Uczenie maszynowe:
wykład 11

Paweł Cichosz

Macierz pomyłek

Macierz pomyłek na zbiorze S:

$$\mathrm{CM}_{S,c}(h)[d_1, d_2] = |S_{c=d_1, h=d_2}|$$

Przypadek dwuklasowy:

	h	
c	0	1
0	TN	FP
1	FN	TP

Miary jakości klasyfikacji

błąd:

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

dokładność:

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

współczynnik prawdziwych pozytywnych (true positive rate):

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

współczynnik fałszywych pozytywnych (false positive rate):

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

Miary jakości

odzysk (recall): = współczynnik prawdziwych pozytywnych

precyzja (precision):

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

miara F (F-measure): 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.

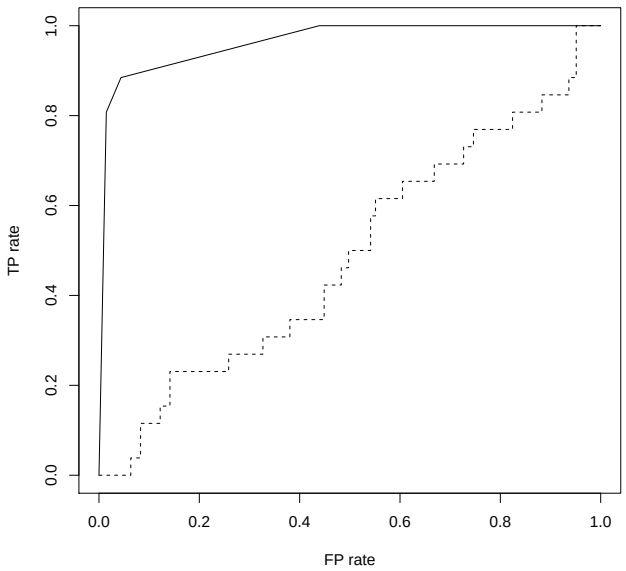
Predykcja prawdopodobieństw (lub funkcji decyzyjnej): krzywa (łamana) łącząca punkty odpowiadające różnym progom odcięcia $P(1|x)$ (lub wartości funkcji decyzyjnej).

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

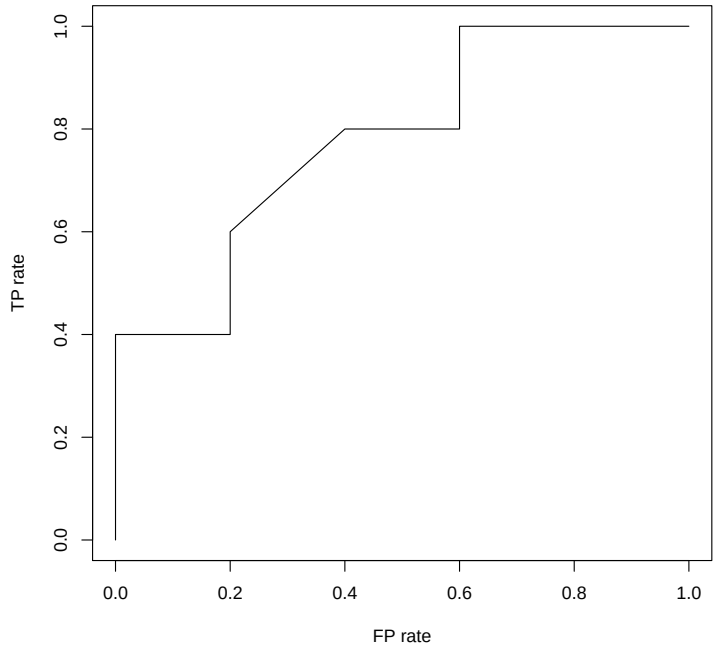
Zagregowana ocena: pole pod krzywą ROC (AUC) – interpretowane jako prawdopodobieństwo, że losowo wybrany przykład klasy 1 uzyska większe $P(1|x)$ (lub wartość funkcji decyzyjnej) niż losowo wybrany przykład klasy 0.

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

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



Ocena macierzy pomyłek dla klasyfikacji wieloklasowej

Binaryzacja: ocena jakości klasyfikacji „1 kontra reszta” (one vs. rest, OvR) albo „1 kontra 1” (one vs. one, OvO).

Makro-uśrednianie: wskaźniki jakości wyznaczane niezależnie dla każdego zadania binarnego i uśredniane (jakość predykcji dla każdej klasy lub pary klas jednakowo ważna).

Mikro-uśrednianie: predykcje i prawdziwe wartości dla każdego zadania binarnego scalane i na ich podstawie wyznaczane wskaźniki jakości (jakość predykcji dla każdej klasy lub pary klas ważna w stopniu proporcjonalnym do jej częstości).

Dodatkowe miary jakości regresji

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

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

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

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

Dodatkowe miary jakości regresji

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

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

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

Współczynnik determinacji:

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

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

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

Procedury oceny

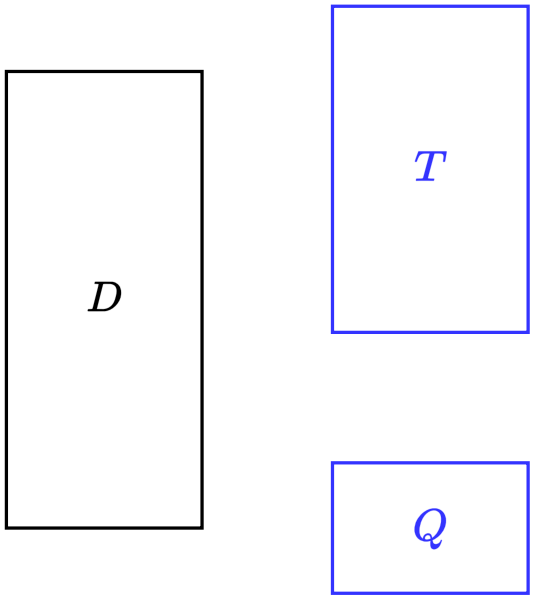
Cel oceny: dostarczenie estymacji prawdziwych wartości wskaźników jakości modelu na nowych danych, do których będzie stosowany w trakcie eksploatacji.

Ocena procedury modelowania: zamiast oceniać konkretny model lepiej ocenić sposób jego tworzenia.

Model oceniany a model użyty: oceniać możemy tylko model budowany na części danych, ale do faktycznego użycia można przekazać inny model – tworzony w taki sam sposób, ale na całości dostępnych danych.

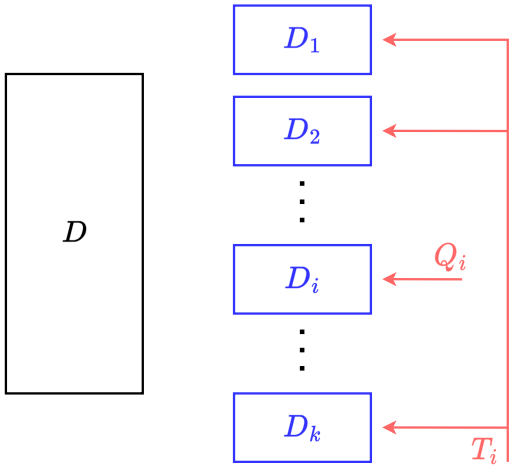
Obciążenie kontra wariancja: zbyt mały zbiór do oceny nie zapewnia wiarygodnej estymacji jakości ze względu na wysoką wariancję, ale odłożenie dużej części danych do oceny wprowadza pesymistyczne obciążenie związane z obniżeniem jakości ocenianego modelu.

Podział na 2 podzbiory (holdout)



- Losowy podział D na T (do budowy modelu) i Q (do oceny jakości modelu – tzw. zbiór testowy lub walidacyjny), $T \cap Q = \emptyset$.
- Słabo rozwiązuje konflikt między obciążeniem i wariancją.
- Redukcja wariancji możliwa przez wielokrotne powtarzanie.

k-krotna walidacja krzyżowa (k-CV)



- Losowy podział D na równoliczne parami rozłączne podzbiory $D_1, D_2, \ldots, D_k$ (z zachowaniem rozkładu klas).
- Dla $i = 1, 2, \ldots, k$:
  - tworzenie modelu h_i na podstawie $T_i = \bigcup_{j \neq i} D_j$,
  - ocena modelu h_i na podstawie $Q_i = D_i$.
- Najczęściej $k \in \{5, 10\}$.
- Małe obciążenie i wariancja.
- Dalsza redukcja wariancji możliwa przez wielokrotne powtarzanie i uśrednianie bądź scalanie wyników: $n \times k$-CV.
- Przypadek specjalny: $k = |D|$ (*leave-one-out*).

k-krotna walidacja krzyżowa (k-CV)

- Estymacja wskaźnika jakości w jednym z dwóch trybów:

  makro-uśrednianie: średnia wartość wskaźników jakości uzyskanych przy ocenie modeli h_i na zbiorach Q_i dla $i = 1, 2, \ldots, k$ – najczęściej stosowane, umożliwia także estymację wariancji wskaźnika jakości (jako wariancji średniej),

  mikro-uśrednianie (scalanie): wartość wskaźnika jakości obliczonego na zbiorze D po scaleniu predykcji poszczególnych modeli h_i na zbiorach Q_i dla $i = 1, 2, \ldots, k$ w jeden wektor przewidywanych klas, porównywanych następnie z prawdziwymi klasami (rzadziej stosowane).