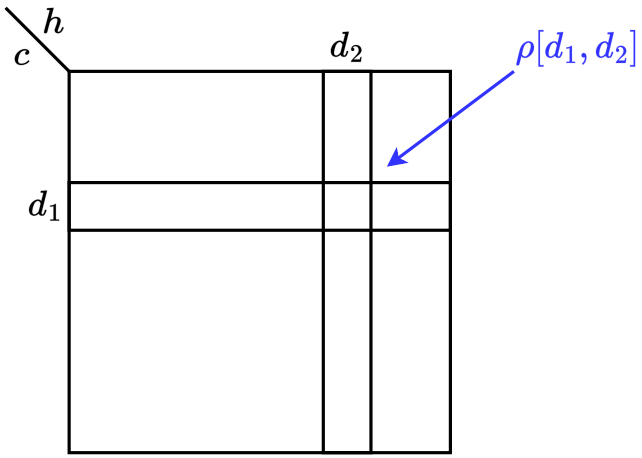
Zaawansowane uczenie maszynowe:
wykład 10

Paweł Cichosz

Reprezentacja kosztów



Macierz kosztów: $\rho[d_1, d_2]$ – koszt predykcji $h(x) = d_2$ gdy $c(x) = d_1$.

- $\rho[d_1, d_2] \geq 0$ dla dowolnych $d_1, d_2 \in C$,
- $\rho[d, d] = 0$ dla dowolnego $d \in C$.

Wektor kosztów: $\rho[d]$ – koszt niepoprawnej predykcji gdy $c(x) = d$ (uproszczona reprezentacja kosztów).

„Spłaszczenie" kosztów:

$$\rho[d] = \frac{1}{|C| - 1} \sum_{d' \in C} \rho[d, d']$$

Reprezentacja kosztów

Koszty na poziomie przykładów: dla każdego przykładu x osobno określone koszty pomyłek $\rho(x)[d_1, d_2]$ lub $\rho(x)[d]$ zależne od wartości atrybutów tego przykładu.

Źródło kosztów:

- obiektywne koszty niepoprawnych predykcji określone na podstawie wiedzy o dziedzinie,
- subiektywnie określone koszty reprezentujące „ważność" klas lub kompensujące ich niezrównoważenie.

Niezrównoważone klasy

Problem: minimalizacja błędu może dawać model „zaniedbujący” klasę lub klasy mniejszościowe.

Uwrażliwianie na klasy mniejszościowe: analogicznie jak uwrażliwianie na klasy o bardziej kosztownych pomyłkach przy tworzeniu modelu.

Koszty równoważące klasy: koszt pomyłek $\rho[d]$ odwrotnie proporcjonalny do częstości występowania klasy d w zbiorze trenującym.

Uwzględnianie kosztów przy predykcji

Oczekiwany koszt pomyłki przy predykcji $h(x) = d$:

$$\sum_{d' \in C} \rho[d', d] P(d'|x)$$

Reguła minimalnego kosztu:

$$h(x) = \arg\min_{d \in C} \sum_{d' \in C} \rho[d', d] P(d'|x)$$

Przypadek dwuklasowy: predykcja klasy 1 wtedy i tylko wtedy, gdy

$$P(0|x)\rho[0] \leq P(1|x)\rho[1]$$

Próg predykcji 1:

$$P(1|x) \geq \ldots$$

Kalibracja prawdopodobieństw: możliwość poprawienia jakości predykcji probabilistycznych (por. wykład 6).

Uwzględnianie kosztów przy tworzeniu modelu

Ważenie: waga przykładu x proporcjonalna do kosztu $\rho[c(x)]$ – minimalizacja ważonego błędu klasyfikacji, który jest proporcjonalny do średniego kosztu pomyłek:

$$\begin{aligned} e_{T,c,w}(h) &= \frac{\sum_{x \in T_{h \neq c}} w_x}{\sum_{x \in T} w_x} \\ &= \frac{|T|}{\sum_{x \in T} w_x} \frac{\sum_{x \in T_{h \neq c}} \rho[c(x)]}{|T|} \\ &= \frac{|T|}{\sum_{x \in T} w_x} \frac{\sum_{x \in T} \rho[c(x), h(x)]}{|T|} \end{aligned}$$

przyjmując, że $w_x = \rho[c(x)]$ oraz

$$\rho[d_1, d_2] = \begin{cases} 0 & \text{jeśli } d_1 = d_2 \\ \rho[d_1] & \text{w przeciwnym przypadku} \end{cases}$$

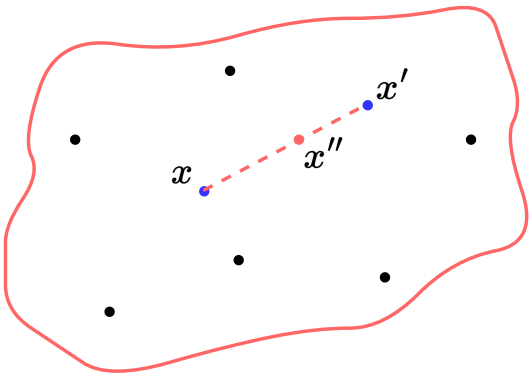
Uwzględnianie kosztów przy tworzeniu modelu

Prawdopodobieństwa *a priori*: prawdopodobieństwo *a priori* klasy d proporcjonalne do kosztu $\rho[d]$ (zamiast prawdopodobieństwa opartego na częstości klasy d w zbiorze trenującym).

Replikacja/próbkowanie: liczba egzemplarzy lub prawdopodobieństwo przykładu x proporcjonalne do kosztu $\rho[c(x)]$, np.:

- replikacja przykładów klas o bardziej kosztownych pomyłkach,
- próbkowanie przykładów klas o mniej kosztownych pomyłkach,
- losowanie próby ze zwracaniem ze zbioru trenującego z prawdopodobieństwami proporcjonalnymi do kosztów pomyłek.

Uwzględnianie kosztów przy tworzeniu modelu



Generowanie sztucznych przykładów: SMOTE (*synthetic minority oversampling technique*)

- dla każdego przykładu trenującego klasy d (mniejszościowej/o większym koszcie pomyłek) wyznaczany zbiór k najbliższych sąsiadów tej samej klasy z użyciem ustalonej miary niepodobieństwa,
- nowy przykład generowany jako losowo wybrany wektor wartości atrybutów leżący pomiędzy x i sąsiadem x',
- liczb generowanych przykładów proporcjonalna do kosztu $\rho[d]$.

Przykładowe realizacje

Wagi klas lub przykładów dla drzew decyzyjnych: zliczanie przykładów w celu wyznaczania kryterium stopu, prawdopodobieństw klas w liściach i kryterium wyboru podziału zastępowane przez sumowanie wag przykładów:

$$|S| \quad \rightsquigarrow \quad \sum_{x \in S} w_x$$

dla dowolnego $S \subseteq T$.

Prawdopodobieństwa *a priori* dla drzew decyzyjnych: przy zliczaniu przykładów w celu wyznaczania kryterium stopu, prawdopodobieństw klas w liściach i kryterium wyboru podziału liczba przykładów każdej klasy mnożona przez iloraz jej prawdopodobieństwa *a priori* i jej częstości występowania w zbiorze trenującym:

$$|S_{c=d}| \quad \rightsquigarrow \quad |S_{c=d}| \frac{P(c=d)}{|T_{c=d}|/|T|}$$

dla dowolnego $S \subseteq T$.

Przykładowe realizacje

Wagi klas dla SVM: wartości zmiennych luzujących uwzględniane przy wyznaczaniu kosztu naruszenia ograniczeń mnożone przez wagi klas:

$$C \sum_{x \in T} \xi_x \quad \rightsquigarrow \quad C \sum_{x \in T} w_{c(x)} \xi_x$$

Próbkowanie dla lasu losowego:

- próby bootstrapowe losowane warstwowo,
- prawdopodobieństwo wylosowania/wielkość próby dla klasy d proporcjonalne do $\rho[d]$.

1. Koszty pomyłek i niezrównoważone klasy

2. Drugie spojrzenie na ocenę jakości modeli
   - Dodatkowe miary jakości
   - Procedury oceny
   - Kryteria informacyjne jakości modeli

Dodatkowe miary jakości

Równoważona dokładność (*balanced accuracy*, BAC):

$$\frac{1}{2}\left(\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}+\frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}\right)$$

– średnia arytmetyczna dokładności dla obu klas (możliwe także uogólnienie dla klasyfikacji wieloklasowej).

Korelacja Matthewsa (*Matthews correlation coefficient*, MCC):

$$\frac{\mathrm{TP}\cdot\mathrm{TN}-\mathrm{FP}\cdot\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}$$

– równoważny współczynnikowi korelacji liniowej między c i h (prawdziwymi i przewidywanymi klasami ze zbioru $\{0,1\}$).

Ocena macierzy pomyłek dla klasyfikacji wieloklasowej

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

„1 kontra reszta" (one vs. rest, OvR): 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.

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.

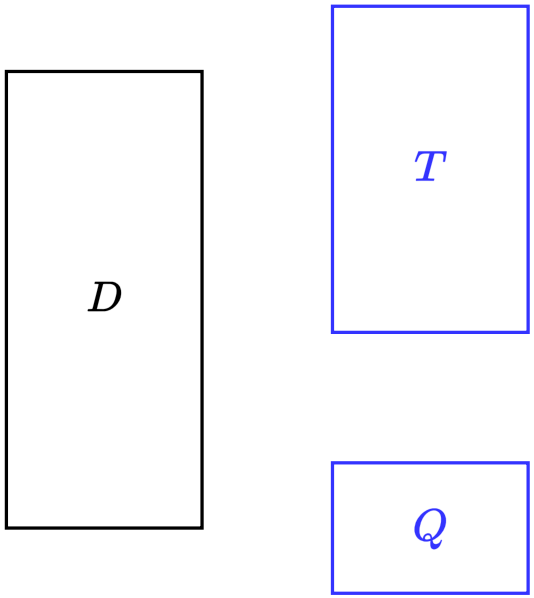
Procedury oparte na próbkowaniu: różne metody wyznaczania podzbiorów trenujących i walidacyjnych/testowych.

Pułapki oceny modeli (1)

Obciążenie kontra wariancja:

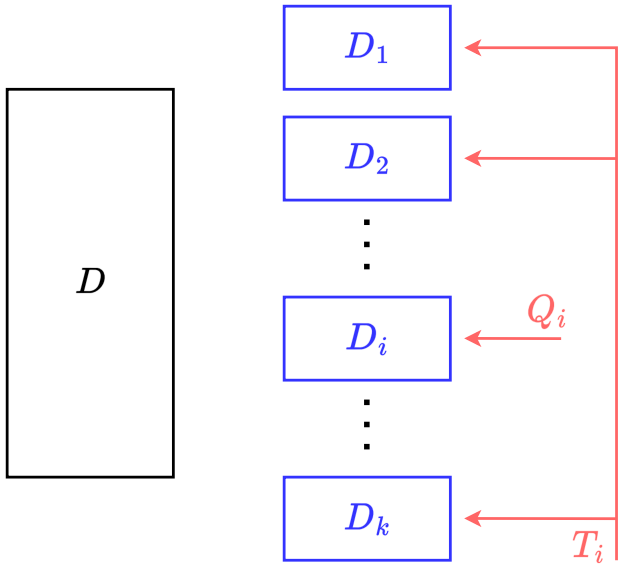
- zbyt mały zbiór do oceny nie zapewnia wiarygodnej estymacji jakości ze względu na wysoką wariancję,
- odłożenie dużej części danych do oceny wprowadza pesymistyczne obciążenie związane z obniżeniem jakości ocenianego modelu,
- procedury oceny realizują pewien kompromis między minimalizacją obciążenia i wariancji a nakładem obliczeń.

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).

Pułapki oceny modeli (2)

Ocena pośrednia kontra końcowa: jeśli pewien zbiór modeli jest oceniany w celu podjęcia decyzji wpływających na końcową postać modelu, dane wykorzystywane do tej oceny pośrednio stają się częścią danych trenujących *w szerszym sensie*, a więc wiarygodna ocena końcowego modelu wymaga osobnego podzbioru danych (zewnętrznej procedury oceny), np.:

- strojenie parametrów: dane użyte do oceny modeli w celu strojenia parametrów nie mogą być podstawą do wiarygodnej oceny jakości modelu zbudowanego z użyciem dobranych na tej podstawie ustawień parametrów,
- selekcja atrybutów: dane użyte do oceny modeli w celu selekcji atrybutów nie mogą być podstawą do wiarygodnej oceny jakości modelu zbudowanego z użyciem wybranych na tej podstawie atrybutów,
- transformacja atrybutów: dane użyte do oceny modeli w celu określenia transformacji atrybutów nie mogą być podstawą do wiarygodnej oceny jakości modelu zbudowanego z użyciem wybranych na tej podstawie transformacji.

Koncepcja kryteriów informacyjnych

Ocena jakości bez „nowych danych”: porównywanie modeli i selekcja najlepszego modelu wyłącznie na podstawie danych trenujących.

Ogólna zasada: ocena dopasowania do danych i złożoności modelu.

Ograniczenia: sensowne porównanie możliwe tylko dla modeli o takiej samej reprezentacji.

Typowe zastosowanie: selekcja najbardziej obiecującego modelu spośród modeli:

- utworzonych za pomocą tego samego algorytmu
- zastosowanego do tych samych danych
- ale różniących się np. ustawieniami parametrów algorytmu, zestawem używanych atrybutów, zastosowanymi transformacjami atrybutów itp.

bez konieczności stosowania procedury oceny dzielącej dane na trenujące i testowe/walidacyjne.

AIC (*Akaike Information Criterion*)

Postać ogólna: minimalizacja (dla modelu parametrycznego o k parametrach):

$$-2\mathrm{LL}_T(h) + 2k$$

gdzie $\mathrm{LL}_T(h)$ – logarytm wiarygodności (*loglikelihood*) modelu.

Właściwości: składnik kary nie zależy od wielkości zbioru danych – przy dużych zbiorach danych mogą być preferowane złożone lecz dobrze dopasowane modele.

Zastosowanie dla regresji logistycznej:

$$\mathrm{LL}_T(h) = \sum_{x \in T} \big(c(x) \ln \pi(x) + (1 - c(x)) \ln(1 - \pi(x))\big)$$

gdzie $\pi(x) = P(1|x)$ – probabilistyczne predykcje modelu.

Zastosowanie dla regresji liniowej:

$$\mathrm{LL}_T(h) = -\frac{1}{2}|T| \ln \mathrm{MSE}_{T,f}(h)$$

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

BIC (*Bayesian Information Criterion*)

Postać ogólna: minimalizacja:

$$-2\mathrm{LL}_T(h) + k \ln |T|$$

(dla modelu parametrycznego o k parametrach)

Właściwości: składnik kary zależy od wielkości zbioru danych – zwiększona preferencja dla prostszych modeli.

Zastosowanie: jak AIC.

MDL (*Minimum Description Length*)

Postać ogólna: minimalizacja długości kodu do reprezentacji modelu i danych trenujących przy znajomości modelu:

$$L(h) + L(T|h)$$

Kodowanie modelu: zależne od reprezentacji – kodowanie struktury i parametrów albo tylko parametrów (przy ustalonej strukturze).

Kodowanie danych przy znajomości modelu: wartości atrybutów wejściowych można pominąć (składnik niezależny od modelu), pozostaje kodowanie informacji, która wraz z modelem pozwala zrekonstruować wartości atrybutu docelowego – korekty predykcji modelu.