'GMC Yukon XL 2500 SLT']

Column correlations
14 0.83  92- 95 Weight (Pounds)
17 0.74 105-106 Width (inches)
```

Analiza składowych głównych – przykład zastosowania (analiza)

```
Component 4 ---------------------------

Examples
Lowest
['Lincoln Town Car Ultimate L 4dr' 'Chrysler Concorde LX 4dr'
'Lincoln Town Car Ultimate 4dr']
Highest
['Mini Cooper S' 'Hummer H2' 'Jeep Wrangler Sahara convertible 2dr']

Column correlations
16 0.74 101-103 Length (inches)
15 0.51  97- 99 Wheel Base (inches)
```

Analiza składowych głównych – przykład zastosowania (analiza)



Analiza składowych głównych – podsumowanie

1. Reprezentacja x uwzględniająca charakterystykę źródła danych
   - Mniej współczynników do reprezentacji i klasyfikacji
   - Podzbiór współczynników zachowuje zakres zmienności danych
   - Analiza zbioru danych – data mining – weryfikacja obecności grup, anomalii, outlierów

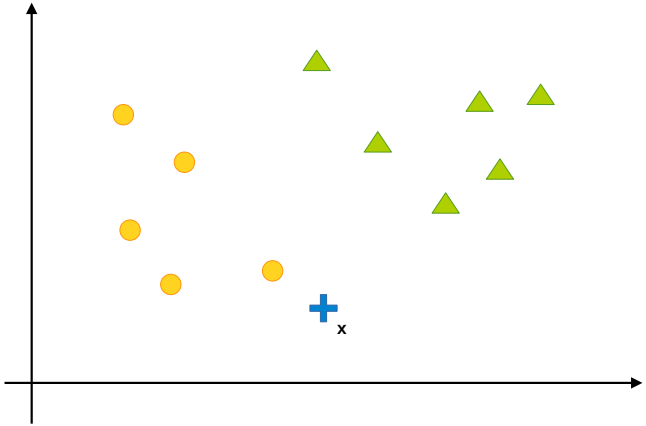
Analiza składowych głównych – podsumowanie

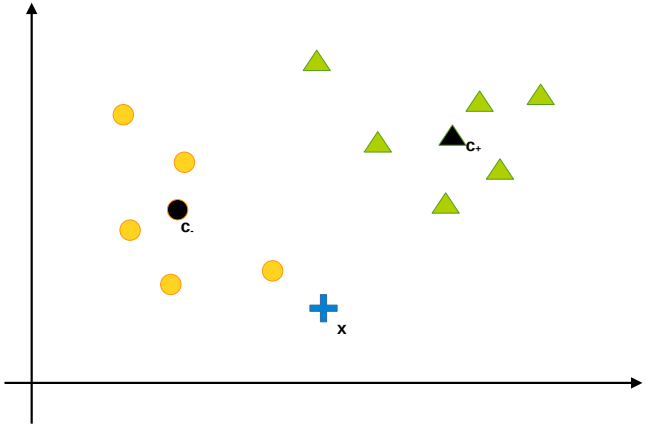
1. Reprezentacja x uwzględniająca charakterystykę źródła danych
   - Mniej współczynników do reprezentacji i klasyfikacji
   - Podzbiór współczynników zachowuje zakres zmienności danych
   - Analiza zbioru danych – data mining – weryfikacja obecności grup, anomalii, outlierów

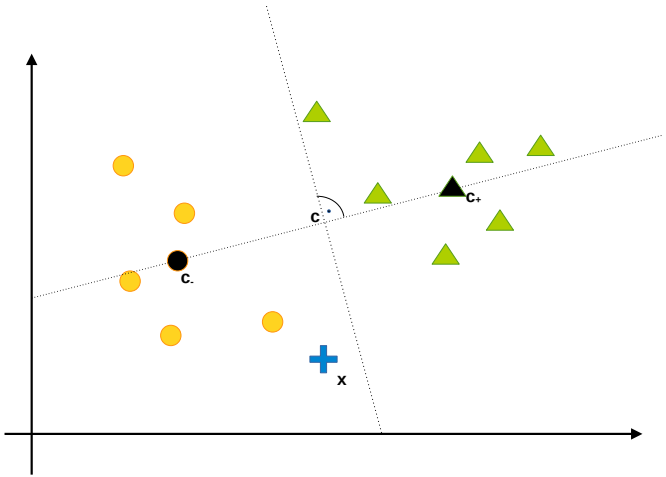
Analiza składowych głównych – podsumowanie

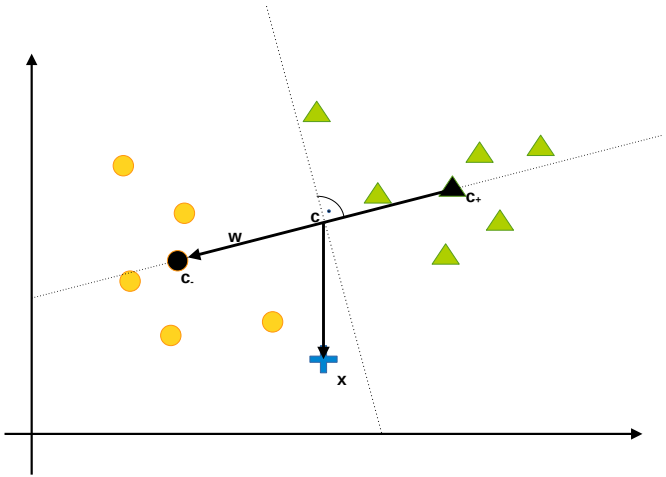
1. Reprezentacja x uwzględniająca charakterystykę źródła danych
   - Mniej współczynników do reprezentacji i klasyfikacji
   - Podzbiór współczynników zachowuje zakres zmienności danych
   - Analiza zbioru danych – data mining – weryfikacja obecności grup, anomalii, outlierów

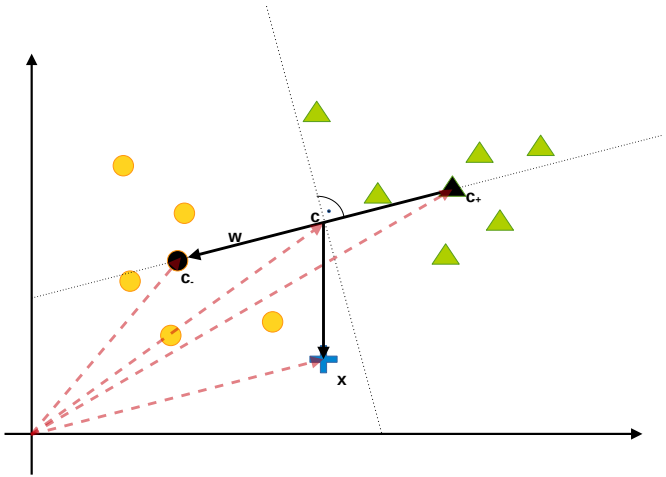
Support Vector Machine

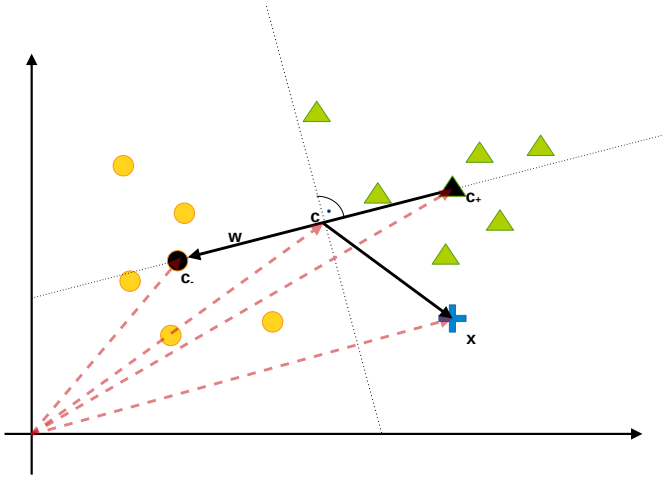

x

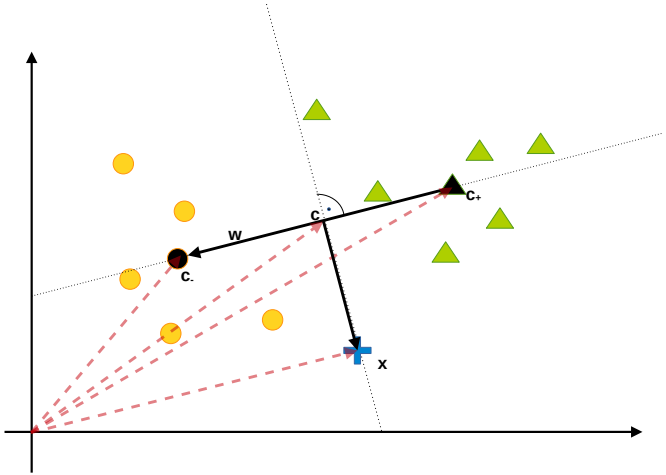

C+
C-
x

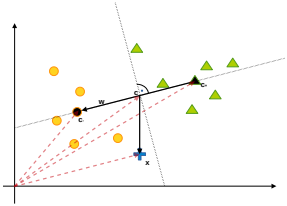

c+
c-
x

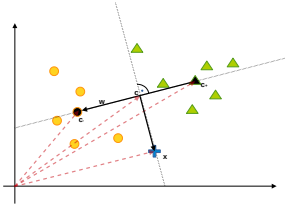

c*
c+
c-
w
x

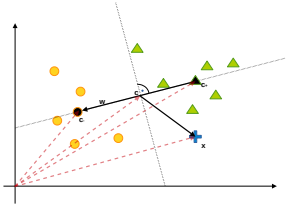

c*
c+
c-
w
x


c+
w
c-
x


c+
w
c-
x





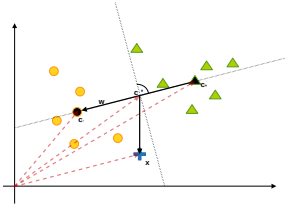


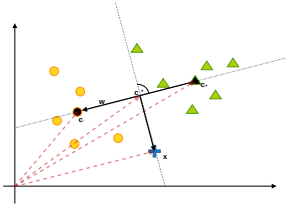
$\mathbf{c}_+ = \frac{1}{n_+}\sum_{i\in\mathcal{I}_+}\mathbf{x}_i \qquad \mathbf{c}_- = \frac{1}{n_-}\sum_{i\in\mathcal{I}_-}\mathbf{x}_i \qquad \mathbf{c} = \frac{\mathbf{c}_- + \mathbf{c}_+}{2} \qquad \mathbf{w} = \mathbf{c}_- - \mathbf{c}_+$

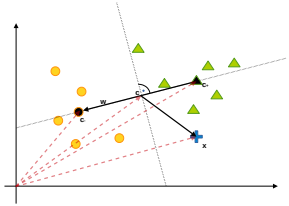
$\mathbf{y} = \operatorname{sgn}\langle(\mathbf{x}-\mathbf{c}),\mathbf{w}\rangle$ iloczyn skalarny $\langle\mathbf{a},\mathbf{b}\rangle = \|\mathbf{a}\|\|\mathbf{b}\|\cos\theta$

$$\mathbf{y} = \operatorname{sgn}\left\langle(\mathbf{x} - \tfrac{\mathbf{c}_- + \mathbf{c}_+}{2}), \mathbf{c}_- - \mathbf{c}_+\right\rangle =$$

$$\mathbf{y} = \operatorname{sgn}(\langle\mathbf{x},\mathbf{c}_-\rangle - \langle\mathbf{x},\mathbf{c}_+\rangle \underbrace{- \left\langle\frac{\mathbf{c}_- + \mathbf{c}_+}{2}, \mathbf{c}_-\right\rangle + \left\langle\frac{\mathbf{c}_- + \mathbf{c}_+}{2}, \mathbf{c}_+\right\rangle}_{\text{nie zależy od } \mathbf{x}})$$





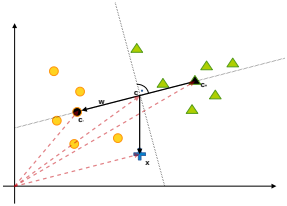


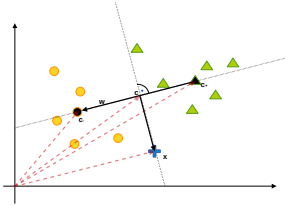
$\mathbf{c}_+ = \frac{1}{n_+}\sum_{i\in\mathcal{I}_+}\mathbf{x}_i \qquad \mathbf{c}_- = \frac{1}{n_-}\sum_{i\in\mathcal{I}_-}\mathbf{x}_i \qquad \mathbf{c} = \frac{\mathbf{c}_-+\mathbf{c}_+}{2} \qquad \mathbf{w} = \mathbf{c}_- - \mathbf{c}_+$

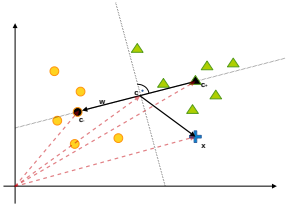
$\mathbf{y} = \operatorname{sgn}\langle(\mathbf{x}-\mathbf{c}),\mathbf{w}\rangle$ iloczyn skalarny $\langle\mathbf{a},\mathbf{b}\rangle = \|\mathbf{a}\|\|\mathbf{b}\|\cos\theta$

$\mathbf{y} = \operatorname{sgn}\left\langle(\mathbf{x}-\frac{\mathbf{c}_-+\mathbf{c}_+}{2}),\mathbf{c}_- - \mathbf{c}_+\right\rangle =$

$\mathbf{y} = \operatorname{sgn}(\langle\mathbf{x},\mathbf{c}_-\rangle - \langle\mathbf{x},\mathbf{c}_+\rangle \underbrace{- \left\langle\frac{\mathbf{c}_-+\mathbf{c}_+}{2},\mathbf{c}_-\right\rangle + \left\langle\frac{\mathbf{c}_-+\mathbf{c}_+}{2},\mathbf{c}_+\right\rangle}_{\text{nie zależy od } \mathbf{x}})$





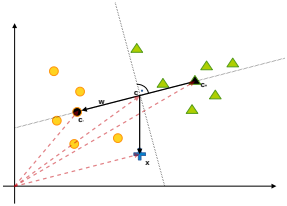


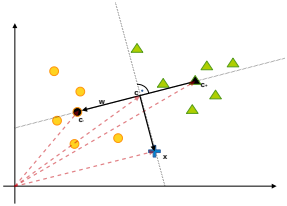
$\mathbf{c}_+ = \frac{1}{n_+}\sum_{i\in\mathcal{I}_+}\mathbf{x}_i \qquad \mathbf{c}_- = \frac{1}{n_-}\sum_{i\in\mathcal{I}_-}\mathbf{x}_i \qquad \mathbf{c} = \frac{\mathbf{c}_- + \mathbf{c}_+}{2} \qquad \mathbf{w} = \mathbf{c}_- - \mathbf{c}_+$

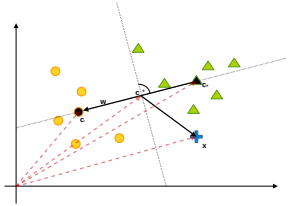
$\mathbf{y} = \operatorname{sgn}\langle(\mathbf{x} - \mathbf{c}), \mathbf{w}\rangle$ iloczyn skalarny $\langle \mathbf{a}, \mathbf{b}\rangle = \|\mathbf{a}\|\|\mathbf{b}\|\cos\theta$

$$\mathbf{y} = \operatorname{sgn}\left\langle(\mathbf{x} - \tfrac{\mathbf{c}_- + \mathbf{c}_+}{2}), \mathbf{c}_- - \mathbf{c}_+\right\rangle =$$

$$\mathbf{y} = \operatorname{sgn}(\langle \mathbf{x}, \mathbf{c}_-\rangle - \langle \mathbf{x}, \mathbf{c}_+\rangle \underbrace{- \left\langle \frac{\mathbf{c}_- + \mathbf{c}_+}{2}, \mathbf{c}_-\right\rangle + \left\langle \frac{\mathbf{c}_- + \mathbf{c}_+}{2}, \mathbf{c}_+\right\rangle}_{\text{nie zależy od } \mathbf{x}})$$





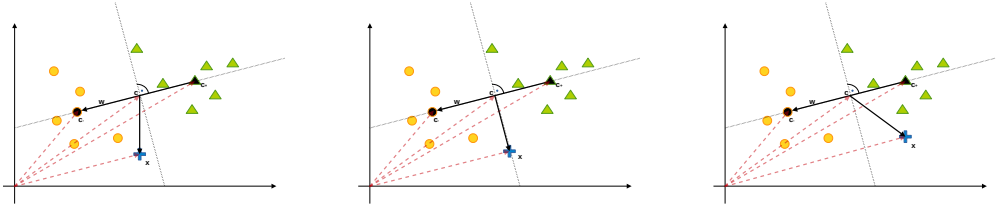


$\mathbf{c}_+ = \frac{1}{n_+}\sum_{i\in\mathcal{I}_+}\mathbf{x}_i \qquad \mathbf{c}_- = \frac{1}{n_-}\sum_{i\in\mathcal{I}_-}\mathbf{x}_i \qquad \mathbf{c} = \frac{\mathbf{c}_- + \mathbf{c}_+}{2} \qquad \mathbf{w} = \mathbf{c}_- - \mathbf{c}_+$

$\mathbf{y} = \operatorname{sgn}\langle(\mathbf{x}-\mathbf{c}),\mathbf{w}\rangle$ iloczyn skalarny $\langle\mathbf{a},\mathbf{b}\rangle = \|\mathbf{a}\|\|\mathbf{b}\|\cos\theta$

$\mathbf{y} = \operatorname{sgn}\left\langle(\mathbf{x} - \frac{\mathbf{c}_- + \mathbf{c}_+}{2}), \mathbf{c}_- - \mathbf{c}_+\right\rangle =$

$\mathbf{y} = \operatorname{sgn}(\langle\mathbf{x},\mathbf{c}_-\rangle - \langle\mathbf{x},\mathbf{c}_+\rangle \underbrace{- \left\langle\frac{\mathbf{c}_- + \mathbf{c}_+}{2}, \mathbf{c}_-\right\rangle + \left\langle\frac{\mathbf{c}_- + \mathbf{c}_+}{2}, \mathbf{c}_+\right\rangle}_{\text{nie zależy od } \mathbf{x}})$

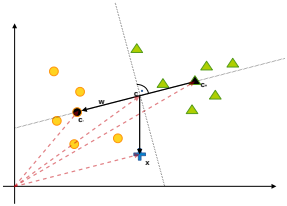


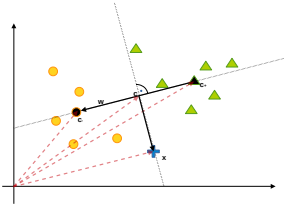
$$\mathbf{c}_+ = \frac{1}{n_+}\sum_{i\in\mathcal{I}_+}\mathbf{x}_i \qquad \mathbf{c}_- = \frac{1}{n_-}\sum_{i\in\mathcal{I}_-}\mathbf{x}_i \qquad \mathbf{c} = \frac{\mathbf{c}_- + \mathbf{c}_+}{2} \qquad \mathbf{w} = \mathbf{c}_- - \mathbf{c}_+$$

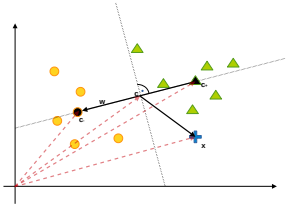
$$\mathbf{y} = \operatorname{sgn}\langle(\mathbf{x}-\mathbf{c}),\mathbf{w}\rangle \qquad \text{iloczyn skalarny } \langle\mathbf{a},\mathbf{b}\rangle = \|\mathbf{a}\|\|\mathbf{b}\|\cos\theta$$

$$\mathbf{y} = \operatorname{sgn}\left\langle(\mathbf{x} - \frac{\mathbf{c}_- + \mathbf{c}_+}{2}), \mathbf{c}_- - \mathbf{c}_+\right\rangle =$$

$$\mathbf{y} = \operatorname{sgn}(\langle\mathbf{x},\mathbf{c}_-\rangle - \langle\mathbf{x},\mathbf{c}_+\rangle \underbrace{- \left\langle\frac{\mathbf{c}_- + \mathbf{c}_+}{2}, \mathbf{c}_-\right\rangle + \left\langle\frac{\mathbf{c}_- + \mathbf{c}_+}{2}, \mathbf{c}_+\right\rangle}_{\text{nie zależy od } \mathbf{x}})$$





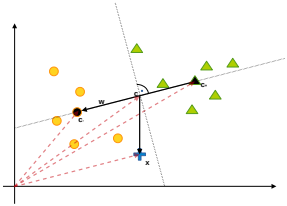


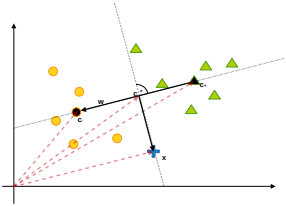
$$\mathbf{y} = \operatorname{sgn}(\langle \mathbf{x}, \mathbf{c}_- \rangle - \langle \mathbf{x}, \mathbf{c}_+ \rangle + b) \qquad \mathbf{c}_- = \frac{1}{n_-} \sum_{i \in \mathcal{I}_-} \mathbf{x}_i$$

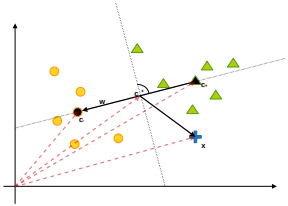
$$\mathbf{y} = \operatorname{sgn}\left(\frac{1}{n_-} \sum_{i \in \mathcal{I}_-} \langle \mathbf{x}_i, \mathbf{x} \rangle - \frac{1}{n_+} \sum_{i \in \mathcal{I}_+} \langle \mathbf{x}_i, \mathbf{x} \rangle + b \right)$$

$$\mathbf{y} = \operatorname{sgn}\left(- \sum_i y_i \langle \mathbf{x}_i, \mathbf{x} \rangle + b \right)$$

$$\mathbf{y} = \operatorname{sgn}\left(\sum_i \alpha_i \langle \mathbf{x}_i, \mathbf{x} \rangle + b \right)$$





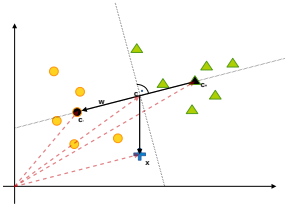


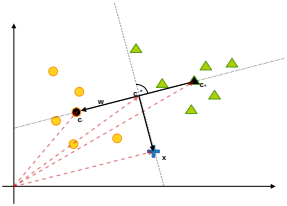
$$\mathbf{y} = \operatorname{sgn}(\langle \mathbf{x}, \mathbf{c}_- \rangle - \langle \mathbf{x}, \mathbf{c}_+ \rangle + b) \qquad \mathbf{c}_- = \frac{1}{n_-} \sum_{i \in \mathcal{I}_-} \mathbf{x}_i$$

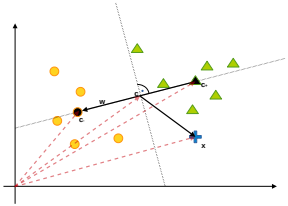
$$\mathbf{y} = \operatorname{sgn}\left(\frac{1}{n_-} \sum_{i \in \mathcal{I}_-} \langle \mathbf{x}_i, \mathbf{x} \rangle - \frac{1}{n_+} \sum_{i \in \mathcal{I}_+} \langle \mathbf{x}_i, \mathbf{x} \rangle + b \right)$$

$$\mathbf{y} = \operatorname{sgn}\left(- \sum_i y_i \langle \mathbf{x}_i, \mathbf{x} \rangle + b \right)$$

$$\mathbf{y} = \operatorname{sgn}\left(\sum_i \alpha_i \langle \mathbf{x}_i, \mathbf{x} \rangle + b \right)$$





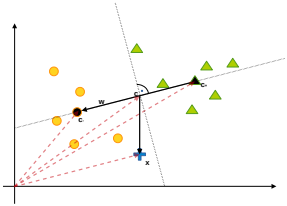


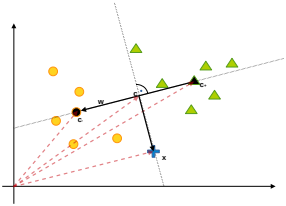
$$\mathbf{y} = \operatorname{sgn}(\langle \mathbf{x}, \mathbf{c}_- \rangle - \langle \mathbf{x}, \mathbf{c}_+ \rangle + b) \qquad \mathbf{c}_- = \frac{1}{n_-} \sum_{i \in \mathcal{I}_-} \mathbf{x}_i$$

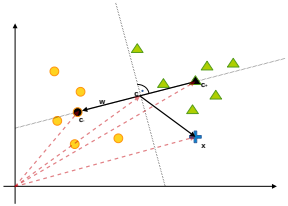
$$\mathbf{y} = \operatorname{sgn}\left(\frac{1}{n_-} \sum_{i \in \mathcal{I}_-} \langle \mathbf{x}_i, \mathbf{x} \rangle - \frac{1}{n_+} \sum_{i \in \mathcal{I}_+} \langle \mathbf{x}_i, \mathbf{x} \rangle + b \right)$$

$$\mathbf{y} = \operatorname{sgn}\left(- \sum_i y_i \langle \mathbf{x}_i, \mathbf{x} \rangle + b \right)$$

$$\mathbf{y} = \operatorname{sgn}\left(\sum_i \alpha_i \langle \mathbf{x}_i, \mathbf{x} \rangle + b \right)$$







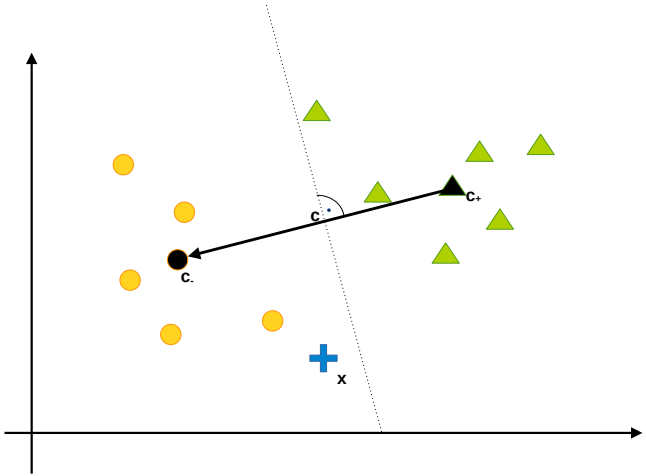
$$\mathbf{y} = \operatorname{sgn}(\langle \mathbf{x}, \mathbf{c}_- \rangle - \langle \mathbf{x}, \mathbf{c}_+ \rangle + b) \qquad \mathbf{c}_- = \frac{1}{n_-} \sum_{i \in \mathcal{I}_-} \mathbf{x}_i$$

$$\mathbf{y} = \operatorname{sgn}\left(\frac{1}{n_-} \sum_{i \in \mathcal{I}_-} \langle \mathbf{x}_i, \mathbf{x} \rangle - \frac{1}{n_+} \sum_{i \in \mathcal{I}_+} \langle \mathbf{x}_i, \mathbf{x} \rangle + b \right)$$

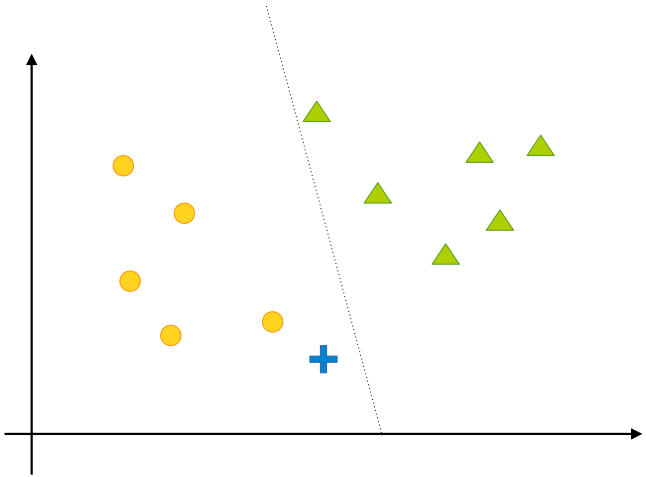
$$\mathbf{y} = \operatorname{sgn}\left(- \sum_i y_i \langle \mathbf{x}_i, \mathbf{x} \rangle + b \right)$$

$$\mathbf{y} = \operatorname{sgn}\left(\sum_i \alpha_i \langle \mathbf{x}_i, \mathbf{x} \rangle + b \right)$$

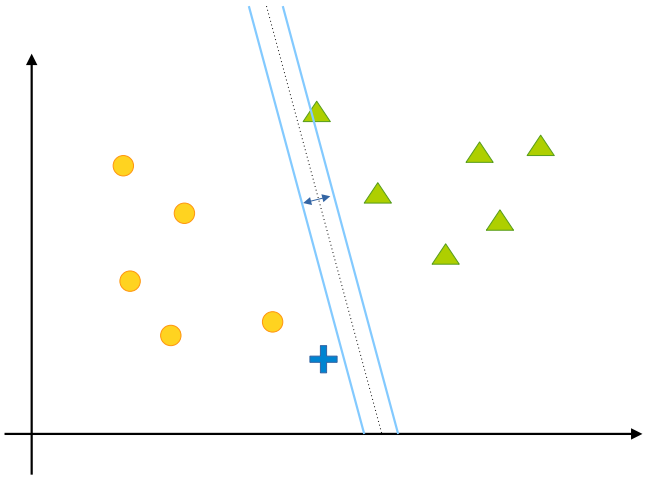
Margines



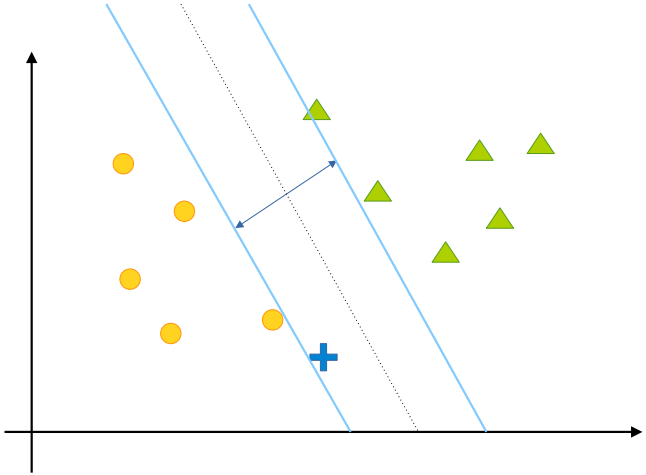
Margines



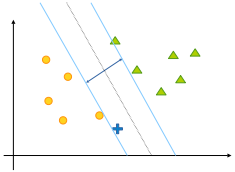
Margines



Margines



SVM

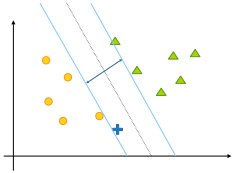


$$\mathbf{y} = \operatorname{sgn}\left(\sum_i \alpha_i \langle \mathbf{x}_i, \mathbf{x} \rangle + b\right)$$

$$\underset{\boldsymbol{\alpha} \in \mathbb{R}^n}{\text{maximize}} W(\boldsymbol{\alpha}) = \sum_i \alpha_i - \frac{1}{2} \sum_{ij} \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle$$

ograniczenia $\alpha_i \geqslant 0, \sum_i \alpha_i y_i = 0$

SVM

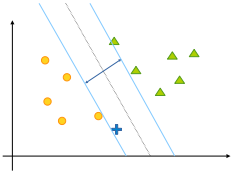


$$\mathbf{y} = \operatorname{sgn}\left(\sum_i \alpha_i \langle \mathbf{x}_i, \mathbf{x} \rangle + b\right)$$

zalety Brak [hiper]parametrów!
(Parametry α_i z optymalizacji, ale nie potrzeba np. learning rate, rozmiaru warstw, ...)

wady Tylko dla separowalnych zbiorów danych, których klasy da się rozdzielić hiperpłaszczyzną

SVM

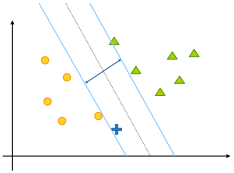


$$\mathbf{y} = \operatorname{sgn}\left(\sum_i \alpha_i \langle \mathbf{x}_i, \mathbf{x} \rangle + b\right)$$

zalety Brak [hiper]parametrów!
(Parametry α_i z optymalizacji, ale nie potrzeba np. learning rate, rozmiaru warstw, ...)

wady Tylko dla separowalnych zbiorów danych, których klasy da się rozdzielić hiperpłaszczyzną

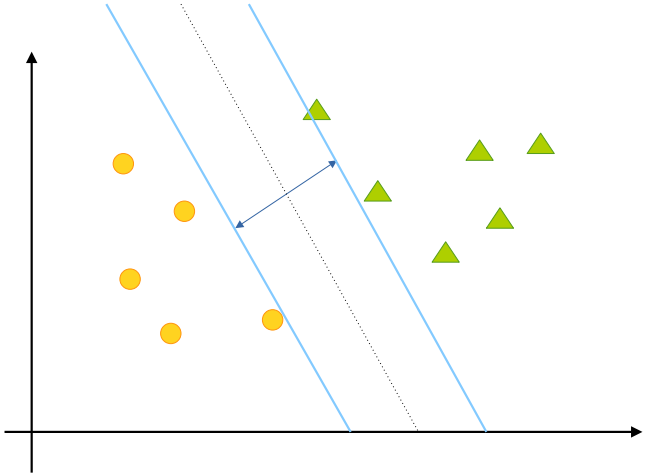
SVM

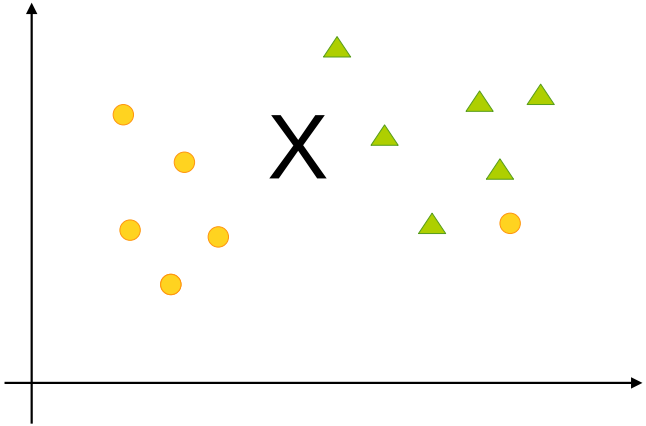


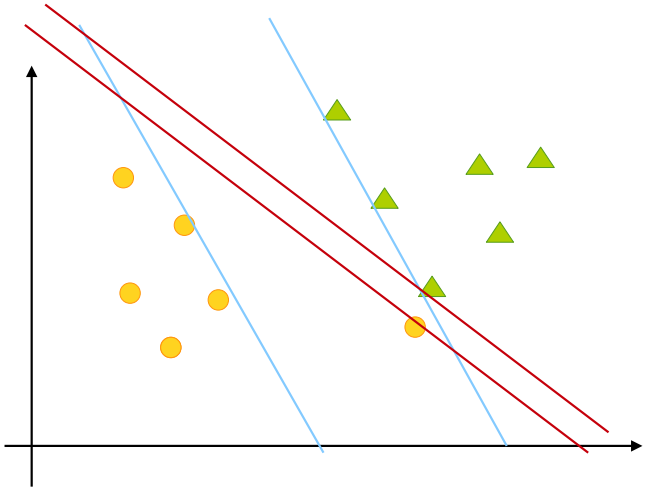
$$\mathbf{y} = \operatorname{sgn}\left(\sum_i \alpha_i \langle \mathbf{x}_i, \mathbf{x} \rangle + b\right)$$

zalety Brak [hiper]parametrów!
(Parametry α_i z optymalizacji, ale nie potrzeba np. learning rate, rozmiaru warstw, ...)

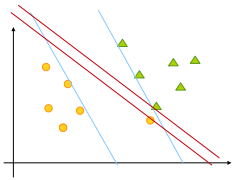
wady Tylko dla separowalnych zbiorów danych, których klasy da się rozdzielić hiperpłaszczyzną







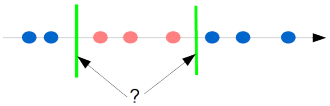
C-SVM



- Wprowadzenie do optymalizacji parametrów ξ_i – błędów klasyfikacji dla konkretnych przykładów ze zbioru treningowego
- Hiperparametr C – sterowanie dopasowania do danych vs wielkość marginesu
- C-SVM może zarówno klasyfikować zbiory niemożliwe do rozdzielenia hiperpłaszczyzną, jak i takie, w których pojedyncze przykłady zmniejszają margines
  - (Wciąż liniowe rozdzielenie $\rightarrow$ metody nieliniowe)

Metody nieliniowe

Kernel SVM

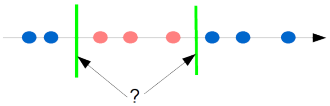


$$x \to x^2$$

Mapujemy dane do innej przestrzeni ($\{x_1, x_2, \dots\} \to \{x_1^2, x_2^2, \dots\}$), mapowanie może być nieliniowe, ale w tej nowej przestrzeni istnieje możliwość liniowego rozdzielenia klas.

W klasyfikatorze SVM, decyzja o klasyfikacji i optymalizacja wykorzystują wyłącznie iloczyn skalarny $\langle \cdot, \cdot \rangle$.

1. Możemy przetwarzać różne dane (nie tylko wektory), o ile dla elementów naszego zbioru danych $x_i, x_j \in \mathcal{X}$ możemy policzyć $\langle x_i, x_j \rangle$
2. Możemy przetwarzać przekształcenia danych $x_i \to \Phi(x_i)$, o ile możemy policzyć $\langle \Phi(x_i), \Phi(x_j) \rangle \dots$ niekoniecznie musimy wyliczać $\Phi(x_i)$!

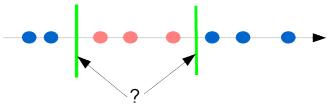


$$x \to x^2$$

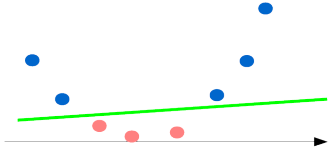
Mapujemy dane do innej przestrzeni ($\{x_1, x_2, \dots\} \to \{x_1^2, x_2^2, \dots\}$), mapowanie może być nieliniowe, ale w tej nowej przestrzeni istnieje możliwość liniowego rozdzielenia klas.

W klasyfikatorze SVM, decyzja o klasyfikacji i optymalizacja wykorzystują wyłącznie iloczyn skalarny $\langle \cdot, \cdot \rangle$.

1. Możemy przetwarzać różne dane (nie tylko wektory), o ile dla elementów naszego zbioru danych $x_i, x_j \in \mathcal{X}$ możemy policzyć $\langle x_i, x_j \rangle$
2. Możemy przetwarzać przekształcenia danych $x_i \to \Phi(x_i)$, o ile możemy policzyć $\langle \Phi(x_i), \Phi(x_j) \rangle \dots$ niekoniecznie musimy wyliczać $\Phi(x_i)$!



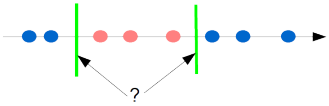
$x \to x^2$



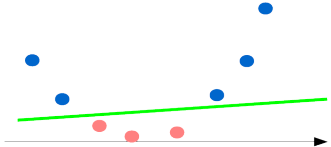
Mapujemy dane do innej przestrzeni ($\{x_1, x_2, \dots\} \to \{x_1^2, x_2^2, \dots\}$), mapowanie może być nieliniowe, ale w tej nowej przestrzeni istnieje możliwość liniowego rozdzielenia klas.

W klasyfikatorze SVM, decyzja o klasyfikacji i optymalizacja wykorzystują wyłącznie iloczyn skalarny $\langle \cdot, \cdot \rangle$.

1. Możemy przetwarzać różne dane (nie tylko wektory), o ile dla elementów naszego zbioru danych $x_i, x_j \in \mathcal{X}$ możemy policzyć $\langle x_i, x_j \rangle$
2. Możemy przetwarzać przekształcenia danych $x_i \to \Phi(x_i)$, o ile możemy policzyć $\langle \Phi(x_i), \Phi(x_j) \rangle \dots$ niekoniecznie musimy wyliczać $\Phi(x_i)$!



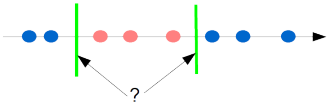
$x \to x^2$



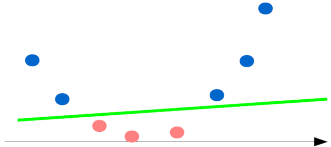
Mapujemy dane do innej przestrzeni ($\{x_1, x_2, \dots\} \to \{x_1^2, x_2^2, \dots\}$), mapowanie może być nieliniowe, ale w tej nowej przestrzeni istnieje możliwość liniowego rozdzielenia klas.

W klasyfikatorze SVM, decyzja o klasyfikacji i optymalizacja wykorzystują wyłącznie iloczyn skalarny $\langle \cdot, \cdot \rangle$.

1. Możemy przetwarzać różne dane (nie tylko wektory), o ile dla elementów naszego zbioru danych $x_i, x_j \in \mathcal{X}$ możemy policzyć $\langle x_i, x_j \rangle$
2. Możemy przetwarzać przekształcenia danych $x_i \to \Phi(x_i)$, o ile możemy policzyć $\langle \Phi(x_i), \Phi(x_j) \rangle \dots$ niekoniecznie musimy wyliczać $\Phi(x_i)$!



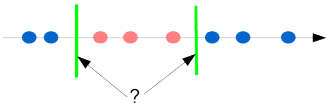
$x \to x^2$



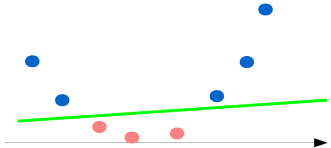
Mapujemy dane do innej przestrzeni ($\{x_1, x_2, \dots\} \to \{x_1^2, x_2^2, \dots\}$), mapowanie może być nieliniowe, ale w tej nowej przestrzeni istnieje możliwość liniowego rozdzielenia klas.

W klasyfikatorze SVM, decyzja o klasyfikacji i optymalizacja wykorzystują wyłącznie iloczyn skalarny $\langle \cdot, \cdot \rangle$.

1. Możemy przetwarzać różne dane (nie tylko wektory), o ile dla elementów naszego zbioru danych $x_i, x_j \in \mathcal{X}$ możemy policzyć $\langle x_i, x_j \rangle$
2. Możemy przetwarzać przekształcenia danych $x_i \to \Phi(x_i)$, o ile możemy policzyć $\langle \Phi(x_i), \Phi(x_j) \rangle \dots$ niekoniecznie musimy wyliczać $\Phi(x_i)$!



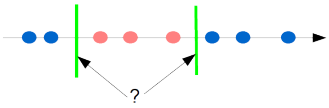
$x \to x^2$



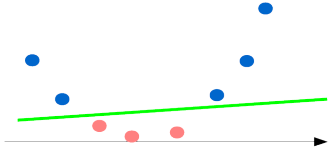
Mapujemy dane do innej przestrzeni ($\{x_1, x_2, \dots\} \to \{x_1^2, x_2^2, \dots\}$), mapowanie może być nieliniowe, ale w tej nowej przestrzeni istnieje możliwość liniowego rozdzielenia klas.

W klasyfikatorze SVM, decyzja o klasyfikacji i optymalizacja wykorzystują wyłącznie iloczyn skalarny $\langle \cdot, \cdot \rangle$.

1. Możemy przetwarzać różne dane (nie tylko wektory), o ile dla elementów naszego zbioru danych $x_i, x_j \in \mathcal{X}$ możemy policzyć $\langle x_i, x_j \rangle$
2. Możemy przetwarzać przekształcenia danych $x_i \to \Phi(x_i)$, o ile możemy policzyć $\langle \Phi(x_i), \Phi(x_j) \rangle \dots$ niekoniecznie musimy wyliczać $\Phi(x_i)$!



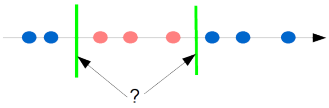
$x \to x^2$



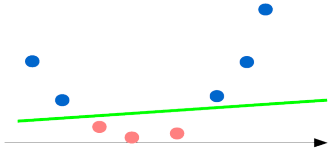
Mapujemy dane do innej przestrzeni ($\{x_1, x_2, \dots\} \to \{x_1^2, x_2^2, \dots\}$), mapowanie może być nieliniowe, ale w tej nowej przestrzeni istnieje możliwość liniowego rozdzielenia klas.

W klasyfikatorze SVM, decyzja o klasyfikacji i optymalizacja wykorzystują wyłącznie iloczyn skalarny $\langle \cdot, \cdot \rangle$.

1. Możemy przetwarzać różne dane (nie tylko wektory), o ile dla elementów naszego zbioru danych $x_i, x_j \in \mathcal{X}$ możemy policzyć $\langle x_i, x_j \rangle$
2. Możemy przetwarzać przekształcenia danych $x_i \to \Phi(x_i)$, o ile możemy policzyć $\langle \Phi(x_i), \Phi(x_j) \rangle \dots$ niekoniecznie musimy wyliczać $\Phi(x_i)$!



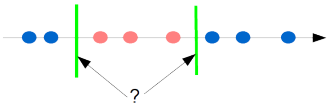
$x \to x^2$



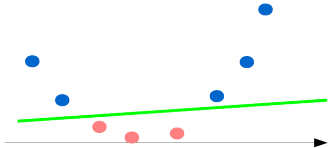
Mapujemy dane do innej przestrzeni ($\{x_1, x_2, \dots\} \to \{x_1^2, x_2^2, \dots\}$), mapowanie może być nieliniowe, ale w tej nowej przestrzeni istnieje możliwość liniowego rozdzielenia klas.

W klasyfikatorze SVM, decyzja o klasyfikacji i optymalizacja wykorzystują wyłącznie iloczyn skalarny $\langle \cdot, \cdot \rangle$.

1. Możemy przetwarzać różne dane (nie tylko wektory), o ile dla elementów naszego zbioru danych $x_i, x_j \in \mathcal{X}$ możemy policzyć $\langle x_i, x_j \rangle$
2. Możemy przetwarzać przekształcenia danych $x_i \to \Phi(x_i)$, o ile możemy policzyć $\langle \Phi(x_i), \Phi(x_j) \rangle \dots$ niekoniecznie musimy wyliczać $\Phi(x_i)$!



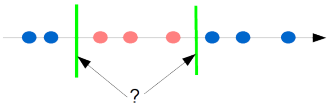
$x \to x^2$



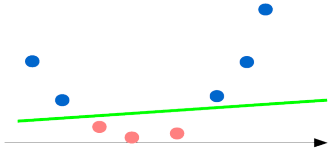
Mapujemy dane do innej przestrzeni ($\{x_1, x_2, \dots\} \to \{x_1^2, x_2^2, \dots\}$), mapowanie może być nieliniowe, ale w tej nowej przestrzeni istnieje możliwość liniowego rozdzielenia klas.

W klasyfikatorze SVM, decyzja o klasyfikacji i optymalizacja wykorzystują wyłącznie iloczyn skalarny $\langle \cdot, \cdot \rangle$.

1. Możemy przetwarzać różne dane (nie tylko wektory), o ile dla elementów naszego zbioru danych $x_i, x_j \in \mathcal{X}$ możemy policzyć $\langle x_i, x_j \rangle$
2. Możemy przetwarzać przekształcenia danych $x_i \to \Phi(x_i)$, o ile możemy policzyć $\langle \Phi(x_i), \Phi(x_j) \rangle \dots$ niekoniecznie musimy wyliczać $\Phi(x_i)$!



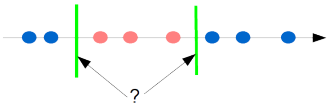
$x \to x^2$



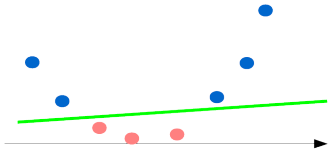
Mapujemy dane do innej przestrzeni ($\{x_1, x_2, \dots\} \to \{x_1^2, x_2^2, \dots\}$), mapowanie może być nieliniowe, ale w tej nowej przestrzeni istnieje możliwość liniowego rozdzielenia klas.

W klasyfikatorze SVM, decyzja o klasyfikacji i optymalizacja wykorzystują wyłącznie iloczyn skalarny $\langle \cdot, \cdot \rangle$.

1. Możemy przetwarzać różne dane (nie tylko wektory), o ile dla elementów naszego zbioru danych $x_i, x_j \in \mathcal{X}$ możemy policzyć $\langle x_i, x_j \rangle$
2. Możemy przetwarzać przekształcenia danych $x_i \to \Phi(x_i)$, o ile możemy policzyć $\langle \Phi(x_i), \Phi(x_j) \rangle$ . . . niekoniecznie musimy wyliczać $\Phi(x_i)$!



$x \to x^2$



Mapujemy dane do innej przestrzeni ($\{x_1, x_2, \dots\} \to \{x_1^2, x_2^2, \dots\}$), mapowanie może być nieliniowe, ale w tej nowej przestrzeni istnieje możliwość liniowego rozdzielenia klas.

W klasyfikatorze SVM, decyzja o klasyfikacji i optymalizacja wykorzystują wyłącznie iloczyn skalarny $\langle \cdot, \cdot \rangle$.

1. Możemy przetwarzać różne dane (nie tylko wektory), o ile dla elementów naszego zbioru danych $x_i, x_j \in \mathcal{X}$ możemy policzyć $\langle x_i, x_j \rangle$
2. Możemy przetwarzać przekształcenia danych $x_i \to \Phi(x_i)$, o ile możemy policzyć $\langle \Phi(x_i), \Phi(x_j) \rangle \dots$ niekoniecznie musimy wyliczać $\Phi(x_i)$!

$\langle x_i, x_j \rangle = x_i x_j$

$\left\langle x_i^2, x_j^2 \right\rangle = x_i^2 x_j^2 = (x_i x_j)^2 = \langle x_i, x_j \rangle^2$

A może przestrzeń funkcji $f, g \in \mathcal{H}, f(x) \in \mathbb{R}, x \in \mathcal{X} \ldots \langle f, g \rangle$?

Rozważmy funkcje w postaci $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R} \qquad (x_i, x_j) \mapsto k(x_i, x_j)$ o cechach:

- symetryczna $k(x_i, x_j) = k(x_j, x_i)$
- dodatnio określona $\sum_{i,j=1}^{n} c_i c_j k(x_i, x_j) \geqslant 0 \quad \forall n \in \mathbb{N} \, \forall x_1, \ldots, x_n \in \mathcal{X} \, \forall c_1, \ldots, c_n \in \mathbb{R}$

Na przykład $k(x_i, x_j) = \langle x_i, x_j \rangle^2$:) (a ogólnie $k(x_j, x_i) = (\langle x_i, x_j \rangle + c)^d, c \geqslant 0, d \in \mathbb{N}$)

Dla $x_1 = 42$, możemy zbudować funkcję $k_1(x) = \langle x, 42 \rangle^2 = 1764x^2$

- Dla punktów danych $x_i \in \mathcal{X}$ mamy funkcje $k(x_i, x) \in \mathbb{R}$ – przestrzeń funkcji
- $\langle k(x_i, x), k(x_j, x) \rangle$?

$$\langle x_i, x_j \rangle = x_i x_j$$

$$\left\langle x_i^2, x_j^2 \right\rangle = x_i^2 x_j^2 = (x_i x_j)^2 = \langle x_i, x_j \rangle^2$$

A może przestrzeń funkcji $f, g \in \mathcal{H}, f(x) \in \mathbb{R}, x \in \mathcal{X} \ldots \langle f, g \rangle$?

Rozważmy funkcje w postaci $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R} \qquad (x_i, x_j) \mapsto k(x_i, x_j)$ o cechach:

- symetryczna $k(x_i, x_j) = k(x_j, x_i)$
- dodatnio określona $\sum_{i,j=1}^{n} c_i c_j k(x_i, x_j) \geqslant 0 \quad \forall n \in \mathbb{N}\, \forall x_1, \ldots, x_n \in \mathcal{X}\, \forall c_1, \ldots, c_n \in \mathbb{R}$

Na przykład $k(x_i, x_j) = \langle x_i, x_j \rangle^2$:) (a ogólnie $k(x_j, x_i) = (\langle x_i, x_j \rangle + c)^d, c \geqslant 0, d \in \mathbb{N}$)

Dla $x_1 = 42$, możemy zbudować funkcję $k_1(x) = \langle x, 42 \rangle^2 = 1764x^2$

- Dla punktów danych $x_i \in \mathcal{X}$ mamy funkcje $k(x_i, x) \in \mathbb{R}$ – przestrzeń funkcji
- $\langle k(x_i, x), k(x_j, x) \rangle$?

$\langle x_i, x_j \rangle = x_i x_j$

$\left\langle x_i^2, x_j^2 \right\rangle = x_i^2 x_j^2 = (x_i x_j)^2 = \langle x_i, x_j \rangle^2$

A może przestrzeń funkcji $f, g \in \mathcal{H}, f(x) \in \mathbb{R}, x \in \mathcal{X} \ldots \langle f, g \rangle$?

Rozważmy funkcje w postaci $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R} \qquad (x_i, x_j) \mapsto k(x_i, x_j)$ o cechach:

- symetryczna $k(x_i, x_j) = k(x_j, x_i)$
- dodatnio określona $\sum_{i,j=1}^{n} c_i c_j k(x_i, x_j) \geqslant 0 \quad \forall n \in \mathbb{N}\, \forall x_1, \ldots, x_n \in \mathcal{X}\, \forall c_1, \ldots, c_n \in \mathbb{R}$

Na przykład $k(x_i, x_j) = \langle x_i, x_j \rangle^2$:) (a ogólnie $k(x_j, x_i) = (\langle x_i, x_j \rangle + c)^d, c \geqslant 0, d \in \mathbb{N}$)

Dla $x_1 = 42$, możemy zbudować funkcję $k_1(x) = \langle x, 42 \rangle^2 = 1764x^2$

- Dla punktów danych $x_i \in \mathcal{X}$ mamy funkcje $k(x_i, x) \in \mathbb{R}$ – przestrzeń funkcji
- $\langle k(x_i, x), k(x_j, x) \rangle$?

$\langle x_i, x_j \rangle = x_i x_j$

$\left\langle x_i^2, x_j^2 \right\rangle = x_i^2 x_j^2 = (x_i x_j)^2 = \langle x_i, x_j \rangle^2$

A może przestrzeń funkcji $f, g \in \mathcal{H}, f(x) \in \mathbb{R}, x \in \mathcal{X} \ldots \langle f, g \rangle$?

Rozważmy funkcje w postaci $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R} \qquad (x_i, x_j) \mapsto k(x_i, x_j)$ o cechach:

- symetryczna $k(x_i, x_j) = k(x_j, x_i)$
- dodatnio określona $\sum_{i,j=1}^{n} c_i c_j k(x_i, x_j) \geqslant 0 \quad \forall n \in \mathbb{N}\, \forall x_1, \ldots, x_n \in \mathcal{X}\, \forall c_1, \ldots, c_n \in \mathbb{R}$

Na przykład $k(x_i, x_j) = \langle x_i, x_j \rangle^2$:) (a ogólnie $k(x_j, x_i) = (\langle x_i, x_j \rangle + c)^d, c \geqslant 0, d \in \mathbb{N}$)

Dla $x_1 = 42$, możemy zbudować funkcję $k_1(x) = \langle x, 42 \rangle^2 = 1764x^2$

- Dla punktów danych $x_i \in \mathcal{X}$ mamy funkcje $k(x_i, x) \in \mathbb{R}$ – przestrzeń funkcji
- $\langle k(x_i, x), k(x_j, x) \rangle$?

$\langle x_i, x_j \rangle = x_i x_j$

$\left\langle x_i^2, x_j^2 \right\rangle = x_i^2 x_j^2 = (x_i x_j)^2 = \langle x_i, x_j \rangle^2$

A może przestrzeń funkcji $f, g \in \mathcal{H}, f(x) \in \mathbb{R}, x \in \mathcal{X} \ldots \langle f, g \rangle$?

Rozważmy funkcje w postaci $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R} \qquad (x_i, x_j) \mapsto k(x_i, x_j)$ o cechach:

- symetryczna $k(x_i, x_j) = k(x_j, x_i)$
- dodatnio określona $\sum_{i,j=1}^{n} c_i c_j k(x_i, x_j) \geqslant 0 \quad \forall n \in \mathbb{N}\, \forall x_1, \ldots, x_n \in \mathcal{X}\, \forall c_1, \ldots, c_n \in \mathbb{R}$

Na przykład $k(x_i, x_j) = \langle x_i, x_j \rangle^2$:) (a ogólnie $k(x_j, x_i) = (\langle x_i, x_j \rangle + c)^d, c \geqslant 0, d \in \mathbb{N}$)

Dla $x_1 = 42$, możemy zbudować funkcję $k_1(x) = \langle x, 42 \rangle^2 = 1764x^2$

- Dla punktów danych $x_i \in \mathcal{X}$ mamy funkcje $k(x_i, x) \in \mathbb{R}$ – przestrzeń funkcji
- $\langle k(x_i, x), k(x_j, x) \rangle$?

$\langle x_i, x_j \rangle = x_i x_j$

$\left\langle x_i^2, x_j^2 \right\rangle = x_i^2 x_j^2 = (x_i x_j)^2 = \langle x_i, x_j \rangle^2$

A może przestrzeń funkcji $f, g \in \mathcal{H}, f(x) \in \mathbb{R}, x \in \mathcal{X} \ldots \langle f, g \rangle$?

Rozważmy funkcje w postaci $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R} \qquad (x_i, x_j) \mapsto k(x_i, x_j)$ o cechach:

- symetryczna $k(x_i, x_j) = k(x_j, x_i)$
- dodatnio określona $\sum_{i,j=1}^{n} c_i c_j k(x_i, x_j) \geqslant 0 \quad \forall n \in \mathbb{N}\, \forall x_1, \ldots, x_n \in \mathcal{X}\, \forall c_1, \ldots, c_n \in \mathbb{R}$

Na przykład $k(x_i, x_j) = \langle x_i, x_j \rangle^2$:) (a ogólnie $k(x_j, x_i) = (\langle x_i, x_j \rangle + c)^d, c \geqslant 0, d \in \mathbb{N}$)

Dla $x_1 = 42$, możemy zbudować funkcję $k_1(x) = \langle x, 42 \rangle^2 = 1764x^2$

- Dla punktów danych $x_i \in \mathcal{X}$ mamy funkcje $k(x_i, x) \in \mathbb{R}$ – przestrzeń funkcji
- $\langle k(x_i, x), k(x_j, x) \rangle$?

$\langle x_i, x_j \rangle = x_i x_j$

$\left\langle x_i^2, x_j^2 \right\rangle = x_i^2 x_j^2 = (x_i x_j)^2 = \langle x_i, x_j \rangle^2$

A może przestrzeń funkcji $f, g \in \mathcal{H}, f(x) \in \mathbb{R}, x \in \mathcal{X} \ldots \langle f, g \rangle$?

Rozważmy funkcje w postaci $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R} \qquad (x_i, x_j) \mapsto k(x_i, x_j)$ o cechach:

- symetryczna $k(x_i, x_j) = k(x_j, x_i)$
- dodatnio określona $\sum_{i,j=1}^{n} c_i c_j k(x_i, x_j) \geqslant 0 \quad \forall n \in \mathbb{N}\, \forall x_1, \ldots, x_n \in \mathcal{X}\, \forall c_1, \ldots, c_n \in \mathbb{R}$

Na przykład $k(x_i, x_j) = \langle x_i, x_j \rangle^2$:) (a ogólnie $k(x_j, x_i) = (\langle x_i, x_j \rangle + c)^d, c \geqslant 0, d \in \mathbb{N}$)

Dla $x_1 = 42$, możemy zbudować funkcję $k_1(x) = \langle x, 42 \rangle^2 = 1764x^2$

- Dla punktów danych $x_i \in \mathcal{X}$ mamy funkcje $k(x_i, x) \in \mathbb{R}$ – przestrzeń funkcji
- $\langle k(x_i, x), k(x_j, x) \rangle$?

$\langle k(x_i, x), k(x_j, x) \rangle$?

Jeżeli funkcja w postaci $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R} \quad (x_i, x_j) \mapsto k(x_i, x_j)$ jest symetryczna i dodatnio określona, to (twierdzenie Moore'a–Aronszajna) istnieje przestrzeń Hilberta funkcji określonych na $\mathcal{X}$ w której k jest reprodukującym jądrem (reproducing kernel).

- Przestrzeń Hilberta – przestrzeń liniowa (dodawanie, mnożenie przez skalar) + zdefiniowany iloczyn skalarny który indukuje metrykę (możemy wyznaczyć odległość między elementami zbioru)
- Przestrzeń Hilberta z reprodukującym jądrem (RKHS) – dla każdego $x_i \in \mathcal{X}$ istnieje funkcja $k_{x_i} \in \mathcal{H}$ taka że dla każdej funkcji z tej przestrzeni $f \in \mathcal{H}$ zachodzi $\langle f, k_{x_i} \rangle = f(x_i)$

$\langle k(x_i, x), k(x_j, x) \rangle = k(x_i, x_j)$

$\langle k(x_i, x), k(x_j, x) \rangle$?

Jeżeli funkcja w postaci $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R} \quad (x_i, x_j) \mapsto k(x_i, x_j)$ jest symetryczna i dodatnio określona, to (twierdzenie Moore'a–Aronszajna) istnieje przestrzeń Hilberta funkcji określonych na $\mathcal{X}$ w której k jest reprodukującym jądrem (reproducing kernel).

- Przestrzeń Hilberta – przestrzeń liniowa (dodawanie, mnożenie przez skalar) + zdefiniowany iloczyn skalarny który indukuje metrykę (możemy wyznaczyć odległość między elementami zbioru)
- Przestrzeń Hilberta z reprodukującym jądrem (RKHS) – dla każdego $x_i \in \mathcal{X}$ istnieje funkcja $k_{x_i} \in \mathcal{H}$ taka że dla każdej funkcji z tej przestrzeni $f \in \mathcal{H}$ zachodzi $\langle f, k_{x_i} \rangle = f(x_i)$

$\langle k(x_i, x), k(x_j, x) \rangle = k(x_i, x_j)$

$\langle k(x_i, x), k(x_j, x) \rangle$?

Jeżeli funkcja w postaci $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R} \quad (x_i, x_j) \mapsto k(x_i, x_j)$ jest symetryczna i dodatnio określona, to (twierdzenie Moore'a–Aronszajna) istnieje przestrzeń Hilberta funkcji określonych na $\mathcal{X}$ w której k jest reprodukującym jądrem (reproducing kernel).

- Przestrzeń Hilberta – przestrzeń liniowa (dodawanie, mnożenie przez skalar) + zdefiniowany iloczyn skalarny który indukuje metrykę (możemy wyznaczyć odległość między elementami zbioru)
- Przestrzeń Hilberta z reprodukującym jądrem (RKHS) – dla każdego $x_i \in \mathcal{X}$ istnieje funkcja $k_{x_i} \in \mathcal{H}$ taka że dla każdej funkcji z tej przestrzeni $f \in \mathcal{H}$ zachodzi $\langle f, k_{x_i} \rangle = f(x_i)$

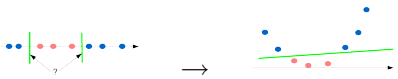
$\langle k(x_i, x), k(x_j, x) \rangle = k(x_i, x_j)$

$\langle k(x_i, x), k(x_j, x) \rangle$?

Jeżeli funkcja w postaci $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R} \quad (x_i, x_j) \mapsto k(x_i, x_j)$ jest symetryczna i dodatnio określona, to (twierdzenie Moore'a–Aronszajna) istnieje przestrzeń Hilberta funkcji określonych na $\mathcal{X}$ w której k jest reprodukującym jądrem (reproducing kernel).

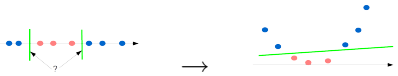
- Przestrzeń Hilberta – przestrzeń liniowa (dodawanie, mnożenie przez skalar) + zdefiniowany iloczyn skalarny który indukuje metrykę (możemy wyznaczyć odległość między elementami zbioru)
- Przestrzeń Hilberta z reprodukującym jądrem (RKHS) – dla każdego $x_i \in \mathcal{X}$ istnieje funkcja $k_{x_i} \in \mathcal{H}$ taka że dla każdej funkcji z tej przestrzeni $f \in \mathcal{H}$ zachodzi $\langle f, k_{x_i} \rangle = f(x_i)$

$\langle k(x_i, x), k(x_j, x) \rangle = k(x_i, x_j)$



Podsumowanie:

1. Mamy zbiór danych $\mathcal{X}$, o złożonej strukturze (nieliniowa granica decyzyjna)
2. Mamy narzędzie liniowej klasyfikacji, oparte o iloczyn skalarny
3. Szukamy nieliniowego przekształcenia danych do innej przestrzeni, gdzie możemy liniowo rozdzielić klasy
4. Jeżeli znajdziemy dobrą funkcję k, to dzięki własnościom RKHS możemy odwzorować dane w innej przestrzeni $\mathcal{H}$, opartej o k, w której wyliczymy iloczyn skalarny jako $k(x_i, x_j)$
5. Ponieważ odwzorowanie jest nieliniowe, liniowe rozdzielenie w $\mathcal{H}$ przekłada się na nieliniową granicę decyzyjną w $\mathcal{X}$

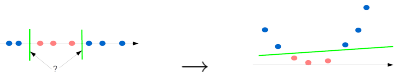


SVM $\mathbf{y} = \operatorname{sgn}\left(\sum_i \alpha_i \langle \mathbf{x}_i, \mathbf{x} \rangle + b\right)$ $\quad \rightarrow \quad$ kernel-SVM $\mathbf{y} = \operatorname{sgn}\left(\sum_i \alpha_i k(\mathbf{x}_i, \mathbf{x}) + b\right)$

Popularne funkcje k:

- RBF $k(\mathbf{x}_i, \mathbf{x}_j) = e^{-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2}$
- Wielomianowa $k(\mathbf{x}_i, \mathbf{x}_j) = \left(\langle \mathbf{x}_i, \mathbf{x}_j \rangle + c\right)^d, c \geqslant 0, d \in \mathbb{N}$
- Sigmoid $k(\mathbf{x}_i, \mathbf{x}_j) = \ldots$

Najczęściej wystarczy jedna dobra funkcja – RBF



SVM $\mathbf{y} = \operatorname{sgn}\left(\sum_i \alpha_i \langle \mathbf{x}_i, \mathbf{x} \rangle + b\right)$ $\quad \rightarrow \quad$ kernel-SVM $\mathbf{y} = \operatorname{sgn}\left(\sum_i \alpha_i k(\mathbf{x}_i, \mathbf{x}) + b\right)$

Popularne funkcje k:

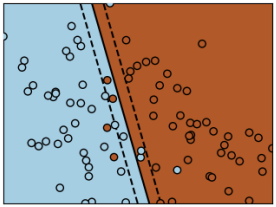
- RBF $k(\mathbf{x}_i, \mathbf{x}_j) = e^{-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2}$
- Wielomianowa $k(\mathbf{x}_i, \mathbf{x}_j) = \left(\langle \mathbf{x}_i, \mathbf{x}_j \rangle + c\right)^d, c \geqslant 0, d \in \mathbb{N}$
- Sigmoid $k(\mathbf{x}_i, \mathbf{x}_j) = \ldots$

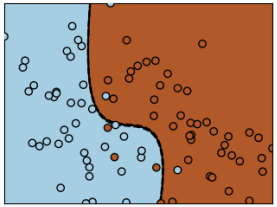
Najczęściej wystarczy jedna dobra funkcja – RBF

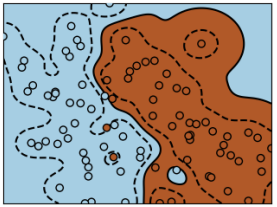
The idea of SVMs: map the training data into a higher-dimensional feature space via Φ, and construct a separating hyperplane with maximum margin there. This yields a nonlinear decision boundary in input space. By the use of a kernel function, it is possible to compute the separating hyperplane without explicitly carrying out the map into the feature space

Learning with Kernels, str. 29

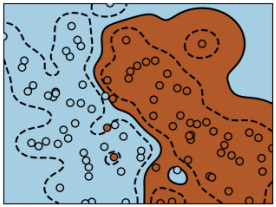
Przykłady funkcji jądrowych: liniowa, wielomianowa, RBF

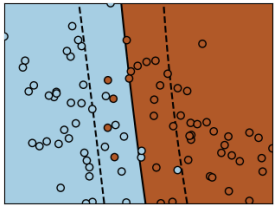


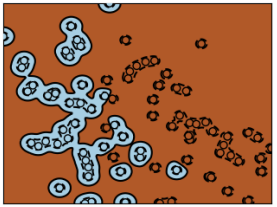




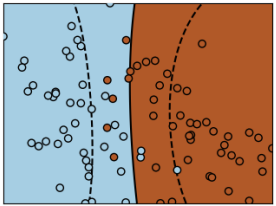
Przykłady funkcji jądrowej RBF dla różnych wartości parametru γ,
kolejno wartości γ: dopasowana do zbioru danych, niska, wysoka

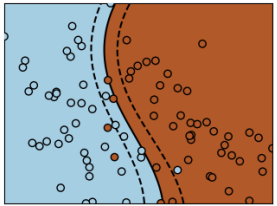


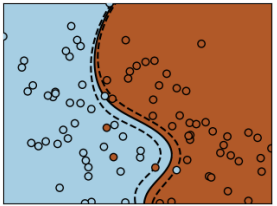




Przykłady funkcji jądrowej RBF dla dopasowanej do danych wartości γ, kolejno wartości C: niska, dopasowana, wysoka

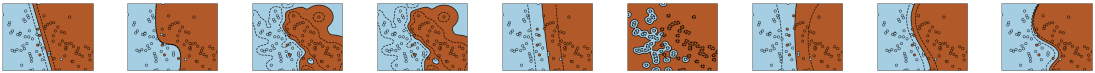






Kernel SVM:

- Skuteczny w praktyce klasyfikator
- Bardzo dobrze opracowany teoretycznie
- Dwa – tylko dwa – hiperparametry, o łatwej interpretacji
- „Kernel trick" – możliwość uogólnienia algorytmów dających się sformułować w oparciu o iloczyn skalarny (np. PCA → kernel PCA)



Pomiar wydajności klasyfikatorów

- Skąd wiemy, jak dobry jest klasyfikator?
- Kiedy jeden klasyfikator jest lepszy od drugiego?
- Jak diagnozować problem z wydajnością/skutecznością?

Miary wydajności

- Skuteczność (accuracy) $\frac{T}{T+F}$
  - Intuicyjna, naturalna
  - Wrażliwa na różną liczebność klas
  - Wrażliwa na rodzaj błędów
- Macierz błędów (confusion matrix)

	Predicted Positive	Predicted Negative
Actual Positive	TP (True Positive)	FN (False Negative)
Actual Negative	FP (False Positive)	TN (True Negative)

- Precison $\frac{TP}{TP+FP}$ oraz recall $\frac{TP}{TP+FN}$
- Sensitivity (czułość) $\frac{TP}{TP+FN}$ oraz specificity (swoistość) $\frac{TN}{TN+FP}$
- True positive ratio $\frac{TP}{TP+FN}$ oraz false positive ratio $\frac{FP}{FP+TN}$
- F1 score $\frac{2}{p^{-1}+r^{-1}} = 2\frac{p \times r}{p+r} = \frac{2TP}{2TP+FP+FN}$

- Skuteczność (accuracy) $\frac{T}{T+F}$
  - Intuicyjna, naturalna
  - Wrażliwa na różną liczebność klas
  - Wrażliwa na rodzaj błędów
- Macierz błędów (confusion matrix)

	Predicted Positive	Predicted Negative
Actual Positive	TP (True Positive)	FN (False Negative)
Actual Negative	FP (False Positive)	TN (True Negative)

- Precison $\frac{TP}{TP+FP}$ oraz recall $\frac{TP}{TP+FN}$
- Sensitivity (czułość) $\frac{TP}{TP+FN}$ oraz specificity (swoistość) $\frac{TN}{TN+FP}$
- True positive ratio $\frac{TP}{TP+FN}$ oraz false positive ratio $\frac{FP}{FP+TN}$
- F1 score $\frac{2}{p^{-1}+r^{-1}} = 2\frac{p \times r}{p+r} = \frac{2TP}{2TP+FP+FN}$

- Skuteczność (accuracy) $\frac{T}{T+F}$
  - Intuicyjna, naturalna
  - Wrażliwa na różną liczebność klas
  - Wrażliwa na rodzaj błędów
- Macierz błędów (confusion matrix)

	Predicted Positive	Predicted Negative
Actual Positive	TP (True Positive)	FN (False Negative)
Actual Negative	FP (False Positive)	TN (True Negative)

- Precison $\frac{TP}{TP+FP}$ oraz recall $\frac{TP}{TP+FN}$
- Sensitivity (czułość) $\frac{TP}{TP+FN}$ oraz specificity (swoistość) $\frac{TN}{TN+FP}$
- True positive ratio $\frac{TP}{TP+FN}$ oraz false positive ratio $\frac{FP}{FP+TN}$
- F1 score $\frac{2}{p^{-1}+r^{-1}} = 2\frac{p \times r}{p+r} = \frac{2TP}{2TP+FP+FN}$

- Skuteczność (accuracy) $\frac{T}{T+F}$
  - Intuicyjna, naturalna
  - Wrażliwa na różną liczebność klas
  - Wrażliwa na rodzaj błędów
- Macierz błędów (confusion matrix)

	Predicted Positive	Predicted Negative
Actual Positive	TP (True Positive)	FN (False Negative)
Actual Negative	FP (False Positive)	TN (True Negative)

- Precison $\frac{TP}{TP+FP}$ oraz recall $\frac{TP}{TP+FN}$
- Sensitivity (czułość) $\frac{TP}{TP+FN}$ oraz specificity (swoistość) $\frac{TN}{TN+FP}$
- True positive ratio $\frac{TP}{TP+FN}$ oraz false positive ratio $\frac{FP}{FP+TN}$
- F1 score $\frac{2}{p^{-1}+r^{-1}} = 2\frac{p \times r}{p+r} = \frac{2TP}{2TP+FP+FN}$

- Skuteczność (accuracy) $\frac{T}{T+F}$
  - Intuicyjna, naturalna
  - Wrażliwa na różną liczebność klas
  - Wrażliwa na rodzaj błędów
- Macierz błędów (confusion matrix)

	Predicted Positive	**Predicted Negative**
Actual Positive	TP (True Positive)	FN (False Negative)
Actual Negative	FP (False Positive)	TN (True Negative)

- Precison $\frac{TP}{TP+FP}$ oraz recall $\frac{TP}{TP+FN}$
- Sensitivity (czułość) $\frac{TP}{TP+FN}$ oraz specificity (swoistość) $\frac{TN}{TN+FP}$
- True positive ratio $\frac{TP}{TP+FN}$ oraz false positive ratio $\frac{FP}{FP+TN}$
- F1 score $\frac{2}{p^{-1}+r^{-1}} = 2\frac{p \times r}{p+r} = \frac{2TP}{2TP+FP+FN}$

- Skuteczność (accuracy) $\frac{T}{T+F}$ $\frac{TP+TN}{TP+FN+FP+TN}$
  - Intuicyjna, naturalna
  - Wrażliwa na różną liczebność klas
  - Wrażliwa na rodzaj błędów
- Macierz błędów (confusion matrix)

	Predicted Positive	**Predicted Negative**
Actual Positive	TP (True Positive)	FN (False Negative)
Actual Negative	FP (False Positive)	TN (True Negative)

- Precison $\frac{TP}{TP+FP}$ oraz recall $\frac{TP}{TP+FN}$
- Sensitivity (czułość) $\frac{TP}{TP+FN}$ oraz specificity (swoistość) $\frac{TN}{TN+FP}$
- True positive ratio $\frac{TP}{TP+FN}$ oraz false positive ratio $\frac{FP}{FP+TN}$
- F1 score $\frac{2}{p^{-1}+r^{-1}} = 2\frac{p \times r}{p+r} = \frac{2TP}{2TP+FP+FN}$

- Skuteczność (accuracy) $\frac{T}{T+F}$ $\frac{TP+TN}{TP+FN+FP+TN}$
  - Intuicyjna, naturalna
  - Wrażliwa na różną liczebność klas
  - Wrażliwa na rodzaj błędów
- Macierz błędów (confusion matrix)

	Predicted Positive	**Predicted Negative**
Actual Positive	TP (True Positive)	FN (False Negative)
Actual Negative	FP (False Positive)	TN (True Negative)

- Precison $\frac{TP}{TP+FP}$ oraz recall $\frac{TP}{TP+FN}$
- Sensitivity (czułość) $\frac{TP}{TP+FN}$ oraz specificity (swoistość) $\frac{TN}{TN+FP}$
- True positive ratio $\frac{TP}{TP+FN}$ oraz false positive ratio $\frac{FP}{FP+TN}$
- F1 score $\frac{2}{p^{-1}+r^{-1}} = 2\frac{p \times r}{p+r} = \frac{2TP}{2TP+FP+FN}$

- Skuteczność (accuracy) $\frac{T}{T+F}$ $\frac{TP+TN}{TP+FN+FP+TN}$
  - Intuicyjna, naturalna
  - Wrażliwa na różną liczebność klas
  - Wrażliwa na rodzaj błędów
- Macierz błędów (confusion matrix)

	Predicted Positive	**Predicted Negative**
Actual Positive	TP (True Positive)	FN (False Negative)
Actual Negative	FP (False Positive)	TN (True Negative)

- Precison $\frac{TP}{TP+FP}$ oraz recall $\frac{TP}{TP+FN}$
- Sensitivity (czułość) $\frac{TP}{TP+FN}$ oraz specificity (swoistość) $\frac{TN}{TN+FP}$
- True positive ratio $\frac{TP}{TP+FN}$ oraz false positive ratio $\frac{FP}{FP+TN}$
- F1 score $\frac{2}{p^{-1}+r^{-1}} = 2\frac{p \times r}{p+r} = \frac{2TP}{2TP+FP+FN}$

- Skuteczność (accuracy) $\frac{T}{T+F}$ $\frac{TP+TN}{TP+FN+FP+TN}$
  - Intuicyjna, naturalna
  - Wrażliwa na różną liczebność klas
  - Wrażliwa na rodzaj błędów
- Macierz błędów (confusion matrix)

	Predicted Positive	**Predicted Negative**
Actual Positive	TP (True Positive)	FN (False Negative)
Actual Negative	FP (False Positive)	TN (True Negative)

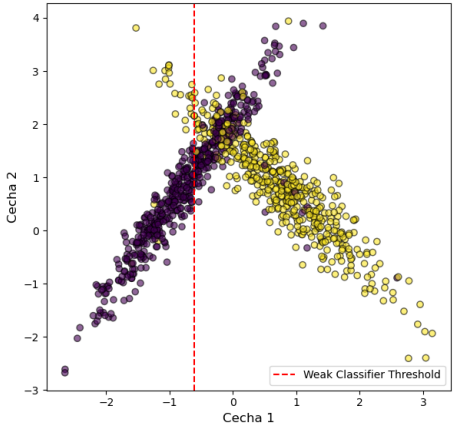
- Precison $\frac{TP}{TP+FP}$ oraz recall $\frac{TP}{TP+FN}$
- Sensitivity (czułość) $\frac{TP}{TP+FN}$ oraz specificity (swoistość) $\frac{TN}{TN+FP}$
- True positive ratio $\frac{TP}{TP+FN}$ oraz false positive ratio $\frac{FP}{FP+TN}$
- F1 score $\frac{2}{p^{-1}+r^{-1}} = 2\frac{p \times r}{p+r} = \frac{2TP}{2TP+FP+FN}$

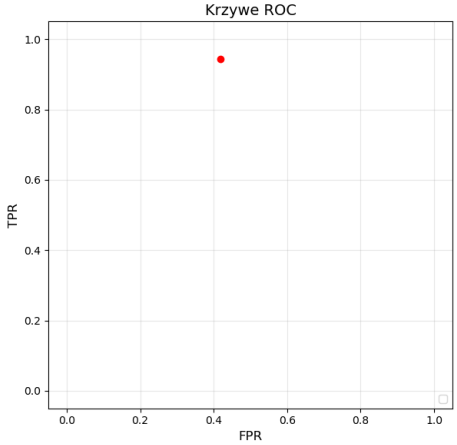
Jeden klasyfikator, wydajność – niezależnie od hiperparametrów?

Prosty klasyfikator (średnie klas), progowanie szacowanego prawopodobieństwa $T = 0.1$

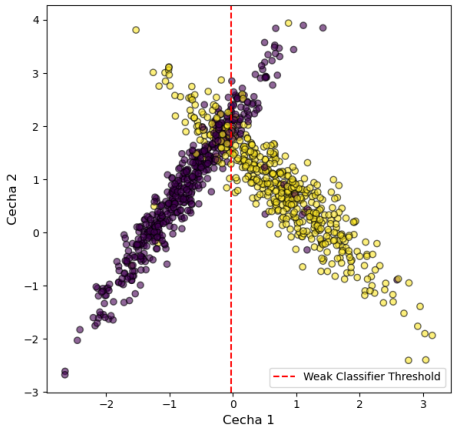
Jeden klasyfikator, wydajność – niezależnie od hiperparametrów?

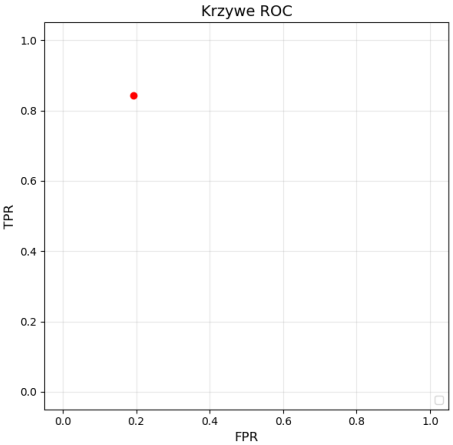
Prosty klasyfikator (średnie klas), progowanie szacowanego prawopodobieństwa $T = 0.1$



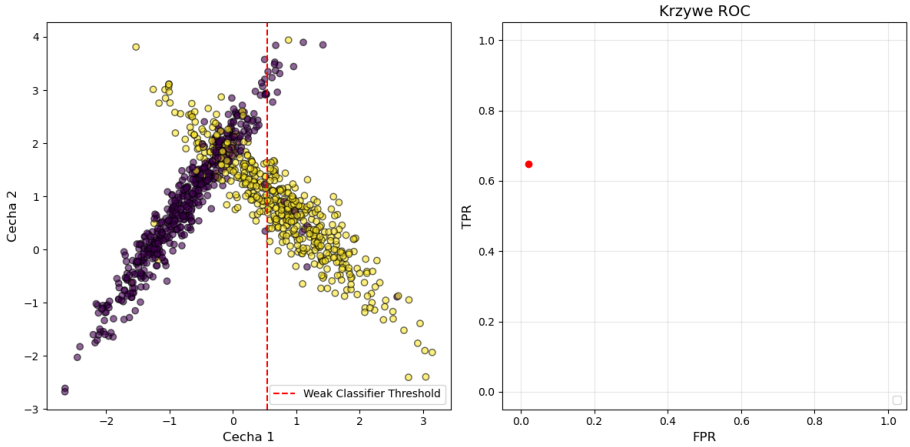


Prosty klasyfikator (średnie klas), progowanie szacowanego prawopodobieństwa $T = 0.5$

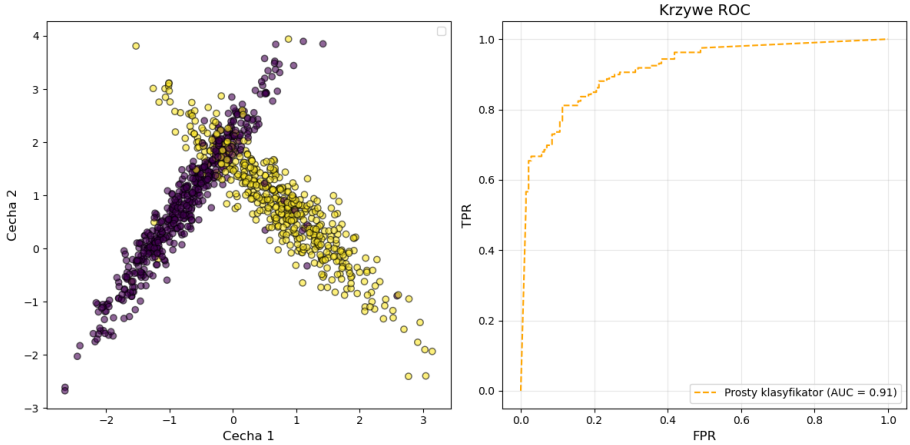




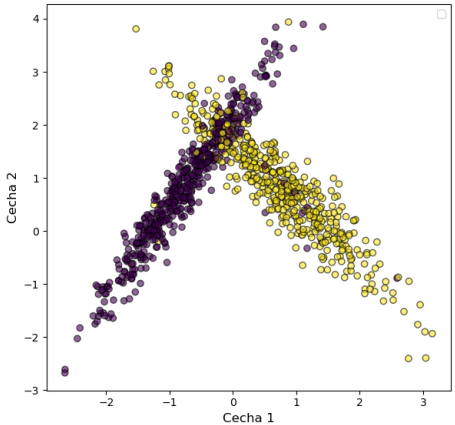
Prosty klasyfikator (średnie klas), progowanie szacowanego prawopodobieństwa $T = 0.9$

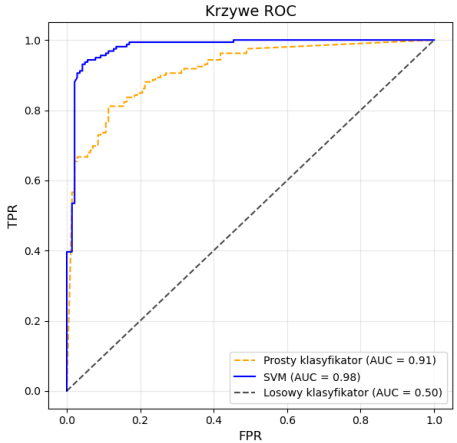


Prosty klasyfikator (średnie klas), progowanie szacowanego prawopodobieństwa – krzywa charakterystyki operatora (receiver operator characteristic, ROC)



Prosty klasyfikator vs SVM – krzywe ROC





Dwa klasyfikatory, wydajność, porównanie?

Np. 88% vs 89% - a co jeżeli to przypadek, i na innej części zbioru danych będzie 89% vs 88%?

(alternatywnie – który zestaw hiperparametrów wybrać?)

1. Testy statystyczne (→ wykłady dr hab. Hanna Wojewódka-Ściążko)
2. Pary wyników (dla konkretnego punktu danych wyniki obu klasyfikatorów) – różnice
3. Czy mają rozkład normalny (test Shapiro-Wilk)?
4. Jeżeli tak (r.n.) – t-test dla par
   - Hipoteza zerowa tego testu mówi o równych średnich, a alternatywna o różnych, zatem odrzucenie hipotezy zerowej świadczy o tym, że wyniki jednej z metod są istotnie statystycznie różne od drugiej
   - Jednostronnie lub dwustronnie
5. Jeżeli nie (r.n.) – Wilcoxon signed-rank test
6. Testy pozwalają określić szansę (przez p-wartość) odpowiedniości lub przewagi klasyfikatorów

Dwa klasyfikatory, wydajność, porównanie?

Np. 88% vs 89% - a co jeżeli to przypadek, i na innej części zbioru danych będzie 89% vs 88%?

(alternatywnie – który zestaw hiperparametrów wybrać?)

1. Testy statystyczne (→ wykłady dr hab. Hanna Wojewódka-Ściążko)
2. Pary wyników (dla konkretnego punktu danych wyniki obu klasyfikatorów) – różnice
3. Czy mają rozkład normalny (test Shapiro-Wilk)?
4. Jeżeli tak (r.n.) – t-test dla par
   - Hipoteza zerowa tego testu mówi o równych średnich, a alternatywna o różnych, zatem odrzucenie hipotezy zerowej świadczy o tym, że wyniki jednej z metod są istotnie statystycznie różne od drugiej
   - Jednostronnie lub dwustronnie
5. Jeżeli nie (r.n.) – Wilcoxon signed-rank test
6. Testy pozwalają określić szansę (przez p-wartość) odpowiedniości lub przewagi klasyfikatorów

Dwa klasyfikatory, wydajność, porównanie?

Np. 88% vs 89% - a co jeżeli to przypadek, i na innej części zbioru danych będzie 89% vs 88%?

(alternatywnie – który zestaw hiperparametrów wybrać?)

1. Testy statystyczne (→ wykłady dr hab. Hanna Wojewódka-Ściążko)
2. Pary wyników (dla konkretnego punktu danych wyniki obu klasyfikatorów) – różnice
3. Czy mają rozkład normalny (test Shapiro-Wilk)?
4. Jeżeli tak (r.n.) – t-test dla par
   - Hipoteza zerowa tego testu mówi o równych średnich, a alternatywna o różnych, zatem odrzucenie hipotezy zerowej świadczy o tym, że wyniki jednej z metod są istotnie statystycznie różne od drugiej
   - Jednostronnie lub dwustronnie
5. Jeżeli nie (r.n.) – Wilcoxon signed-rank test
6. Testy pozwalają określić szansę (przez p-wartość) odpowiedniości lub przewagi klasyfikatorów

Dwa klasyfikatory, wydajność, porównanie?

Np. 88% vs 89% - a co jeżeli to przypadek, i na innej części zbioru danych będzie 89% vs 88%?

(alternatywnie – który zestaw hiperparametrów wybrać?)

1. Testy statystyczne (→ wykłady dr hab. Hanna Wojewódka-Ściążko)
2. Pary wyników (dla konkretnego punktu danych wyniki obu klasyfikatorów) – różnice
3. Czy mają rozkład normalny (test Shapiro-Wilk)?
4. Jeżeli tak (r.n.) – t-test dla par
   - Hipoteza zerowa tego testu mówi o równych średnich, a alternatywna o różnych, zatem odrzucenie hipotezy zerowej świadczy o tym, że wyniki jednej z metod są istotnie statystycznie różne od drugiej
   - Jednostronnie lub dwustronnie
5. Jeżeli nie (r.n.) – Wilcoxon signed-rank test
6. Testy pozwalają określić szansę (przez p-wartość) odpowiedniości lub przewagi klasyfikatorów

Dwa klasyfikatory, wydajność, porównanie?

Np. 88% vs 89% - a co jeżeli to przypadek, i na innej części zbioru danych będzie 89% vs 88%?

(alternatywnie – który zestaw hiperparametrów wybrać?)

1. Testy statystyczne (→ wykłady dr hab. Hanna Wojewódka-Ściążko)
2. Pary wyników (dla konkretnego punktu danych wyniki obu klasyfikatorów) – różnice
3. Czy mają rozkład normalny (test Shapiro-Wilk)?
4. Jeżeli tak (r.n.) – t-test dla par
   - Hipoteza zerowa tego testu mówi o równych średnich, a alternatywna o różnych, zatem odrzucenie hipotezy zerowej świadczy o tym, że wyniki jednej z metod są istotnie statystycznie różne od drugiej
   - Jednostronnie lub dwustronnie
5. Jeżeli nie (r.n.) – Wilcoxon signed-rank test
6. Testy pozwalają określić szansę (przez p-wartość) odpowiedniości lub przewagi klasyfikatorów

Dwa klasyfikatory, wydajność, porównanie?

Np. 88% vs 89% - a co jeżeli to przypadek, i na innej części zbioru danych będzie 89% vs 88%?

(alternatywnie – który zestaw hiperparametrów wybrać?)

1. Testy statystyczne (→ wykłady dr hab. Hanna Wojewódka-Ściążko)
2. Pary wyników (dla konkretnego punktu danych wyniki obu klasyfikatorów) – różnice
3. Czy mają rozkład normalny (test Shapiro-Wilk)?
4. Jeżeli tak (r.n.) – t-test dla par
   - Hipoteza zerowa tego testu mówi o równych średnich, a alternatywna o różnych, zatem odrzucenie hipotezy zerowej świadczy o tym, że wyniki jednej z metod są istotnie statystycznie różne od drugiej
   - Jednostronnie lub dwustronnie
5. Jeżeli nie (r.n.) – Wilcoxon signed-rank test
6. Testy pozwalają określić szansę (przez p-wartość) odpowiedniości lub przewagi klasyfikatorów

Dwa klasyfikatory, wydajność, porównanie?

Np. 88% vs 89% - a co jeżeli to przypadek, i na innej części zbioru danych będzie 89% vs 88%?

(alternatywnie – który zestaw hiperparametrów wybrać?)

1. Testy statystyczne (→ wykłady dr hab. Hanna Wojewódka-Ściążko)
2. Pary wyników (dla konkretnego punktu danych wyniki obu klasyfikatorów) – różnice
3. Czy mają rozkład normalny (test Shapiro-Wilk)?
4. Jeżeli tak (r.n.) – t-test dla par
   - Hipoteza zerowa tego testu mówi o równych średnich, a alternatywna o różnych, zatem odrzucenie hipotezy zerowej świadczy o tym, że wyniki jednej z metod są istotnie statystycznie różne od drugiej
   - Jednostronnie lub dwustronnie
5. Jeżeli nie (r.n.) – Wilcoxon signed-rank test
6. Testy pozwalają określić szansę (przez p-wartość) odpowiedniości lub przewagi klasyfikatorów

Dwa klasyfikatory, wydajność, porównanie?

Np. 88% vs 89% - a co jeżeli to przypadek, i na innej części zbioru danych będzie 89% vs 88%?

(alternatywnie – który zestaw hiperparametrów wybrać?)

1. Testy statystyczne (→ wykłady dr hab. Hanna Wojewódka-Ściążko)
2. Pary wyników (dla konkretnego punktu danych wyniki obu klasyfikatorów) – różnice
3. Czy mają rozkład normalny (test Shapiro-Wilk)?
4. Jeżeli tak (r.n.) – t-test dla par
   - Hipoteza zerowa tego testu mówi o równych średnich, a alternatywna o różnych, zatem odrzucenie hipotezy zerowej świadczy o tym, że wyniki jednej z metod są istotnie statystycznie różne od drugiej
   - Jednostronnie lub dwustronnie
5. Jeżeli nie (r.n.) – Wilcoxon signed-rank test
6. Testy pozwalają określić szansę (przez p-wartość) odpowiedniości lub przewagi klasyfikatorów

Dwa klasyfikatory, wydajność, porównanie?

Np. 88% vs 89% - a co jeżeli to przypadek, i na innej części zbioru danych będzie 89% vs 88%?

(alternatywnie – który zestaw hiperparametrów wybrać?)

1. Testy statystyczne (→ wykłady dr hab. Hanna Wojewódka-Ściążko)
2. Pary wyników (dla konkretnego punktu danych wyniki obu klasyfikatorów) – różnice
3. Czy mają rozkład normalny (test Shapiro-Wilk)?
4. Jeżeli tak (r.n.) – t-test dla par
   - Hipoteza zerowa tego testu mówi o równych średnich, a alternatywna o różnych, zatem odrzucenie hipotezy zerowej świadczy o tym, że wyniki jednej z metod są istotnie statystycznie różne od drugiej
   - Jednostronnie lub dwustronnie
5. Jeżeli nie (r.n.) – Wilcoxon signed-rank test
6. Testy pozwalają określić szansę (przez p-wartość) odpowiedniości lub przewagi klasyfikatorów

Walidacja krzyżowa

Klasyfikator, skuteczność 89% . . . oczekiwania? Spodziewany wynik „w działaniu", najlepsze możliwe oszacowanie, prawdziwa skuteczność klasyfikacji, powtarzalność, uogólnianie. . .

- Zbiór danych – reprezentatywny dla problemu
- Wydzielenie zbioru testowego:
  - Brak – oszacowanie obciążone (skuteczność zawyżona)
  - 50% trening, 50% test – oszacowanie obciążone (skuteczność zaniżona)
  - Wydzielenie pojedynczego przykładu (leave one out) + powtarzanie – duża wariancja, złożone obliczeniowo
    - Ale – leave one patient out!
  - Wydzielenie części (np. 5, 10, 20%) i powtórzenie – walidacja krzyżowa, optymalne
  - Wydzielenie zbioru testowego – niezależna weryfikacja wydajności
- Wydzielenie zbioru walidacyjnego
  - Skąd wiemy jakie hiperparametry ustawić?
  - Wewnętrzna pętla walidacji żeby ustalić wartości hiperparametrów
  - Dwa poziomy walidacji (zewnętrzna – zbiór testowy, wewnętrzna – zbiór walidacyjny) – niezależna weryfikacja wydajności uwzględniająca niepewność doboru hiperparametrów
    - Testowanie metody vs testowanie metody z hiperparametrami dobranymi przez eksperta

Klasyfikator, skuteczność 89% . . . oczekiwania? Spodziewany wynik „w działaniu", najlepsze możliwe oszacowanie, prawdziwa skuteczność klasyfikacji, powtarzalność, uogólnianie. . .

- Zbiór danych – reprezentatywny dla problemu
- Wydzielenie zbioru testowego:
  - Brak – oszacowanie obciążone (skuteczność zawyżona)
  - 50% trening, 50% test – oszacowanie obciążone (skuteczność zaniżona)
  - Wydzielenie pojedynczego przykładu (leave one out) + powtarzanie – duża wariancja, złożone obliczeniowo
    - Ale – leave one patient out!
  - Wydzielenie części (np. 5, 10, 20%) i powtórzenie – walidacja krzyżowa, optymalne
  - Wydzielenie zbioru testowego – niezależna weryfikacja wydajności
- Wydzielenie zbioru walidacyjnego
  - Skąd wiemy jakie hiperparametry ustawić?
  - Wewnętrzna pętla walidacji żeby ustalić wartości hiperparametrów
  - Dwa poziomy walidacji (zewnętrzna – zbiór testowy, wewnętrzna – zbiór walidacyjny) – niezależna weryfikacja wydajności uwzględniająca niepewność doboru hiperparametrów
    - Testowanie metody vs testowanie metody z hiperparametrami dobranymi przez eksperta

Klasyfikator, skuteczność 89% . . . oczekiwania? Spodziewany wynik „w działaniu", najlepsze możliwe oszacowanie, prawdziwa skuteczność klasyfikacji, powtarzalność, uogólnianie. . .

- Zbiór danych – reprezentatywny dla problemu
- Wydzielenie zbioru testowego:
  - Brak – oszacowanie obciążone (skuteczność zawyżona)
  - 50% trening, 50% test – oszacowanie obciążone (skuteczność zaniżona)
  - Wydzielenie pojedynczego przykładu (leave one out) + powtarzanie – duża wariancja, złożone obliczeniowo
    - Ale – leave one patient out!
  - Wydzielenie części (np. 5, 10, 20%) i powtórzenie – walidacja krzyżowa, optymalne
  - Wydzielenie zbioru testowego – niezależna weryfikacja wydajności
- Wydzielenie zbioru walidacyjnego
  - Skąd wiemy jakie hiperparametry ustawić?
  - Wewnętrzna pętla walidacji żeby ustalić wartości hiperparametrów
  - Dwa poziomy walidacji (zewnętrzna – zbiór testowy, wewnętrzna – zbiór walidacyjny) – niezależna weryfikacja wydajności uwzględniająca niepewność doboru hiperparametrów
    - Testowanie metody vs testowanie metody z hiperparametrami dobranymi przez eksperta

Klasyfikator, skuteczność 89% . . . oczekiwania? Spodziewany wynik „w działaniu", najlepsze możliwe oszacowanie, prawdziwa skuteczność klasyfikacji, powtarzalność, uogólnianie. . .

- Zbiór danych – reprezentatywny dla problemu
- Wydzielenie zbioru testowego:
  - Brak – oszacowanie obciążone (skuteczność zawyżona)
  - 50% trening, 50% test – oszacowanie obciążone (skuteczność zaniżona)
  - Wydzielenie pojedynczego przykładu (leave one out) + powtarzanie – duża wariancja, złożone obliczeniowo
    - Ale – leave one patient out!
  - Wydzielenie części (np. 5, 10, 20%) i powtórzenie – walidacja krzyżowa, optymalne
  - Wydzielenie zbioru testowego – niezależna weryfikacja wydajności
- Wydzielenie zbioru walidacyjnego
  - Skąd wiemy jakie hiperparametry ustawić?
  - Wewnętrzna pętla walidacji żeby ustalić wartości hiperparametrów
  - Dwa poziomy walidacji (zewnętrzna – zbiór testowy, wewnętrzna – zbiór walidacyjny) – niezależna weryfikacja wydajności uwzględniająca niepewność doboru hiperparametrów
    - Testowanie metody vs testowanie metody z hiperparametrami dobranymi przez eksperta

Klasyfikator, skuteczność 89% . . . oczekiwania? Spodziewany wynik „w działaniu", najlepsze możliwe oszacowanie, prawdziwa skuteczność klasyfikacji, powtarzalność, uogólnianie. . .

- Zbiór danych – reprezentatywny dla problemu
- Wydzielenie zbioru testowego:
  - Brak – oszacowanie obciążone (skuteczność zawyżona)
  - 50% trening, 50% test – oszacowanie obciążone (skuteczność zaniżona)
  - Wydzielenie pojedynczego przykładu (leave one out) + powtarzanie – duża wariancja, złożone obliczeniowo
    - Ale – leave one patient out!
  - Wydzielenie części (np. 5, 10, 20%) i powtórzenie – walidacja krzyżowa, optymalne
  - Wydzielenie zbioru testowego – niezależna weryfikacja wydajności
- Wydzielenie zbioru walidacyjnego
  - Skąd wiemy jakie hiperparametry ustawić?
  - Wewnętrzna pętla walidacji żeby ustalić wartości hiperparametrów
  - Dwa poziomy walidacji (zewnętrzna – zbiór testowy, wewnętrzna – zbiór walidacyjny) – niezależna weryfikacja wydajności uwzględniająca niepewność doboru hiperparametrów
    - Testowanie metody vs testowanie metody z hiperparametrami dobranymi przez eksperta

Klasyfikator, skuteczność 89% . . . oczekiwania? Spodziewany wynik „w działaniu", najlepsze możliwe oszacowanie, prawdziwa skuteczność klasyfikacji, powtarzalność, uogólnianie. . .

- Zbiór danych – reprezentatywny dla problemu
- Wydzielenie zbioru testowego:
  - Brak – oszacowanie obciążone (skuteczność zawyżona)
  - 50% trening, 50% test – oszacowanie obciążone (skuteczność zaniżona)
  - Wydzielenie pojedynczego przykładu (leave one out) + powtarzanie – duża wariancja, złożone obliczeniowo
    - Ale – leave one patient out!
  - Wydzielenie części (np. 5, 10, 20%) i powtórzenie – walidacja krzyżowa, optymalne
  - Wydzielenie zbioru testowego – niezależna weryfikacja wydajności
- Wydzielenie zbioru walidacyjnego
  - Skąd wiemy jakie hiperparametry ustawić?
  - Wewnętrzna pętla walidacji żeby ustalić wartości hiperparametrów
  - Dwa poziomy walidacji (zewnętrzna – zbiór testowy, wewnętrzna – zbiór walidacyjny) – niezależna weryfikacja wydajności uwzględniająca niepewność doboru hiperparametrów
    - Testowanie metody vs testowanie metody z hiperparametrami dobranymi przez eksperta

Klasyfikator, skuteczność 89% . . . oczekiwania? Spodziewany wynik „w działaniu", najlepsze możliwe oszacowanie, prawdziwa skuteczność klasyfikacji, powtarzalność, uogólnianie. . .

- Zbiór danych – reprezentatywny dla problemu
- Wydzielenie zbioru testowego:
  - Brak – oszacowanie obciążone (skuteczność zawyżona)
  - 50% trening, 50% test – oszacowanie obciążone (skuteczność zaniżona)
  - Wydzielenie pojedynczego przykładu (leave one out) + powtarzanie – duża wariancja, złożone obliczeniowo
    - Ale – leave one patient out!
  - Wydzielenie części (np. 5, 10, 20%) i powtórzenie – walidacja krzyżowa, optymalne
  - Wydzielenie zbioru testowego – niezależna weryfikacja wydajności
- Wydzielenie zbioru walidacyjnego
  - Skąd wiemy jakie hiperparametry ustawić?
  - Wewnętrzna pętla walidacji żeby ustalić wartości hiperparametrów
  - Dwa poziomy walidacji (zewnętrzna – zbiór testowy, wewnętrzna – zbiór walidacyjny) – niezależna weryfikacja wydajności uwzględniająca niepewność doboru hiperparametrów
    - Testowanie metody vs testowanie metody z hiperparametrami dobranymi przez eksperta

Analiza jakościowa (danych i wyników)

Klasyfikator, 89% zweryfikowana walidacją krzyżową, dobór parametrów OK – co dalej?
Analiza ilościowa – miary skuteczności
Analiza jakościowa – przegląd wyników pod kątem outlierów, regularności, wzorców itp.

- Szukaj outlierów i anomalii
- Zobacz jak się grupują przykłady (np. w obrębie klas)
- Stosuj redukcję wymiarowości (dla wizualizacji)
- Próbuj prostych metod
- Wizualizuj co tylko się da

- Powodzenia :)

Klasyfikator, 89% zweryfikowana walidacją krzyżową, dobór parametrów OK – co dalej?

Analiza ilościowa – miary skuteczności

Analiza jakościowa – przegląd wyników pod kątem outlierów, regularności, wzorców itp.

- Szukaj outlierów i anomalii
- Zobacz jak się grupują przykłady (np. w obrębie klas)
- Stosuj redukcję wymiarowości (dla wizualizacji)
- Próbuj prostych metod
- Wizualizuj co tylko się da

- Powodzenia :)

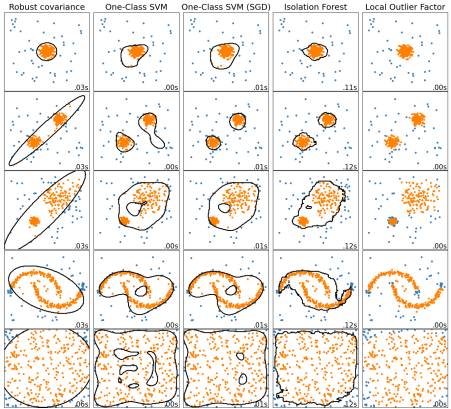
Klasyfikator, 89% zweryfikowana walidacją krzyżową, dobór parametrów OK – co dalej?
Analiza ilościowa – miary skuteczności
Analiza jakościowa – przegląd wyników pod kątem outlierów, regularności, wzorców itp.

- Szukaj outlierów i anomalii
- Zobacz jak się grupują przykłady (np. w obrębie klas)
- Stosuj redukcję wymiarowości (dla wizualizacji)
- Próbuj prostych metod
- Wizualizuj co tylko się da

- Powodzenia :)

Klasyfikator, 89% zweryfikowana walidacją krzyżową, dobór parametrów OK – co dalej?
Analiza ilościowa – miary skuteczności
Analiza jakościowa – przegląd wyników pod kątem outlierów, regularności, wzorców itp.

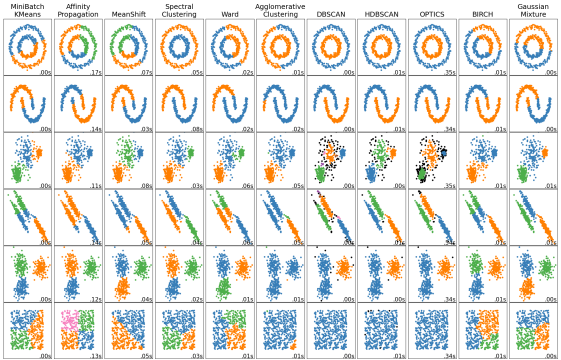
- Szukaj outlierów i anomalii



dokumentacja sklearn, BSD

Klasyfikator, 89% zweryfikowana walidacją krzyżową, dobór parametrów OK – co dalej?
Analiza ilościowa – miary skuteczności
Analiza jakościowa – przegląd wyników pod kątem outlierów, regularności, wzorców itp.

- Szukaj outlierów i anomalii
- Zobacz jak się grupują przykłady (np. w obrębie klas)



dokumentacja sklearn, BSD

- Stosuj redukcje wymiarowości (dla wizualizacji)

Klasyfikator, 89% zweryfikowana walidacją krzyżową, dobór parametrów OK – co dalej?
Analiza ilościowa – miary skuteczności
Analiza jakościowa – przegląd wyników pod kątem outlierów, regularności, wzorców itp.

- Szukaj outlierów i anomalii
- Zobacz jak się grupują przykłady (np. w obrębie klas)
- Stosuj redukcję wymiarowości (dla wizualizacji)
  - PCA
  - tSNE
  - kernel PCA, ICA, NMF, ...
- Próbuj prostych metod
- Wizualizuj co tylko się da
- Powodzenia :)

Klasyfikator, 89% zweryfikowana walidacją krzyżową, dobór parametrów OK – co dalej?
Analiza ilościowa – miary skuteczności
Analiza jakościowa – przegląd wyników pod kątem outlierów, regularności, wzorców itp.

- Szukaj outlierów i anomalii
- Zobacz jak się grupują przykłady (np. w obrębie klas)
- Stosuj redukcję wymiarowości (dla wizualizacji)
- Próbuj prostych metod
  - SVM, RF, MLP
  - kNN, 1-NN
  - Regresja liniowa
- Wizualizuj co tylko się da
- Powodzenia :)

Klasyfikator, 89% zweryfikowana walidacją krzyżową, dobór parametrów OK – co dalej?
Analiza ilościowa – miary skuteczności
Analiza jakościowa – przegląd wyników pod kątem outlierów, regularności, wzorców itp.

- Szukaj outlierów i anomalii
- Zobacz jak się grupują przykłady (np. w obrębie klas)
- Stosuj redukcję wymiarowości (dla wizualizacji)
- Próbuj prostych metod
- Wizualizuj co tylko się da
  - Histogramy/wykresy pudełkowe błędów
  - Przykłady dobrze i źle sklasyfikowane
  - Test wizualnej identyfikacji wzorców
  - Zależność od zewnętrznych parametrów

- Powodzenia :)

Klasyfikator, 89% zweryfikowana walidacją krzyżową, dobór parametrów OK – co dalej?
Analiza ilościowa – miary skuteczności
Analiza jakościowa – przegląd wyników pod kątem outlierów, regularności, wzorców itp.

- Szukaj outlierów i anomalii
- Zobacz jak się grupują przykłady (np. w obrębie klas)
- Stosuj redukcję wymiarowości (dla wizualizacji)
- Próbuj prostych metod
- Wizualizuj co tylko się da

- Powodzenia :)