Linear regression

Machine Learning 2021-2022 - UMONS
Souhaib Ben Taieb

1

Consider the following optimization problem:

$$(\beta_0^*, \beta_1^*) = \underset{(\beta_0, \beta_1) \in \mathbb{R}^2}{\operatorname{argmin}} E_{\text{out}}(\beta_0, \beta_1) := \mathbb{E}_{x,y}[(y - (\beta_0 + \beta_1 x))^2]$$

where $x, y \in \mathbb{R}$.

Show that the solution is given by

$$\begin{aligned} \beta_1^* &= \frac{\operatorname{Cov}(x,y)}{\operatorname{Var}(x)}, \\ \beta_0^* &= \mathbb{E}[y] - \beta_1 \mathbb{E}[x]. \end{aligned}$$

2

Consider a dataset $\mathscr{D} = \{(x_i, y_i)\}_{i=1}^n$ with $x_i, y_i \in \mathbb{R}$, and the following optimization problem:

$$(\hat{\beta}_0, \hat{\beta}_1) = \underset{(\beta_0, \beta_1) \in \mathbb{R}^2}{\operatorname{argmin}} E_{\text{in}}(\beta_0, \beta_1),$$

where

$$E_{\text{in}}(\beta_0, \beta_1) = \frac{1}{n} \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2.$$

Prove that the minimizing values $\hat{\beta}_1$ and $\hat{\beta}_0$ are given by

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2} \\
\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x},
\end{aligned}$$

where $\bar{y} = \frac{1}{n}\sum_{i=1}^n y_i$ and $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$.

3

We now assume the data has been generated by the following model

$$y_i = f(x_i) + \varepsilon_i,$$

where x_i is fixed (non-random), ε_i are i.i.d. with $E[\varepsilon_i] = 0$, $\mathrm{Var}(\varepsilon_i) = \sigma^2$.

Show that the variance of $\hat{\beta}_1$ is given by $\mathrm{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2}$. You can use the following equalities

$$\begin{aligned}
\sum_i (x_i - \bar{x})(y_i - \bar{y}) &= \sum_i (x_i - \bar{x}) y_i - \sum_i (x_i - \bar{x}) \bar{y} \\
&= \sum_i (x_i - \bar{x}) y_i \\
&= \sum_i (x_i - \bar{x})(\beta_0 + \beta_1 x_i + \varepsilon_i)
\end{aligned}$$

Show that the variance of $\hat{\beta}_0$ is given by

$$\mathrm{Var}(\hat{\beta}_0) = \frac{\sigma^2}{n} + \frac{\bar{x}^2 \sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

4

Under the same set of assumptions as the previous exercise, show that the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased, i.e. $\text{Bias}(\hat{\beta}_0) = 0$ and $\text{Bias}(\hat{\beta}_1) = 0$.