Zaawansowane uczenie maszynowe:
wykład 7

Paweł Cichosz

Regresja logistyczna z regularyzacją

Funkcja celu do estymacji parametrów (L2): minimalizacja

$$
\begin{aligned}
J_{T,c}(\mathbf{w}) &= -\frac{1}{|T|}\mathrm{LL}_{T,c}(\pi) + \frac{1}{2}\lambda\sum_{i=1}^{n} w_i^2 \\
&= -\frac{1}{|T|}\sum_{x\in T}\big(c(x)\ln\pi(x) + (1-c(x))\ln(1-\pi(x))\big) + \frac{1}{2}\lambda\sum_{i=1}^{n} w_i^2
\end{aligned}
$$

Gradient funkcji celu (L2):

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

Regresja logistyczna z regularyzacją

Gradientowa reguła aktualizacji parametrów (L2):

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

Funkcja celu do estymacji parametrów (L1): minimalizacja

$$J_{T,c}(\mathbf{w}) = -\frac{1}{|T|}\sum_{x\in T}\big(c(x)\ln\pi(x) + (1-c(x))\ln(1-\pi(x))\big) + \lambda\sum_{i=1}^{n}|w_i|$$

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

Mnożniki Lagrange'a

Metoda mnożników Lagrange'a: wyznaczanie ekstremów warunkowych funkcji różniczkowalnych przez uwzględnianie ograniczeń w funkcji celu.

Lagrangian (twardy margines):

$$\mathcal{L}(\mathbf{w}, \alpha) = \frac{1}{2}\|\mathbf{w}_{1:n}\|^2 - \sum_{x \in T} \alpha_x (c_-(x)\mathbf{w} \circ \mathbf{a}(x) - 1)$$

Mnożniki Lagrange'a: α_x – mnożnik dla przykładu x.

Postać dualna

Transformacja zadania optymalizacji:

$$\Theta_{\mathcal{P}}(\mathbf{w}) = \max_{\alpha \geq 0} \mathcal{L}(\mathbf{w}, \alpha)$$
$$\Theta_{\mathcal{D}}(\alpha) = \min_{\mathbf{w}} \mathcal{L}(\mathbf{w}, \alpha)$$

- Minimalizacja $\Theta_{\mathcal{P}}$ względem $\mathbf{w}$ równoważna minimalizacji $\frac{1}{2}\|\mathbf{w}_{1:n}\|^2$ jeśli spełnione są ograniczenia, gdyż:

$$\Theta_{\mathcal{P}}(\mathbf{w}) = \begin{cases} \frac{1}{2}\|\mathbf{w}_{1:n}\|^2 & \text{jeśli } (\forall x \in T)\ c_-(x)\mathbf{w} \circ \mathbf{a}(x) \geq 1 \\ \infty & \text{w przeciwnym przypadku} \end{cases}$$

- Maksymalizacja $\Theta_{\mathcal{D}}$ różni się od minimalizacji $\Theta_{\mathcal{P}}$ zamianą kolejności $\min/\max$.

Postać dualna

- Można udowodnić, że:

$$\max_{\alpha \geq 0} \Theta_{\mathcal{D}}(\alpha) \leq \min_{\mathbf{w}} \Theta_{\mathcal{P}}(\mathbf{w})$$

- Równość jeśli istnieje $\mathbf{w}$ spełniające ograniczenia, przy czym dla rozwiązań optymalnych $\mathbf{w}^*$ i α^*:

$$(\forall x \in T) \quad \alpha_x^*(c_-(x)\mathbf{w}^* \circ \mathbf{a}(x) - 1) = 0$$

tzn. jeśli $\alpha_x^* > 0$, to $c_-(x)\mathbf{w}^* \circ \mathbf{a}(x) = 1$, czyli x jest wektorem podpierającym – „leży na marginesie” (warunki KKT: Karush-Kuhn-Tucker).

Minimalizacja Lagrangianu względem $\mathbf{w}$

- Minimalizacja Lagrangianu względem $\mathbf{w}$, wyznaczenie stąd $\mathbf{w}$ i wstawienie do $\mathcal{L}(\mathbf{w}, \alpha)$ prowadzi do funkcji celu dla zadania dualnego.
- Przyrównanie gradientu Lagrangianu $\nabla_{\mathbf{w}} \mathcal{L}(\mathbf{w}, \alpha)$ do 0:

$$\nabla_{\mathbf{w}_{1:n}} \mathcal{L}(\mathbf{w}, \alpha) = \mathbf{w}_{1:n} - \sum_{x \in T} \alpha_x c_-(x) \mathbf{a}_{1:n}(x) = 0$$

$$\mathbf{w}_{1:n} = \sum_{x \in T} \alpha_x c_-(x) \mathbf{a}_{1:n}(x)$$

$$\frac{\partial \mathcal{L}(\mathbf{w}, \alpha)}{\partial w_{n+1}} = -\sum_{x \in T} \alpha_x c_-(x) = 0$$

Minimalizacja Lagrangianu względem $\mathbf{w}$

Po wstawieniu w miejsce $\mathbf{w}$ wyrażeń wyznaczonych wcześniej przez minimalizację względem $\mathbf{w}$:

$$\begin{aligned}\min_{\mathbf{w}} \mathcal{L}(\mathbf{w}, \alpha) = &\frac{1}{2} \sum_{x_1 \in T} \sum_{x_2 \in T} \alpha_{x_1} \alpha_{x_2} c_-(x_1) c_-(x_2) \mathbf{a}_{1:n}(x_1) \circ \mathbf{a}_{1:n}(x_2) \\ &- \sum_{x_1 \in T} \sum_{x_2 \in T} \alpha_{x_1} \alpha_{x_2} c_-(x_1) c_-(x_2) \mathbf{a}_{1:n}(x_1) \circ \mathbf{a}_{1:n}(x_2) \\ &- w_{n+1} \sum_{x \in T} \alpha_x c_-(x) + \sum_{x \in T} \alpha_x \\ = &-\frac{1}{2} \sum_{x_1 \in T} \sum_{x_2 \in T} \alpha_{x_1} \alpha_{x_2} c_-(x_1) c_-(x_2) \mathbf{a}_{1:n}(x_1) \circ \mathbf{a}_{1:n}(x_2) + \sum_{x \in T} \alpha_x\end{aligned}$$

gdyż

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

$$\sum_{x \in T} \alpha_x c_-(x) = 0$$

Zadanie dualne: twardy margines

maksymalizacja:

$$-\frac{1}{2}\sum_{x_1\in T}\sum_{x_2\in T} c_-(x_1)c_-(x_2)\alpha_{x_1}\alpha_{x_2}\mathbf{a}_{1:n}(x_1)\circ\mathbf{a}_{1:n}(x_2)+\sum_{x\in T}\alpha_x$$

przy ograniczeniach:

$$\sum_{x\in T} c_-(x)\alpha_x = 0$$

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

Zadanie dualne: miękki margines

maksymalizacja:

$$-\frac{1}{2}\sum_{x_1\in T}\sum_{x_2\in T} c_-(x_1)c_-(x_2)\alpha_{x_1}\alpha_{x_2}\mathbf{a}_{1:n}(x_1)\circ\mathbf{a}_{1:n}(x_2)+\sum_{x\in T}\alpha_x$$

przy ograniczeniach:

$$\sum_{x\in T} c_-(x)\alpha_x = 0$$

$$(\forall x\in T)\quad 0\le\alpha_x\le\mathcal{C}$$

Mnożniki Lagrange'a jako parametry modelu

Wektory podpierające: przykłady, dla których $\alpha_x > 0$ („leżące na marginesie" jeśli $0 < \alpha_x < C$) – z warunków KKT (Karush-Kuhn-Tucker).

Parametry postaci prymalnej:

$$\mathbf{w}_{1:n} = \sum_{x \in T} \alpha_x c_-(x) \mathbf{a}_{1:n}(x)$$

Predykcja:

$$g(x) = \mathbf{w} \circ \mathbf{a}(x) = \sum_{x' \in T} \alpha_{x'} c_-(x') \mathbf{a}_{1:n}(x') \circ \mathbf{a}_{1:n}(x) + w_{n+1}$$

przy czym w_{n+1} jest wyznaczane z ograniczenia postaci prymalnej (dla przykładu x_s „leżącego na marginesie" $c_-(x_s)\mathbf{w} \circ \mathbf{a}(x_s) = 1$).

Sztuczka jądrowa

Postać dualna a iloczyn skalarny: użycie przykładów w czasie tworzenia modelu i predykcji *wyłącznie jako argumentów iloczynu skalarnego*.

Sztuczka jądrowa: niejawna transformacja atrybutów przez zastąpienie iloczynu skalarnego $\mathbf{a}_{1:n}(x_1) \circ \mathbf{a}_{1:n}(x_2)$ wartością funkcji jądrowej $K(\mathbf{a}_{1:n}(x_1), \mathbf{a}_{1:n}(x_2))$ takiej, że

$$K(\mathbf{a}_{1:n}(x_1), \mathbf{a}_{1:n}(x_2)) = \mathbf{a}'_{1:N}(x_1) \circ \mathbf{a}'_{1:N}(x_2)$$

w pewnej nowej reprezentacji z atrybutami $a'_1, a'_2, \ldots, a'_N$.

Sztuczka jądrowa

Korzyści:

- $K(\mathbf{a}_{1:n}(x_1), \mathbf{a}_{1:n}(x_2))$ oblicza się bezpośrednio na podstawie $\mathbf{a}_{1:n}(x_1)$ i $\mathbf{a}_{1:n}(x_2)$ bez jawnego wyznaczania $\mathbf{a}'_{1:N}(x_1)$ i $\mathbf{a}'_{1:N}(x_2)$,
- niejawna nieliniowa transformacja atrybutów bez konieczności faktycznego definiowania nowych atrybutów i obliczania ich wartości.

Prosty przykład:

$$K(\mathbf{a}_{1:n}(x_1), \mathbf{a}_{1:n}(x_2)) = (\mathbf{a}_{1:2}(x_1) \circ \mathbf{a}_{1:2}(x_2))^2$$

$$\begin{aligned} K(\mathbf{a}_{1:n}(x_1), \mathbf{a}_{1:n}(x_2)) &= a_1^2(x_1)a_1^2(x_2) + a_2^2(x_1)a_2^2(x_2) \\ &\quad + 2a_1(x_1)a_2(x_1)a_1(x_2)a_2(x_2) \end{aligned}$$

$$a'_1(x) = a_1^2(x), \quad a'_2(x) = a_2^2(x), \quad a'_3(x) = \sqrt{2}a_1(x)a_2(x)$$

Typy nieliniowych funkcji jądrowych

Jądro wielomianowe:

$$K(\mathbf{a}_{1:n}(x_1), \mathbf{a}_{1:n}(x_2)) = (\gamma \mathbf{a}_{1:n}(x_1) \circ \mathbf{a}_{1:n}(x_2) + b)^p$$

gdzie $\gamma > 0$, $b \geq 0$ i $p > 1$ (najczęściej $p = 2$ lub $p = 3$),

Jądro radialne:

$$K(\mathbf{a}_{1:n}(x_1), \mathbf{a}_{1:n}(x_2)) = e^{-\gamma \|\mathbf{a}_{1:n}(x_1) - \mathbf{a}_{1:n}(x_2)\|^2}$$

gdzie $\gamma > 0$.

Warunek Mercera: K jest funkcją jądrową wtedy i tylko wtedy, gdy dla dowolnego zbioru $V \subset \mathcal{R}^n$ macierz wartości $K(\mathbf{v}_1, \mathbf{v}_2)$ dla $\mathbf{v}_1, \mathbf{v}_2 \in V$ jest symetryczna i dodatnio półokreślona.

Funkcja straty SVR

SVM dla zadania regresji: *support vector regression* (SVR)

Tolerancja małych residuów:

$$|f(x) - h(x)| \leq \epsilon$$

Strata ϵ-niewrażliwa:

$$\mathcal{L}(f(x), h(x)) = \begin{cases} 0 & \text{jeśli } |f(x) - h(x)| \leq \epsilon \\ |f(x) - h(x)| - \epsilon & \text{w przeciwnym przypadku} \end{cases}$$

Preferencja dla „płaskich" aproksymacji: minimalizacja $\|\mathbf{w}_{1:n}\|$ ogranicza wrażliwość na wartości atrybutów i zmniejsza ryzyko nadmiernego dopasowania.

Zadanie optymalizacji SVR

minimalizacja:

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

przy ograniczeniach:

$$(\forall x \in T) \quad f(x) - \mathbf{w} \circ \mathbf{a}(x) \leq \epsilon$$
$$(\forall x \in T) \quad \mathbf{w} \circ \mathbf{a}(x) - f(x) \leq \epsilon$$

Zadanie optymalizacji SVR z rozluźnionymi ograniczeniami

minimalizacja:

$$\frac{1}{2}\|\mathbf{w}_{1:n}\|^2 + \mathcal{C}\sum_{x\in T}(\xi_x^+ + \xi_x^-)$$

przy ograniczeniach:

$$\begin{aligned}
(\forall x \in T) \quad & f(x) - \mathbf{w} \circ \mathbf{a}(x) \leq \epsilon + \xi_x^+ \\
(\forall x \in T) \quad & \mathbf{w} \circ \mathbf{a}(x) - f(x) \leq \epsilon + \xi_x^- \\
(\forall x \in T) \quad & \xi_x^+ \geq 0 \\
(\forall x \in T) \quad & \xi_x^- \geq 0
\end{aligned}$$

Postać dualna SVR

maksymalizacja:

$$-\frac{1}{2}\sum_{x_1\in T}\sum_{x_2\in T}(\alpha^+_{x_1}-\alpha^-_{x_1})(\alpha^+_{x_2}-\alpha^-_{x_2})\mathbf{a}_{1:n}(x_1)\circ\mathbf{a}_{1:n}(x_2)$$
$$+\sum_{x\in T}(\alpha^+_x-\alpha^-_x)f(x)-\sum_{x\in T}(\alpha^+_x+\alpha^-_x)\epsilon$$

przy ograniczeniach:

$$(\forall x\in T)\quad \sum_{x\in T}(\alpha^+_x-\alpha^-_x)=0$$
$$(\forall x\in T)\quad 0\le\alpha^+_x\le\mathcal{C}$$
$$(\forall x\in T)\quad 0\le\alpha^-_x\le\mathcal{C}$$

Funkcje jądrowe: analogicznie jak dla klasyfikacji.

Właściwości algorytmów SVM i SVR

- Należą do najbardziej skutecznych i często używanych algorytmów dla zadań klasyfikacji i regresji:
  - zwiększona odporność na nadmierne dopasowanie nawet przy dużej liczbie atrybutów,
  - możliwość reprezentowania nieliniowych zależności.
- Wymagają staranności przy doborze parametrów (kosztu, ϵ, typu i parametrów funkcji jądrowej).
- Brak możliwości bezpośredniej interpretacji modeli.