Zaawansowane uczenie maszynowe:
wykład 3

Paweł Cichosz

Naiwny klasyfikator bayesowski

Wzór Bayesa:

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)}$$

Zastosowanie do predykcji prawdopodobieństw klas:

$$P(d|x) = P(c = d \mid a_1 = a_1(x), a_2 = a_2(x), \ldots, a_n = a_n(x))$$

Prawdopodobieństwo klasy na podstawie wartości atrybutów:

$$\begin{aligned} &P(c = d \mid a_1 = v_1, \ldots, a_n = v_n) \\ &\quad = \frac{P(c = d)P(a_1 = v_1, \ldots, a_n = v_n \mid c = d)}{P(a_1 = v_1, \ldots, a_n = v_n)} \end{aligned}$$

Założenie o niezależności:

$$P(a_1 = v_1, \ldots, a_n = v_n \mid c = d) = \prod_{i=1}^{n} P(a_i = v_i|c = d)$$

Wymiar VC: liniowo zależny od n – ograniczone ryzyko nadmiernego dopasowania.

Naiwny klasyfikator bayesowski: estymacja prawdopodobieństw

Prawdopodobieństwo *a priori* klasy:

$$P(c = d) = P_T(c = d) = \frac{|T_{c=d}|}{|T|}$$

Prawdopodobieństwo warunkowe wartości atrybutu:

$$P(a_i = v_i | c = d) = P_{T_{c=d}}(a_i = v_i) = \frac{|T_{c=d,a_i=v_i}|}{|T_{c=d}|}$$

(zwykle stosowane wygładzanie Cestnika lub Laplace'a)

Mianownik: stała normalizująca:

$$P(a_1 = v_1, \ldots, a_n = v_n) = \sum_{d \in C} P(c = d) P(a_1 = v_1, \ldots, a_n = v_n \mid c = d)$$

Naiwny klasyfikator bayesowski: atrybuty ciągłe

Funkcja gęstości: zastąpienie $P(a_i = v_i|c = d)$ przez $g_{d,i}(v_i)$, gdzie $g_{d,i}$ jest funkcją gęstości atrybutu a_i w klasie d – najczęściej zakładany rozkład normalny, parametry estymowane jako:

średnia a_i w klasie d: $m_{T_{c=d}}(a_i)$,
odchylenie standardowe a_i w klasie d: $s_{T_{c=d}}(a_i)$.

Dyskretyzacja: często lepsze, chociaż bardziej pracochłonne podejście.

Naiwny klasyfikator bayesowski: brakujące wartości

Tworzenie modelu: pomijanie brakujących wartości przy estymacji prawdopodobieństw $P(a_i = v_i|c = d)$.

Predykcja: pomijanie prawdopodobieństw $P(a_i = v_i|c = d)$ jeśli wartość a_i nie jest znana dla klasyfikowanego przykładu.

Naiwny klasyfikator bayesowski: właściwości

- Prosty koncepcyjnie, implementacyjnie i obliczeniowo.
- Dość odporny na nadmierne dopasowanie.
- Niewymagający strojenia parametrów.
- Naruszenie założenia o niezależności nie wyklucza wartości predykcyjnej, chociaż predykcje prawdopodobieństw klas nie są skalibrowane.
- Skuteczny jeśli:
  - trzeba uwzględnić nieznaczny wpływ znacznej liczby atrybutów (np. klasyfikacja tekstu),
  - liczba przykładów jest stosunkowo mała w porównaniu z liczbą atrybutów,
  - występują liczne brakujące wartości.

Regresja liniowa

Funkcja reprezentacji:

$$h(x) = \sum_{i=1}^{n} w_i a_i(x) + w_{n+1} = \sum_{i=1}^{n+1} w_i a_i(x) = \mathbf{w} \circ \mathbf{a}(x)$$

gdzie $\mathbf{w}$ – wektor parametrów $w_1, w_2, \ldots, w_n, w_{n+1}$, $\mathbf{a}(x)$ – wektor wartości atrybutów $a_1(x), a_2(x), \ldots, a_n(x), a_{n+1}(x)$, przy czym $a_{n+1}(x) \equiv 1$, $\circ$ – symbol iloczynu skalarnego.

Minimalizacja straty kwadratowej:

$$E_{T,f}(h) = \frac{1}{2} \sum_{x \in T} (f(x) - h(x))^2$$

Gradientowa reguła aktualizacji parametrów:

$$\mathbf{w} := \mathbf{w} + \beta(f(x) - h(x))\mathbf{a}(x)$$

Metoda najmniejszych kwadratów

- Wyznaczanie parametrów za pomocą zamkniętej formuły, bez wielokrotnych aktualizacji.
- Postulat dopasowania do danych trenujących jako nadokreślony układ równań liniowych:

$$a_1(x)w_1 + a_2(x)w_2 + \cdots + a_n(x)w_n + a_{n+1}(x)w_{n+1} = f(x)$$

  dla każdego $x \in T$, przy założeniu $n+1 < |T|$ (zwykle $n \ll |T|$, często zaleca się np. $|T| \geq 5n$ lub $|T| \geq 10n$ itp.).
- W zapisie wektorowo-macierzowym:

$$\mathbf{a}(T)\mathbf{w} = \mathbf{f}(T)$$

  gdzie $\mathbf{a}(T)$ – macierz $|T| \times (n+1)$ wartości atrybutów $a_1(x), \ldots, a_{n+1}(x)$ dla $x \in T$, $\mathbf{f}(T)$ – wektor wartości $f(x)$ dla $x \in T$.
- Rozwiązanie przez pseudo-inwersję:

$$\begin{aligned} \mathbf{a}^\mathsf{T}(T)\mathbf{a}(T)\mathbf{w} &= \mathbf{a}^\mathsf{T}(T)\mathbf{f}(T) \\ \mathbf{w} &= \left(\mathbf{a}^\mathsf{T}(T)\mathbf{a}(T)\right)^{-1}\mathbf{a}^\mathsf{T}(T)\mathbf{f}(T) \end{aligned}$$

- Wymagana liniowa niezależność atrybutów aby macierz $\mathbf{a}^\mathsf{T}(T)\mathbf{a}(T)$ była nieosobliwa.

Klasyfikacja liniowo-progowa

Wewnętrzna reprezentacja liniowa (funkcja decyzyjna):

$$g(x) = \mathbf{w} \circ \mathbf{a}(x)$$

Zewnętrzna funkcja progowa:

$$h(x) = \begin{cases} 1 & \text{jeśli } g(x) \geq 0 \\ 0 & \text{w przeciwnym przypadku} \end{cases}$$

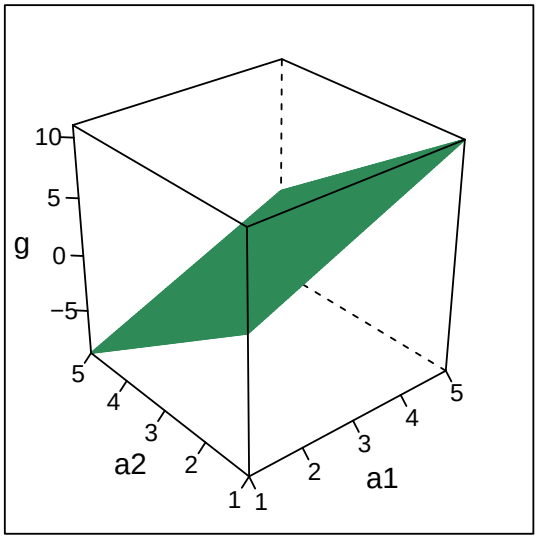
Reguła aktualizacji parametrów (*prosty perceptron*):

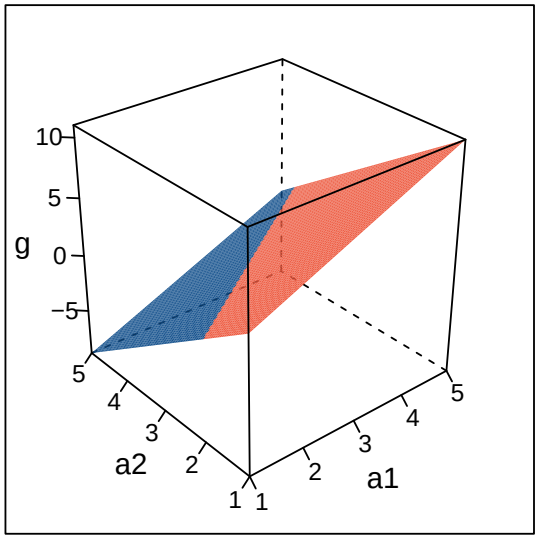
$$\mathbf{w} := \begin{cases} \mathbf{w} + c_{-}(x)\mathbf{a}(x) & \text{jeśli } h(x) \neq c(x) \\ \mathbf{w} & \text{w przeciwnym przypadku} \end{cases}$$

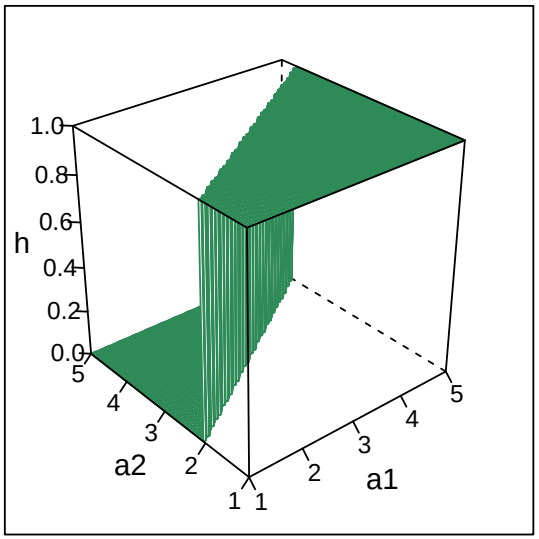
gdzie $c_{-}(x) = 2c(x) - 1$.

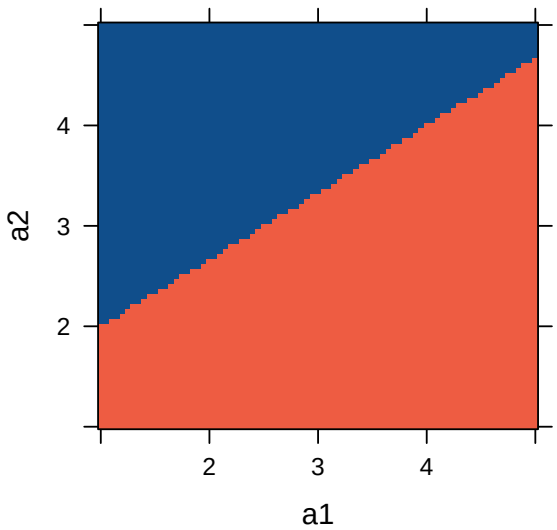
Wymiar VC: $n + 1$ – ograniczone ryzyko nadmiernego dopasowania.

Klasyfikacja liniowo-progowa









Regresja logistyczna

Wewnętrzna reprezentacja liniowa (funkcja decyzyjna):

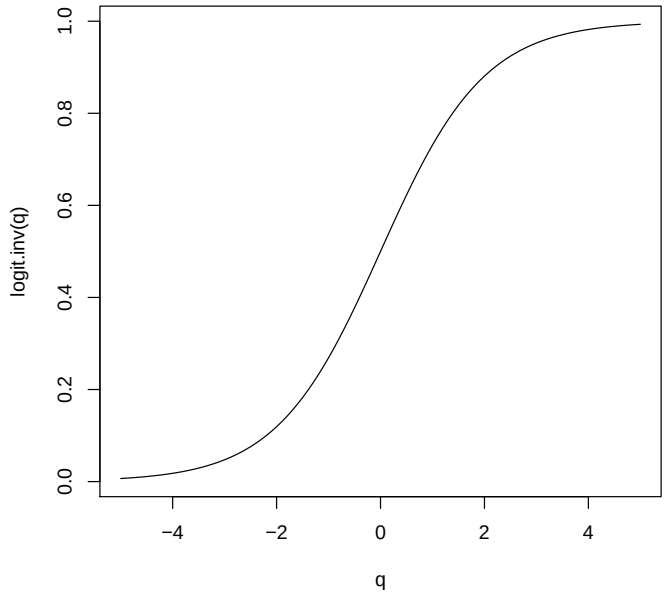
$$g(x) = \sum_{i=1}^{n} w_i a_i(x) + w_{n+1} = \mathbf{w} \circ \mathbf{a}(x)$$

Zewnętrzna logistyczna funkcja łącząca:

$$\text{logit}(p) = \ln \frac{p}{1-p}$$

$$\text{logit}^{-1}(q) = \frac{e^q}{e^q + 1} = \frac{1}{1 + e^{-q}}$$

$$\text{logit}^{-1}(g(x)) = \pi(x) = P(1|x)$$



Maksymalizacja logarytmu wiarygodności:

$$\text{LL}_{T,c}(\pi) = \ln P(T, c; \pi) = \ln \prod_{x \in T} P(c(x)|x; \pi)$$

Reguła aktualizacji parametrów: $\mathbf{w} := \mathbf{w} + \beta(c(x) - \pi(x))\mathbf{a}(x)$

Modele liniowe: właściwości

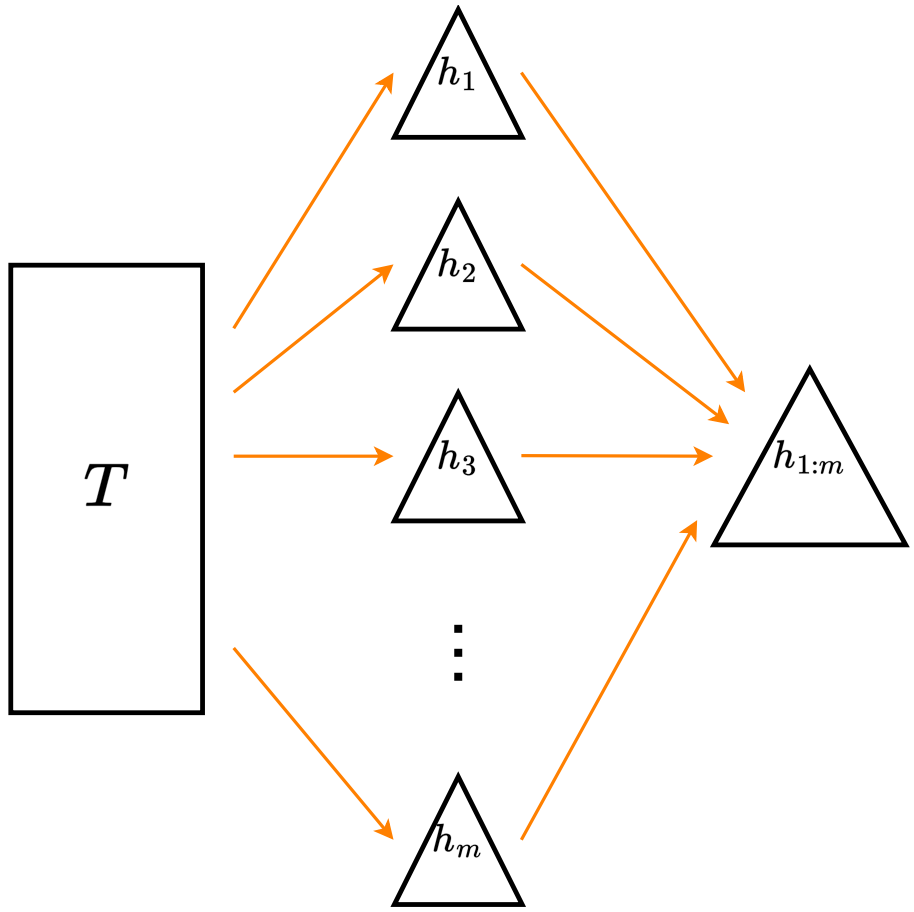
- Proste koncepcyjnie, implementacyjnie i obliczeniowo.
- Dość odporne na nadmierne dopasowanie.
- Niewymagające strojenia parametrów.
- Przydatność ograniczona przez liniową reprezentację, ale ograniczenie te można przezwyciężyć (o czym będzie mowa za kilka tygodni).

Koncepcja modelowania zespołowego

Motywacja: wzmocnienie siły predykcyjnej i kompensacja niedoskonałości przez łączenie modeli.

Modele bazowe: różne modele $h_1, h_2, \ldots, h_m$ dla tego samego zadania klasyfikacji lub regresji.

Łączenie predykcji: predykcja $h_{1:m}(x)$ na podstawie $h_1(x), h_2(x), \ldots, h_m(x)$.



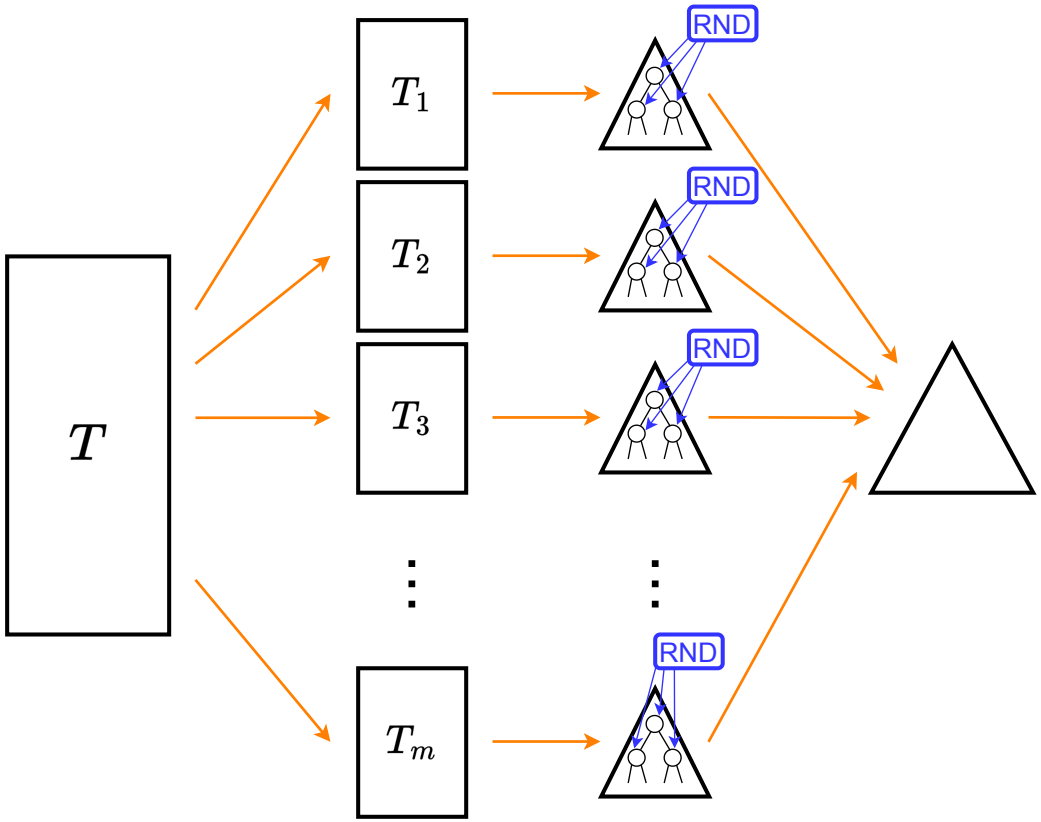
Las losowy

Modele bazowe: drzewa decyzyjne/drzewa regresji tworzone na podstawie prób *bootstrapowych* przez algorytm randomizowany w celu zwiększenia zróżnicowania:

- **wybór podziału:** spośród podziałów opartych na podzbiorze atrybutów niezależnie losowanym w każdym węźle (zazwyczaj losuje się $\lfloor \sqrt{n} \rfloor$ spośród wszystkich n atrybutów).
- **rozbudowane drzewa:** późno działające kryteria stopu, brak przycinania (więcej węzłów – więcej okazji do zróżnicowania).

Łączenie predykcji: zwykłe głosowanie/uśrednianie.

Las losowy



Właściwości: zazwyczaj bardzo wysoka jakość predykcji przy niewielkim wysiłku, odporność na nadmierne dopasowanie, ograniczona wrażliwość na ustawienia parametrów, zwykle co najmniej kilkaset modeli bazowych.