Zadanie regresji liniowej

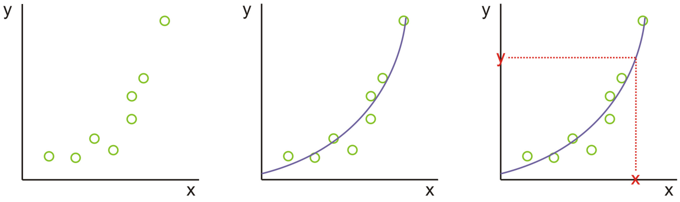
Metody systemowe i decyzyjne

Piotr Dąbrowski

na bazie prezentacji dra. Szymona Zaręby

Problem regresji

- **Zmienne wejściowe** (ang. *input variables*): $\mathbf{x} \in \mathcal{X}$
- **Zmienna wyjściowa** (ang. *target variable*): $y \in \mathbb{R}$
- **Problem:** dla zadanego $\mathcal{D} = \{(\mathbf{x}_n, y_n)\}_{n=1}^{N}$ przewidzieć wartość y dla nowego $\mathbf{x}$.



Przykładowy problem

Mieszkanie o liczbie pokojów równej x kosztuje y tys. złotych. Zebrano następujące dane dotyczące trzech sprzedaży:

x	1	2	3
y	200	400	750

Tabela: Dane dla trzech sprzedaży.

Chcemy przewidywać cenę mieszkania y, jeżeli dane mieszkanie posiada x pokojów.

Ekstrakcja cech

Macierz $= [\phi(\mathbf{x}_1)\ \phi(\mathbf{x}_2) \dots \phi(\mathbf{x}_N)]^{\mathrm{T}}$ (ang. *design matrix*) przyjmuje postać w zależności od definicji cech, np.:

- $\phi(x) = (1 \;\; x \;\; x^2)^{\mathrm{T}}$
- $\phi(x) = \left(1 \;\; e^{-(x-2)^2} \;\; e^{-(x-3)^2}\right)^{\mathrm{T}}$

Ekstrakcja cech

x	1	2	3
y	200	400	750

Tabela: Dane dla trzech klientów.

$$\phi(x) = (1 \;\; x \;\; x^2)^{\mathrm{T}}$$
$$= \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \end{bmatrix}$$

$$\phi(x) = \left(1 \;\; e^{-(x-2)^2} \;\; e^{-(x-3)^2}\right)^{\mathrm{T}}$$
$$= \begin{bmatrix} 1 & e^{-1} & e^{-4} \\ 1 & e^{0} & e^{-1} \\ 1 & e^{-1} & e^{0} \end{bmatrix}$$

Model

Model regresji liniowej przyjmuje następującą postać:

$$\bar{y}(\mathbf{x}) = \phi(\mathbf{x})^{\mathrm{T}}\mathbf{w},$$

gdzie $\mathbf{w} \in \mathbb{R}^D$ jest wektorem parametrów. Przy macierzowej reprezentacji obserwacji

$$\mathbf{X} = [\mathbf{x}_1\ \mathbf{x}_2\ \ldots \mathbf{x}_N]$$

model przyjmuje wygodną postać

$$\bar{\mathbf{y}} = \mathbf{\Phi}\mathbf{w}$$

Predykcja

Dla cech $\phi(x) = (1 \;\; x \;\; x^2)^{\mathrm{T}}$ oraz wektora parametrów $\mathbf{w} = (1 \;\; 2 \;\; 1)^{\mathrm{T}}$, wyznaczyć odpowiedź modelu dla przykładowego wejścia $x = 4$:

$$\phi(4) = (1 \;\; 4 \;\; 16)^{\mathrm{T}}$$

$$\mathbf{w} = (1 \;\; 2 \;\; 1)^{\mathrm{T}}$$

$$\bar{\mathbf{y}} = 1 * 1 + 2 * 4 + 1 * 16 = 25$$

Uczenie

Dla błędu kwadratowego:

$$E(\mathbf{w}) = \frac{1}{2}\|\mathbf{y} - \overline{\mathbf{y}}\|_2^2$$

- wyznacz gradient $\nabla_{\mathbf{w}} E(\mathbf{w})$
- przyrównaj gradient do zera
- wylicz szukaną wartość

Uczenie

$$E(\mathbf{w}) = \frac{1}{2}\|\mathbf{y} - \bar{\mathbf{y}}\|_2^2$$

$$\bar{\mathbf{y}} = \boldsymbol{\Phi}\mathbf{w}$$

$$\begin{aligned}
\nabla_{\mathbf{w}} E(\mathbf{w}) &= \nabla_{\mathbf{w}} \left(\frac{1}{2}\|\mathbf{y} - \boldsymbol{\Phi}\mathbf{w}\|_2^2 \right) \\
&= -2\frac{1}{2}\boldsymbol{\Phi}^T(\mathbf{y} - \boldsymbol{\Phi}\mathbf{w}) \\
&= -\boldsymbol{\Phi}^T(\mathbf{y} - \boldsymbol{\Phi}\mathbf{w})
\end{aligned}$$

Uczenie

Wyznacz wartości parametrów minimalizujące błąd $E(\mathbf{w})$:

$$\nabla_{\mathbf{w}} E(\mathbf{w}) = -\mathbf{\Phi}^T(\mathbf{y} - \mathbf{\Phi}\mathbf{w})$$

$$\mathbf{\Phi}^T(\mathbf{y} - \mathbf{\Phi}\mathbf{w}) = 0$$

$$\mathbf{\Phi}^T\mathbf{y} - \mathbf{\Phi}^T\mathbf{\Phi}\mathbf{w} = 0$$

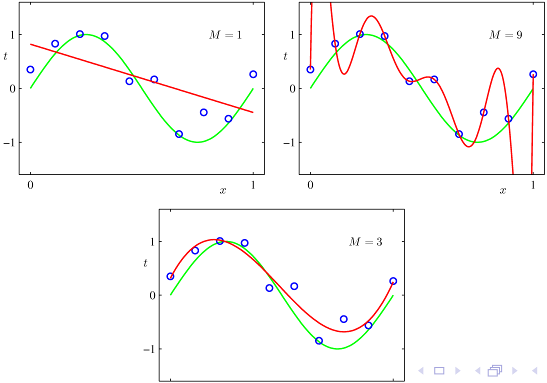
$$\mathbf{\Phi}^T\mathbf{y} = \mathbf{\Phi}^T\mathbf{\Phi}\mathbf{w}$$

$$(\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T\mathbf{y} = \mathbb{I}\mathbf{w}$$

$$\mathbf{w} = (\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T\mathbf{y}$$

Selekcja modelu

- Wybranie zbyt prostego modelu powoduje, że model nie jest w stanie objąć zależności w danych, **underfitting**
- Wybranie modelu zbyt skomplikowanego powoduje, że model uczy się nie tylko natury danych, ale również wszelkich szumów, **overfitting**



Uczenie

Dla błędu kwadratowego z regularyzacją ℓ_2 (Tichonowa):

$$E_\lambda(\mathbf{w}) = \frac{1}{2}\|\mathbf{y} - \bar{\mathbf{y}}\|_2^2 + \frac{\lambda}{2}\|\mathbf{w}\|_2^2$$

gdzie $\lambda > 0$ jest współczynnikiem regularyzacji.

- wyznacz gradient $\nabla_{\mathbf{w}} E_\lambda(\mathbf{w})$
- przyrównaj gradient do zera
- wylicz szukaną wartość

Uczenie

$$E_\lambda(\mathbf{w}) = \frac{1}{2}\|\mathbf{y} - \bar{\mathbf{y}}\|_2^2 + \frac{\lambda}{2}\|\mathbf{w}\|_2^2$$

$$\bar{\mathbf{y}} = \mathbf{\Phi w}$$

$$\begin{aligned}
\nabla_{\mathbf{w}} E(\mathbf{w}) &= \nabla_{\mathbf{w}} \left(\frac{1}{2}\|\mathbf{y} - \mathbf{\Phi w}\|_2^2 + \frac{\lambda}{2}\|\mathbf{w}\|_2^2 \right) \\
&= -2\frac{1}{2}\mathbf{\Phi}^T(\mathbf{y} - \mathbf{\Phi w}) + \frac{\lambda}{2}2\mathbf{w} \\
&= -\mathbf{\Phi}^T(\mathbf{y} - \mathbf{\Phi w}) + \lambda \mathbf{w}
\end{aligned}$$

Uczenie

Wyznacz wartości parametrów minimalizujące błąd $E_\lambda(\mathbf{w})$:

$$\nabla_{\mathbf{w}} E(\mathbf{w}) = -\boldsymbol{\Phi}^T(\mathbf{y} - \boldsymbol{\Phi}\mathbf{w}) + \lambda\mathbf{w}$$

$$-\boldsymbol{\Phi}^T(\mathbf{y} - \boldsymbol{\Phi}\mathbf{w}) + \lambda\mathbf{w} = 0$$

$$\boldsymbol{\Phi}^T\mathbf{y} - \boldsymbol{\Phi}^T\boldsymbol{\Phi}\mathbf{w} - \lambda\mathbf{w} = 0$$

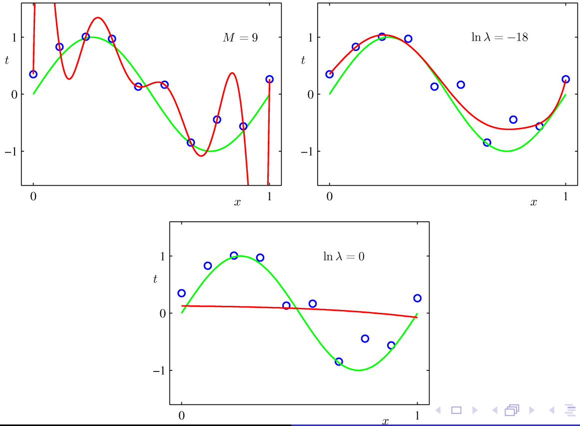
$$\boldsymbol{\Phi}^T\mathbf{y} = \boldsymbol{\Phi}^T\boldsymbol{\Phi}\mathbf{w} + \lambda\mathbf{w}$$

$$\boldsymbol{\Phi}^T\mathbf{y} = (\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \lambda\mathbb{I})\mathbf{w}$$

$$\mathbf{w} = (\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \lambda\mathbb{I})^{-1}\boldsymbol{\Phi}^T\mathbf{y}$$

Regularyzacja ℓ_2 na parametry

- Poszerzenie kryterium uczenia o wyraz mówiący o tym, że wartości parametrów powinny być małe, **bliskie zeru**
- Wprowadzenie parametru regulującego siłę "ściągania" parametrów w kierunku zera



Podział zbioru danych

- 60% - Zbiór **uczący**, na nim dokonujemy znalezienia wartości parametrów w, które optymalizują kryterium jakości
- 20% - Zbiór **walidacyjny**, na nim dokonujemy znalezienia wartości hiperparametrów M oraz λ, hiperparametry wykorzystywane są do stworzenia modelu uczonego na zbiorze uczącym, na zbiorze walidacyjnym testujemy słuszność hiperparametrów
- 20% - Zbiór **testowy**, na nim podajemy ostateczną skuteczność modelu nauczonego na zbiorze uczącym o hiperparametrach sprawdzonych na zbiorze walidacyjnym