Techniki eksploracji danych

Krzysztof Gajowniczek

Rok akademicki: 2020/2021

Section 1

Regresja liniowa

Subsection 1

Estymacja analityczna

- Zwykła regresja liniowa metodą najmniejszych kwadratów próbuje dopasować linię prostą do zbioru punktów danych w taki sposób, aby suma kwadratów błędów była minimalna.

$$f(\boldsymbol{\theta}) = \|\boldsymbol{y} - X\boldsymbol{\theta}\|^2 = \sum_{i=1}^{n}(y_i - \boldsymbol{\theta}^T \boldsymbol{x}_i)^2$$

gdzie

$$X = \begin{vmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_i^T \\ \vdots \\ \boldsymbol{x}_n^T \end{vmatrix}$$

- Nie istnieje dokładne rozwiązanie dla tego problemy, więc staramy się dopasować dane jak najlepiej:

$$\hat{\boldsymbol{\theta}} = \min_{\boldsymbol{\theta}} \|\boldsymbol{y} - X\boldsymbol{\theta}\|^2$$

- Oszacowanie parametrów modelu uzyskuje się za pomocą układu równań normalnych, który wykorzystując notację macierzową ma następującą postać:

$$\hat{\boldsymbol{\theta}} = (X^T X)^{-1} X^T \boldsymbol{y}$$

- Sugeruje to, że $(X^T X)^{-1}$ i $X^T \boldsymbol{y}$ są wystarczającymi statystykami dla $\hat{\boldsymbol{\theta}}$.
- Wymiarowość X to $n \times p$, dlatego $X^T X$ ma wymiar $p \times p$ oraz $X^T \boldsymbol{y}$ ma wymiar $p \times 1$.
- Żadne z powyższych nie jest zależne od n, więc wymiar takiej statystyki nie rośnie wraz z danymi.
- Ilość danych jest często większa niż ich wymiar, a także pozwala na ustalenie określonej ilości miejsca na przechowywanie odpowiednich statystyk, niezależnie od tego, ile danych jest podanych.

- Powyższą estymację analityczną można uzyskać także za pomocą macierzy Gramma $K = XX^T$ o wymiarze $n \times n$, będącą iloczynem skalarnym wszystkich możliwych par obserwacji.
- To jest nieparametryczne podejście. Rośnie wraz z danymi.

$$\hat{\boldsymbol{\theta}} = X^T(XX^T)^{-1}\boldsymbol{y}$$

$$\hat{\boldsymbol{\theta}} = X^T\boldsymbol{\alpha}$$

- gdzie:

$$\boldsymbol{\alpha} = (XX^T)^{-1}\boldsymbol{y}$$

- Przewidywanie wartości nowych obserwacji uzyskuję się za pomocą:

$$\hat{y}_i = \hat{\boldsymbol{\theta}}^T \boldsymbol{x}_i = \boldsymbol{x}_i^T \hat{\boldsymbol{\theta}}$$

- lub (wykorzystując estymaty uzyskane za pomoca macierzy Gramma):

$$\hat{y}_i = \boldsymbol{x}_i^T X^T (X X^T)^{-1} \boldsymbol{y}$$

- lub:

$$\hat{y}_i = \boldsymbol{x}_i^T X^T \boldsymbol{\alpha} = \sum_{j=1}^{n} \boldsymbol{x}_i^T \boldsymbol{x}_j \alpha_j$$

Subsection 2

Estymacja iteracyjna

- Zdefiniujmy na nowo funkcję błędu. Zamiast sumy kwadratów reszt:

$$f(\boldsymbol{\theta}) = \sum_{i=1}^{n} (y_i - \boldsymbol{\theta}^T \boldsymbol{x}_i)^2$$

- wykorzystywać będziemy zmodyfikowany błąd średnio-kwadratowy:

$$f(\boldsymbol{\theta}) = \frac{1}{2n} \sum_{i=1}^{n} (y_i - \boldsymbol{\theta}^T \boldsymbol{x}_i)^2$$

$$f(\boldsymbol{\theta}) = \frac{1}{2n} \sum_{i=1}^{n} (\boldsymbol{\theta}^T \boldsymbol{x}_i - y_i)^2$$

- Głównym pomysłem jest przyjęcie pochodnej cząstkowej funkcji kosztu względem $\boldsymbol{\theta}$.
- Ten gradient pomnożony przez współczynnik uczenia się staje się regułą aktualizacji szacowanych wartości parametrów.
- Dla każdego θ_j olbicza się:

$$\theta_j = \theta_j - \alpha \frac{\partial}{\partial \theta_j} f(\boldsymbol{\theta})$$

- co po przekształceniach daje:

$$\theta_j = \theta_j - \alpha \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{\theta}^T \boldsymbol{x}_i - y_i) * x_i^{(j)}$$

Section 2

Regresja grzbietowa

- Kiedy w zbiorze danych istnieje wiele skorelowanych zmiennych, współczynniki $\boldsymbol{\theta}$ w modelach regresji liniowej, mogą być słabo określone i mieć wiele wariancji.
- Jednym z rozwiązań tego problemu jest ograniczenie parametrów $\boldsymbol{\theta}$ tak, aby nie przekraczały pewnego budżetu C, inaczej mowiąc aby parametry nie miały zbyt dużych wartości.
- Jest to równoważne zastosowaniu regularyzacji $L2$, znanej również jako “rozpad wag”.

Subsection 1

Estymacja analityczna - forma pierwotna

- Funkcja celu przyjmuje teraz postać:

$$\hat{\boldsymbol{\theta}} = \min_{\boldsymbol{\theta}} \|X\boldsymbol{\theta} - \boldsymbol{y}\|^2 + \lambda\|\boldsymbol{\theta}\|^2$$

- gdzie λ jest parametrem regularyzacji.
- Po odpowiednich przekształceniach otrzymuje się:

$$\hat{\boldsymbol{\theta}} = (X^T X + \lambda I)^{-1} X^T \boldsymbol{y}$$

- co jest formą bardzo podobną do zwykłej regresji liniowej, ale tutaj dodaje się λ do każdego elementu diagonalnego $X^T X$.

- Powyższy wzór można zapisać jako:

$$(X^T X + \lambda I)\hat{\boldsymbol{\theta}} = X^T \boldsymbol{y} =$$
$$X^T X\hat{\boldsymbol{\theta}} + \lambda\hat{\boldsymbol{\theta}} = X^T \boldsymbol{y} =$$
$$\lambda\hat{\boldsymbol{\theta}} = X^T \boldsymbol{y} - X^T X\hat{\boldsymbol{\theta}} = X^T(\boldsymbol{y} - X\hat{\boldsymbol{\theta}}) =$$
$$\hat{\boldsymbol{\theta}} = \lambda^{-1}X^T(\boldsymbol{y} - X\hat{\boldsymbol{\theta}}) =$$

$$\hat{\boldsymbol{\theta}} = X^T\boldsymbol{\alpha}$$

- gdzie:

$$\boldsymbol{\alpha} = \lambda^{-1}(\boldsymbol{y} - X\hat{\boldsymbol{\theta}})$$

- Tak samo jak w przypadku regresji liniowej powyższą estymację analityczną można uzyskać także za pomocą macierzy Gramma:

$$\lambda \boldsymbol{\alpha} = \boldsymbol{y} - X\hat{\boldsymbol{\theta}}$$

- posiłkując się podstawieniem ze slajdu nr 8 otrzymujemy (macierz Gramma):

$$\lambda \boldsymbol{\alpha} = \boldsymbol{y} - XX^T\boldsymbol{\alpha}$$
$$XX^T\boldsymbol{\alpha} + \lambda\boldsymbol{\alpha} = \boldsymbol{y}$$
$$(XX^T + \lambda I)\boldsymbol{\alpha} = \boldsymbol{y}$$
$$\boldsymbol{\alpha} = (XX^T + \lambda I)^{-1}\boldsymbol{y}$$

Subsection 2

Estymacja analityczna - wersja jądrowa

- Jądra są używane do obliczania iloczynu skalarnego dwóch wektorów w pewnej przestrzeni cech, nawet bez odwiedzania jej.
- Teraz zastępujemy wszystkie obserwacje, wektorem ich cech w nowej przestrzeni (mogącej mieć mniej lub więcej wymiarów niż oryginalna przestrzeń):

$$\phi : X \to H$$

- Możemy postrzegać jądro jako:

$$k(\boldsymbol{x}_i, \boldsymbol{x}_n) = \langle \phi(\boldsymbol{x}_i), \phi(\boldsymbol{x}_n) \rangle$$

$$K(X, X) = k(\boldsymbol{x}_i, \boldsymbol{x}_n), \forall i, n$$

- chociaż nie wiemy, czym jest $\phi(.)$ wiemy tylko, że istnieje.

- Jądro liniowe:

$$k(\boldsymbol{x}_i, \boldsymbol{x}_n) = \boldsymbol{x}_i^T \boldsymbol{x}_n$$

- Jądro wielomianowe stopnia d i stałej c:

$$k(\boldsymbol{x}_i, \boldsymbol{x}_n) = (\boldsymbol{x}_i^T \boldsymbol{x}_n + c)^d$$

- Jądro o radialnych funkcjach bazowych:

$$k(\boldsymbol{x}_i, \boldsymbol{x}_n) = \exp(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_n\|^2}{2\sigma^2})$$

- Teraz możemy po prostu wziąć rozwiązanie dla regresji grzbietowej i zamienić każde X na $\Phi(X)$:

$$\hat{\boldsymbol{\theta}} = \Phi(X)^T(\Phi(X)\Phi(X)^T + \lambda I)^{-1}\boldsymbol{y}$$

- Macierz Gramma (będąca iloczynem skalarnym) może być zastąpiona dowolną macierzą kwadratową, będącą przekształceniem nieliniowym do nowej przetrzeni:

$$\boldsymbol{\alpha} = (XX^T + \lambda)^{-1}\boldsymbol{y} = (K(X, X) + \lambda I)^{-1}\boldsymbol{y}$$

$$\hat{y}_i = \phi(\boldsymbol{x}_i^T)\Phi(X)^T\boldsymbol{\alpha} = \sum_{j=1}^{n} k(\boldsymbol{x}_i, \boldsymbol{x}_j)\alpha_j$$

- Dlaczego mielibyśmy chcieć zachować macierz K rozmiaru $n \times n$ zamiast pierwotnej macierzy X rozmiaru $n \times p$ lub macierzy $X^T X$ rozmiaru $p \times p$?
- Nie zrobilibyśmy tego, gdybyśmy dokonali rzeczywistej regresji liniowej lub grzebietowej.
- Możemy teraz jednak zastąpić macierz K z dowolną funkcją jądra na regresję grzbietową w innej przestrzeni, uzyskując regresję nieliniową (i nieparametryczną)!

Section 3

Literatura

- *Friedman, J., Hastie, T., Tibshirani, R. (2001). The elements of statistical learning (Vol. 1, No. 10). New York: Springer series in statistics.*