Zaawansowane uczenie maszynowe:
wykład 6

Paweł Cichosz

Rozkład wielomianowy

Rozkład dwumianowy: prawdopodobieństwo liczby sukcesów w serii prób Bernoulliego:

$$P(k) = \frac{N!}{k!(N-k)!} p^k (1-p)^{N-k}$$

gdzie:

- k – liczba sukcesów w N próbach,
- p – prawdopodobieństwo sukcesu w jednej próbie.

Rozkład wielomianowy: prawdopodobieństwo liczby wystąpień poszczególnych wartości dyskretnej zmiennej losowej w serii wielu realizacji:

$$P(k_1, k_2, \ldots, k_n) = \frac{N!}{k_1! \cdot k_2! \cdot \ldots \cdot k_n!} p_1^{k_1} \cdot p_2^{k_2} \cdot \ldots \cdot p_n^{k_n}$$

gdzie dla $i = 1, 2, \ldots, n$:

- k_i – liczba wystąpień wartości i w N realizacjach,
- p_i – prawdopodobieństwo wartości i w jednej realizacji.

Wielomianowy naiwny klasyfikator bayesowski

Motywacja: lepsza estymacja prawdopodobieństw dla atrybutów o rozkładzie wielomianowym reprezentujących liczby wystąpień – np. zdarzeń określonego typu, poszczególnych słów w tekście itp.

Prawopodobieństwo wartości atrybutów przy danej klasie:

$$P(a_1 = v_1, \ldots, a_n = v_n \mid c = d) = \frac{N!}{v_1! \cdot \ldots \cdot v_n!} p_{1,d}^{v_1} \cdot \ldots \cdot p_{n,d}^{v_n}$$

gdzie:

- a_i – atrybut odpowiadający zdarzeniu/słowu i,
- v_i – liczba wystąpień zdarzenia/słowa i,
- $N = v_1 + v_2 + \cdots + v_n$,
- $p_{i,d} = \frac{N_{i,T_{c=d}}}{N_{T_{c=d}}}$ – prawdopodobieństwo pojedynczego wystąpienia zdarzenia/słowa i w klasie d, przy czym $N_{i,T_{c=d}}$ – łączna liczba wystąpień zdarzenia/słowa i w $T_{c=d}$, $N_{T_{c=d}}$ – łączna liczba wystąpień wszystkich zdarzeń/słów w $T_{c=d}$,
- n – liczba możliwych zdarzeń/wielkość słownika.

Dopełnieniowy naiwny klasyfikator bayesowski

Motywacja: lepsza estymacja prawdopodobieństw warunkowych przy klasyfikacji wieloklasowej z niezrównoważonymi klasami.

Realizacja: bayesowskie wyznaczanie dopełnień prawdopodobieństw klas:

$$P(c \neq d \mid a_1 = v_1, \ldots, a_n = v_n) = \frac{P(c \neq d)P(a_1 = v_1, \ldots, a_n = v_n \mid c \neq d)}{P(a_1 = v_1, \ldots, a_n = v_n)}$$

następnie

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

Typowe zastosowanie: klasyfikacja przy wielu niezrównoważonych klasach (zwłaszcza klasyfikacja tekstu – oryginalnie algorytm przedstawiony jako modyfikacja wielomianowego naiwnego klasyfikatora bayesowskiego).

1. Drugie spojrzenie na naiwny klasyfikator bayesowski
   - Model wielomianowy
   - Model dopełnieniowy

2. Drugie spojrzenie na modele liniowe
   - Pokonywanie ograniczenia liniowości
   - Logarytm wiarygodności
   - Drugie spojrzenie na regresję logistyczną
   - Kalibracja predykcji probabilistycznych
   - Regularyzacja modeli liniowych

Pokonywanie ograniczenia liniowości

Reprezentacja kawałkami liniowa: drzewo modeli (por. wykład 5).

Uogólniona reprezentacja liniowa: reprezentacja modelu przez złożenie:

wewnętrznej funkcji liniowej:

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

zewnętrznej nieliniowej funkcji łączącej:

$$L(h(x)) = g(x)$$
$$h(x) = L^{-1}(g(x))$$

gdzie g – wewnętrzna funkcja liniowa, L – nieliniowa funkcja łącząca.

Przypadek szczególny: regresja logistyczna (będzie omawiana dalej).

Transformacja atrybutów: wprowadzane nowe atrybuty $a'_1, a'_2, \ldots, a'_N$ nieliniowo zależne od pierwotnych atrybutów $a_1, a_2, \ldots, a_n$, przy czym zwykle $N \gg n$, np. regresja wielomianowa, funkcje jądrowe (będą omawiane na jednym z kolejnych wykładów).

Logarytm wiarygodności dla regresji

Wiarygodność $f(x)$ dla przykładów trenujących według modelu:

$$\mathrm{L}_{T,f}(h) = P(T, f; h) = \prod_{x \in T} P(f(x)|x; h)$$

$$P(f(x)|x; h) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(f(x)-h(x))^2}{2\sigma^2}}$$

przy założeniu normalnego rozkładu błędów modelu z pewnym ustalonym odchyleniem standardowym σ.

Logarytm wiarygodności:

$$\begin{aligned}\mathrm{LL}_{T,f}(\pi) &= \ln P(T, f; \pi) = \sum_{x \in T} \ln P(f(x)|x; h) \\ &= \sum_{x \in T} \left(\ln \frac{1}{\sigma\sqrt{2\pi}} - \frac{(f(x) - h(x))^2}{2\sigma^2} \right)\end{aligned}$$

Estymacja parametrów: maksymalizacja logarytmu wiarygodności równoważna minimalizacji błędu średniokwadratowego.

Logarytm wiarygodności dla klasyfikacji

Wiarygodność $c(x)$ dla przykładów trenujących według modelu:

$$\mathrm{L}_{T,c}(\pi) = P(T, c; \pi) = \prod_{x \in T} P(c(x)|x; \pi)$$

$$P(c(x)|x; \pi) = \pi(x)^{c(x)}(1 - \pi(x))^{1-c(x)}$$

gdzie $\pi(x)$ – prawdopodobieństwo klasy 1 wyznaczane przez model.

Logarytm wiarygodności:

$$\begin{aligned}\mathrm{LL}_{T,c}(\pi) = \ln P(T, c; \pi) &= \sum_{x \in T} \ln P(c(x)|x; \pi) \\ &= \sum_{x \in T} \big(c(x) \ln \pi(x) + (1 - c(x)) \ln(1 - \pi(x))\big)\end{aligned}$$

Estymacja parametrów: maksymalizacja logarytmu wiarygodności.

Estymacja parametrów regresji logistycznej

Gradient logarytmu wiarygodności:

$$\begin{aligned}
\nabla_{\mathbf{w}} \mathrm{LL}_{T,c}(\pi) &= \sum_{x \in T} \Big(c(x) \frac{1}{\pi(x)} + (1 - c(x)) \frac{-1}{1 - \pi(x)} \Big) \nabla_{\mathbf{w}} \pi(x) \\
&= \sum_{x \in T} \frac{c(x) - \pi(x)}{\pi(x)(1 - \pi(x))} \nabla_{\mathbf{w}} \pi(x) \\
&= \sum_{x \in T} \big(c(x) - \pi(x) \big) \nabla_{\mathbf{w}} g(x) \\
&= \sum_{x \in T} \big(c(x) - \pi(x) \big) \mathbf{a}(x)
\end{aligned}$$

$$\begin{aligned}
\pi(x) &= \mathrm{logit}^{-1}(g(x)) \\
\mathrm{logit}(p) &= \ln \frac{p}{1 - p} \\
\mathrm{logit}^{-1}(q) &= \frac{e^q}{e^q + 1}
\end{aligned}$$

gdyż

$$\begin{aligned}
(\mathrm{logit}^{-1})'(q) &= \mathrm{logit}^{-1}(q)(1 - \mathrm{logit}^{-1}(q)) \\
\nabla_{\mathbf{w}} \pi(x) &= \pi(x)(1 - \pi(x)) \nabla_{\mathbf{w}} g(x) \\
\nabla_{\mathbf{w}} g(x) &= \mathbf{a}(x)
\end{aligned}$$

Minimalizacja straty logarytmicznej

Alternatywne sformułowanie: spadek gradientu zastosowany do minimalizacji *straty logarytmicznej*.

Strata logarytmiczna: zanegowany logarytm wiarygodności:

$$L_{T,c}(\pi) = \sum_{x \in T} \mathcal{L}(c(x), \pi(x))$$

$$\mathcal{L}(c(x), \pi(x)) = -c(x) \ln \pi(x) - (1 - c(x)) \ln(1 - \pi(x))$$

(*logarithmic loss*, *log loss*, *logistic loss*, *cross-entropy loss*)

Właściwości regresji logistycznej

- Zwykle dość dobra jakość klasyfikacji.
- Ograniczone ryzyko nadmiernego dopasowania.
- Skalibrowane predykcje probabilistyczne dzięki maksymalizacji logarytmu wiarygodności.

Skalibrowane predykcje

Skalibrowane predykcje prawdopodobieństw klas: w zbiorze losowo wybranych przykładów, dla których według modelu $P(d|x) = p$, klasa d występuje z częstością zbliżoną do p.

Modele o nieskalibrowanych predykcjach:

- **drzewa decyzyjne:** prawdopodobieństwa w liściach ekstremalizowane („spychane" ku wartościom 0 i 1 ze względu na stosowane kryteria wyboru podziałów i kryteria stopu,
- **naiwny klasyfikator bayesowski:** prawdopodobieństwa niepoprawne ze względu na niespełnione założenie o niezależności.

Modele o skalibrowanych predykcjach:

- **regresja logistyczna:** maksymalizacja logarytmu wiarygodności dopasowuje prawdopodobieństwa do danych.

Skalowanie Platta

Kalibracja prawdopodobieństw: zastosowanie regresji logistycznej do wyjścia modelu o nieskalibrowanych predykcjach („opakowanie" modelu w funkcję logistyczną do wyznaczania predykcji prawdopodobieństwa klasy 1):

$$\pi(x) = \text{logit}^{-1}(a \cdot s(x) + b)$$

gdzie

- $s(x)$ – zwracane przez model nieskalibrowane prawdopodobieństwo klasy 1 (lub dowolny liczbowy *score* – miara „przekonania" modelu, że przykład należy do klasy 1,
- a, b – parametry estymowane analogicznie jak parametry regresji logistycznej w celu maksymalizacji logarytmu wiarygodności.

Skalowanie Platta

Zmniejszenie ryzyka nadmiernego dopasowania:

- do estymacji parametrów a, b używany osobny zbiór przykładów T' (inny niż zbiór trenujący T, na którym tworzono wewnętrzny model),
- zamiast prawdziwych klas $c(x)$ przy estymacji parametrów a, b używane „wygładzone” numeryczne klasy:

$$c'(x) = \begin{cases} \frac{|T'_{c=1}|+1}{|T'_{c=1}|+2} & \text{jeśli } c(x) = 1 \\ \frac{1}{|T'_{c=0}|+2} & \text{jeśli } c(x) = 0 \end{cases}$$

Metody regularyzacji

Cel: ograniczanie ryzyka nadmiernego dopasowania.

Realizacja: dodawanie składnika kary do minimalizowanej przy tworzeniu modelu funkcji straty, np.:

regularyzacja L2 (*ridge*):

$$\frac{1}{2}\lambda \sum_i w_i^2$$

regularyzacja L1 (*lasso*):

$$\lambda \sum_i |w_i|$$

Efekt: preferowanie modeli o mniejszych wartościach parametrów – mniejsza wrażliwość predykcji na zmiany wartości atrybutów.

Metody regularyzacji

Parametr $\lambda \geq 0$: najczęściej ustalany przez strojenie.

Skalowanie atrybutów: zwykle stosowane aby uniknąć nadmiernej kary za wartości parametrów odpowiadających atrybutom o małym zakresie wartości (transformacje atrybutów będą omawiane na jednym z kolejnych wykładów).

L1 vs. L2:

- L1 sprzyja ustaleniu wartości 0 dla wielu parametrów,
- L2 sprzyja bardziej równomiernemu zmniejszaniu parametrów.

Regresja liniowa z regularyzacją L2

Funkcja celu do estymacji parametrów (L2): minimalizacja

$$J_{T,f}(\mathbf{w}) = \frac{1}{2}\frac{1}{|T|}\sum_{x\in T}(f(x)-h(x))^2 + \frac{1}{2}\lambda\sum_{i=1}^{n} w_i^2$$

Gradient funkcji celu (L2):

$$\begin{aligned}
\nabla_{\mathbf{w}_{1:n}} J_{T,f}(\mathbf{w}) &= \frac{1}{|T|}\sum_{x\in T}\big(f(x)-h(x)\big)\big(-\nabla_{\mathbf{w}_{1:n}} h(x)\big) + \lambda\mathbf{w}_{1:n} \\
&= -\frac{1}{|T|}\sum_{x\in T}\big(f(x)-h(x)\big)\mathbf{a}_{1:n}(x) + \lambda\mathbf{w}_{1:n} \\
&= -\frac{1}{|T|}\sum_{x\in T}\Big(\big(f(x)-h(x)\big)\mathbf{a}_{1:n}(x) - \lambda\mathbf{w}_{1:n}\Big) \\
\frac{\partial J_{T,f}(\mathbf{w})}{\partial w_{n+1}} &= -\frac{1}{|T|}\sum_{x\in T}(f(x)-h(x))
\end{aligned}$$

gdzie $\mathbf{w}_{1:n}$, $\mathbf{a}_{1:n}(x)$ – wektory parametrów $w_1, w_2, \ldots, w_n$ i wartości atrybutów $a_1(x), a_2(x), \ldots, a_n(x)$.

Regresja liniowa z regularyzacją L2

Gradientowa reguła aktualizacji parametrów (L2):

$$\mathbf{w}_{1:n} := \mathbf{w}_{1:n} + \beta\Big(\big(f(x) - h(x)\big)\mathbf{a}_{1:n}(x) - \lambda\mathbf{w}_{1:n}\Big)$$
$$w_{n+1} := w_{n+1} + \beta(f(x) - h(x))$$

Metoda regularyzowanych najmniejszych kwadratów (L2): rozwiązanie równania $\nabla_{\mathbf{w}_{1:n}} J_{T,f}(\mathbf{w}) = 0$.

Metoda regularyzowanych najmniejszych kwadratów

- Pomijając dla uproszczenia składnik stały modelu (przyjmując $w_{n+1} = 0$):

$$
\begin{aligned}
&-\frac{1}{|T|}\mathbf{a}_{1:n}^{\mathsf{T}}(T)(\mathbf{f}(T) - \mathbf{a}_{1:n}(T)\mathbf{w}_{1:n}) + \lambda\mathbf{w}_{1:n} = 0 \\
&-\mathbf{a}_{1:n}^{\mathsf{T}}(T)\mathbf{f}(T) + \mathbf{a}_{1:n}^{\mathsf{T}}(T)\mathbf{a}_{1:n}(T)\mathbf{w}_{1:n} + \lambda|T|\mathbf{w}_{1:n} = 0 \\
&\mathbf{w}_{1:n} = \left(\mathbf{a}_{1:n}^{\mathsf{T}}(T)\mathbf{a}_{1:n}(T) + \lambda|T|\mathbf{I}_n\right)^{-1}\mathbf{a}_{1:n}^{\mathsf{T}}(T)\mathbf{f}(T)
\end{aligned}
$$

  gdyż

$$
\sum_{x\in T}(f(x) - h(x))\mathbf{a}_{1:n}(x) = \mathbf{a}_{1:n}^{\mathsf{T}}(T)(\mathbf{f}(T) - \mathbf{a}_{1:n}(T)\mathbf{w}_{1:n})
$$

- $\mathbf{a}_{1:n}(T)$ – macierz $|T| \times n$ wartości atrybutów $a_1(x), a_2(x), \ldots, a_n(x)$ dla $x \in T$,
- $\mathbf{f}(T)$ – wektor kolumnowy wartości $f(x)$ dla $x \in T$,
- $\mathbf{I}_n$ – macierz jednostkowa o wymiarach $n \times n$.

Metoda regularyzowanych najmniejszych kwadratów

- Składnik stały można pominąć (przyjąć $w_{n+1} = 0$) jeśli:
  - dla każdego $i = 1, 2, \ldots, n$ średnia wartość a_i na T, $m_T(a_i)$, jest równa 0, tzn. macierz $\mathbf{a}_{1:n}(T)$ jest *wycentrowana*,
  - średnia wartość f na T, $m_T(f)$, jest równa 0, tzn. wektor $\mathbf{f}(T)$ jest *wycentrowany*.
- Centrowanie $\mathbf{a}_{1:n}(T)$:

$$a'_i(x) = a_i(x) - m_T(a_i)$$

przy czym odpowiednia transformacja atrybutów musi być wówczas także zastosowana dla danych, dla których model jest stosowany do predykcji (odejmowane średnie wyznaczone *na zbiorze trenującym*).

- Centrowanie $\mathbf{f}(T)$:

$$f'(x) = f(x) - m_T(f)$$

przy czym przy predykcji musi być wówczas dodany składnik stały $w_{n+1} = m_T(f)$.

Regresja liniowa z regularyzacją L1

Funkcja celu do estymacji parametrów (L1): minimalizacja

$$J_{T,f}(\mathbf{w}) = \frac{1}{|T|} \sum_{x \in T} (f(x) - h(x))^2 + \lambda \sum_{i=1}^{n} |w_i|$$

Estymacja parametrów (L1): metody subgradientowe (poza zakresem wykładu).

Regresja liniowa z regularyzacją