

# Zaawansowane uczenie maszynowe: *wykład 4*

Paweł Cichosz

- 1 Podstawowe algorytmy uczenia się (część 3)
  - Algorytm SVM
- 2 Ocena jakości modeli

## Algorytm SVM: margines klasyfikacji

Margines geometryczny dla przykładu  $x$ : odległość od granicy decyzyjnej (nieujemna jeśli klasyfikowany poprawnie, ujemna jeśli klasyfikowany niepoprawnie):

$$\gamma_{\mathbf{w}}(x) = c_-(x) \frac{\mathbf{w} \circ \mathbf{a}(x)}{\|\mathbf{w}_{1:n}\|}$$

gdzie  $c_-(x) = 2c(x) - 1$ .

Margines funkcyjny dla przykładu  $x$ :

$$\hat{\gamma}_{\mathbf{w}}(x) = c_-(x) \mathbf{w} \circ \mathbf{a}(x)$$

Margines geometryczny/funkcyjny dla zbioru trenującego:

$$\gamma_{\mathbf{w}}(T) = \min_{x \in T} \gamma_{\mathbf{w}}(x)$$

$$\hat{\gamma}_{\mathbf{w}}(T) = \min_{x \in T} \hat{\gamma}_{\mathbf{w}}(x)$$

## Algorytm SVM: twardy margines

**Założenie:** liniowo separowalny zbiór trenujący, poszukiwana granica decyzyjna poprawnie separująca wszystkie przykłady.

**Postać kanoniczna wektora parametrów:** można ograniczyć się do wektorów parametrów spełniających warunek  $\hat{\gamma}_{\mathbf{w}}(T) = 1$  (granica decyzyjna niewrażliwa na skalowanie  $\mathbf{w}$ ).

**Maksymalizacja marginesu geometrycznego:** równoważna maksymalizacji  $\frac{1}{\|\mathbf{w}_{1:n}\|}$ , co z kolei jest równoważne minimalizacji  $\|\mathbf{w}_{1:n}\|^2$ .

**Zadanie optymalizacji:**

minimalizacja:

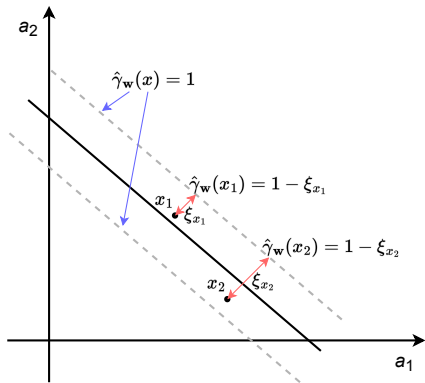
$$\frac{1}{2} \|\mathbf{w}_{1:n}\|^2$$

przy ograniczeniach:

$$(\forall x \in T) \quad c_-(x)\mathbf{w} \circ \mathbf{a}(x) \geq 1$$

**Wektory podpierające:** przykłady, dla których  $c_-(x)\mathbf{w} \circ \mathbf{a}(x) = 1$  („leżące na marginesie”).

# Algorytm SVM: miękki margines



**Założenie:** zbiór trenujący niekoniecznie liniowo separowalny, poszukiwana granica decyzyjna nie musi poprawnie separować wszystkich przykładów.

**Zmienne luzujące:**  $\xi_x$  – wielkość naruszenia ograniczenia dla przykładu  $x$ .

## Algorytm SVM: miękki margines

Zadanie optymalizacji:

minimalizacja:

$$\frac{1}{2} \|\mathbf{w}_{1:n}\|^2 + C \sum_x \xi_x$$

przy ograniczeniach:

$$(\forall x \in T) \quad c_-(x) \mathbf{w} \circ \mathbf{a}(x) \geq 1 - \xi_x$$

$$(\forall x \in T) \quad \xi_x \geq 0$$

„Zawiasowa” funkcja straty (*hinge loss*):

$$\mathcal{L}(c(x), h(x)) = \max\{0, 1 - c_-(x) \mathbf{w} \circ \mathbf{a}(x)\}$$

**Koszt:**  $C$  – koszt naruszenia ograniczenia (stosowane oznaczenie  $C$  zamiast  $C$  dla odróżnienia kosztu od zbioru klas).

**Wektory podpierające:** przykłady, dla których  $c_-(x) \mathbf{w} \circ \mathbf{a}(x) \leq 1$  („leżące na marginesie” jeśli  $c_-(x) \mathbf{w} \circ \mathbf{a}(x) = 1$ ).

## Algorytm SVM: właściwości

- Należy do najbardziej skutecznych i często używanych algorytmów dla zadań klasyfikacji:
  - maksymalizacja marginesu zmniejsza efektywny wymiar VC przestrzeni modeli (VC odwrotnie proporcjonalny do kwadratu marginesu),
  - zwiększona odporność na nadmierne dopasowanie nawet przy dużej liczbie atrybutów,
  - możliwość wyznaczenia modelu dla danych nieseparowalnych liniowo.
- Wymaga staranności przy doborze kosztu naruszenia ograniczeń.
- Brak możliwości bezpośredniej interpretacji modeli.
- Możliwość predykcji prawdopodobieństw klas przez zastosowanie tzw. skalowania Platt'a (o czym będzie mowa na jednym z kolejnych wykładów).
- Możliwość reprezentacji nieliniowych zależności przez zastosowanie funkcji jądrowych (o czym będzie mowa na jednym z kolejnych wykładów).

- 1 Podstawowe algorytmy uczenia się (część 3)
  - Algorytm SVM
- 2 Ocena jakości modeli



# Macierz pomyłek

Macierz pomyłek na zbiorze  $S$ :

$$CM_{S,c}(h)[d_1, d_2] = |S_{c=d_1, h=d_2}|$$

Przypadek dwuklasowy:

	$h$	
$c$	0	1
0	TN	FP
1	FN	TP

## Miary jakości klasyfikacji

współczynnik prawdziwych pozytywnych (true positive rate):

$$\frac{TP}{TP + FN}$$

współczynnik fałszywych pozytywnych (false positive rate):

$$\frac{FP}{TN + FP}$$

odzysk (recall): = współczynnik prawdziwych pozytywnych

precyzja (precision):

$$\frac{TP}{TP + FP}$$

# Miary jakości

miara  $F$ : średnia harmoniczna precyzji i odzysku:

$$F = \frac{1}{\frac{\frac{1}{recall} + \frac{1}{precision}}{2}} = \frac{2 \cdot recall \cdot precision}{recall + precision}$$

czułość (sensitivity): = współczynnik prawdziwych pozytywnych

specyficzność (specificity): =  $1 -$  współczynnik fałszywych pozytywnych

## Analiza ROC

**Graficzna metoda oceny jakości klasyfikacji:** wizualizacja punktów pracy modeli klasyfikacji za pomocą punktów i krzywych w układzie współrzędnych (TP rate $[y]$ , FP rate $[x]$ ).

**Predykcja klas:** pojedynczy punkt.

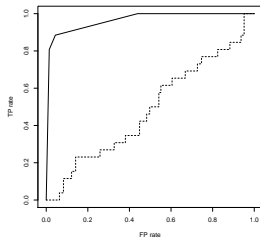
**Predykcja prawdopodobieństw lub funkcji decyzyjnej:** krzywa (łamana) łącząca punkty odpowiadające różnym progom odcięcia  $P(1|x)$  lub wartości funkcji decyzyjnej.

**Liczba punktów pracy:** powiększona o 1 liczba różnych wartości  $P(1|x)$  lub wartości funkcji decyzyjnej.

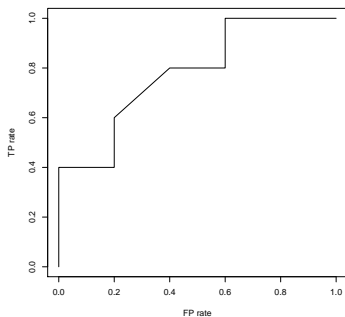
**Zagregowana ocena predykcji probabilistycznych:** pole pod krzywą ROC (AUC) – interpretowane jako prawdopodobieństwo, że losowo wybrany przykład klasy 1 uzyska większe  $P(1|x)$  lub wartość funkcji decyzyjnej niż losowo wybrany przykład klasy 0.

**Sporządzanie wykresu:** sortowanie według  $P(1|x)$  lub wartości funkcji decyzyjnej, wyznaczanie zmiany TP i FP przy każdej zmianie progów odcięcia.

# Analiza ROC



$P(1 x)$	$c(x)$
0.1	0
0.1	0
0.3	1
0.4	0
0.5	0
0.5	1
0.6	1
0.7	0
0.9	1
0.9	1



**Domyślny punkt pracy:** odpowiadający progowi odcięcia domyślnie stosowanemu przy predykcji klas.

**Preferowany punkt pracy:** może być inny niż domyślny i osiągnany przez zmianę progu odcięcia.

**Wybór punktu pracy:** na podstawie wymagań zastosowania lub pewnej heurystyki, np. minimalna odległość od punktu (0,1).

## Ocena macierzy pomyłek dla klasyfikacji wieloklasowej

**Binaryzacja:** ocena jakości klasyfikacji po dekompozycji pierwotnego zadania wieloklasowego na zadania binarne:

„1 kontra reszta” (one vs. rest/all, OvR/OvA): każda klasa traktowana jako pozytywna, pozostałe klasy łącznie traktowane jako negatywna,

„1 kontra 1” (one vs. one, OvO): dla każdej pary klas jedna traktowana jako pozytywna, druga traktowana jako negatywna.

**Makro-uśrednianie:** wskaźniki jakości wyznaczane niezależnie dla każdego zadania binarnego i uśredniane.

**Mikro-uśrednianie:** predykcje i prawdziwe wartości dla każdego zadania binarnego scalane (sumowane wartości TP, FP, TN, FN) i na ich podstawie wyznaczane wskaźniki jakości.

- Dekompozycja OvR lub OvO może również służyć do tworzenia modeli klasyfikacji wieloklasowej z użyciem algorytmów klasyfikacji binarnej.

## Miary jakości regresji

Błąd bezwzględny (MAE, *mean absolute error*):

$$\text{MAE}_{S,f}(h) = \frac{1}{|S|} \sum_{x \in S} |f(x) - h(x)|$$

Błąd średniokwadratowy (MSE, *mean square error*):

$$\text{MSE}_{S,f}(h) = \frac{1}{|S|} \sum_{x \in S} (f(x) - h(x))^2$$

Pierwiastek błędu średniokwadratowego (RMSE, *root mean square error*):

$$\text{RMSE}_{S,f}(h) = \sqrt{\text{mse}_{S,f}(h)}$$

## Miary jakości regresji

Błąd względny (RAE, *relative absolute error*):

$$\text{RAE}_{S,f}(h) = \frac{\sum_{x \in S} |f(x) - h(x)|}{\sum_{x \in S} |f(x) - m_S(f)|} = \frac{\text{MAE}_{S,f}(h)}{\frac{1}{|S|} \sum_{x \in S} |f(x) - m_S(f)|}$$

gdzie  $m_S(f)$  – średnia wartość  $f$  na zbiorze  $S$ .

Współczynnik determinacji:

$$R^2_{S,f}(h) = 1 - \frac{\sum_{x \in S} (f(x) - h(x))^2}{\sum_{x \in S} (f(x) - m_S(f))^2} = 1 - \frac{|S| \text{MSE}_{S,f}(h)}{(|S| - 1) s_S^2(f)}$$

gdzie  $s_S^2(f)$  – wariancja  $f$  na zbiorze  $S$ .

**Korelacja:** współczynnik korelacji liniowej lub rangowej między wartościami  $f$  i  $h$ .



## Procedury oceny

**Podział na 2 podzbiory (*holdout*):** losowy podział  $D$  na  $T$  (do budowy modelu) i  $Q$  (do oceny jakości modelu – tzw. zbiór testowy lub walidacyjny),  $T \cap Q = \emptyset$ .

**$k$ -krotna walidacja krzyżowa ( $k$ -CV):**

- losowy podział  $D$  na równoliczne parami rozłączne podzbiory  $D_1, D_2, \dots, D_k$  (z zachowaniem rozkładu klas),
  - tworzenie modelu  $h_i$  na podstawie  $T_i = \bigcup_{j \neq i} D_j$ ,
  - ocena modelu  $h_i$  na podstawie  $Q_i = D_i$ .
- Drugie spojrzenie na ocenę jakości modeli za jakiś czas...