Machine Learning, Spring 2019
EN 2105
MWF 9-9:50am

Chao Lan
EN 4084A
clan@uwyo.edu

From Density Estimation to Linear Regression

Density estimation is equivalent to linear regression estimation under proper assumptions.

1. MLE = Least Square

Let $y = x^T\beta$ be a linear regression model. We will show that the least square estimate of β is equivalent to the MLE estimate of β for a properly designed probabilistic model.

The probabilistic model is constructed by assuming y is a linear function of x plus a random Gaussian noise $\epsilon \sim \mathcal{N}(0, \sigma^2)$, i.e.,

$$y = x^T\beta + \epsilon. \tag{1}$$

If β is fixed (but unknown), then y becomes a random variable with distribution

$$y \sim N(x^T\beta, \sigma^2). \tag{2}$$

[*Exercise*] Verify (2).

If ϵ does not depend on x. The log-likelihood function of β over a sample $y_1, \ldots, y_n$ is

$$\begin{aligned} L_n(\beta) = \sum_{i=1}^n \log p(y_i) &= \sum_{i=1}^n \log \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(x_i^T\beta - y_i)^2\right\} \\ &= \sum_{i=1}^n \log \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2}(x_i^T\beta - y_i)^2 \\ &= n \log \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2}\sum_{i=1}^n (x_i^T\beta - y_i)^2 \\ &= C_n - \frac{1}{2\sigma^2}\sum_{i=1}^n (x_i^T\beta - y_i)^2, \end{aligned} \tag{3}$$

where $C_n = n \log \frac{1}{\sqrt{2\pi\sigma^2}}$.

The MLE of β is the one that maximizes $L_n(\beta)$, i.e.,

$$\hat{\beta}_{mle} = \arg\max_\beta L_n'(\beta) = \arg\min_\beta \sum_{i=1}^n (x_i^T\beta - y_i)^2. \tag{4}$$

[*Exercise*] Verify the last equality in (4).

The right-most side of (4) suggests $\hat{\beta}_{mle}$ is also the least square estimate of β.

[*Discussions*] What assumptions did we make to derive (4)?

2. MLE = Weighted Least Square

In the previous analysis, if ϵ depends on x, then MLE is equivalent to weighted least square.

Let $\epsilon_i \sim N(0, \sigma_i^2)$ be the random noise and

$$y_i = x_i^T\beta + \epsilon_i. \tag{5}$$

Then

$$y_i \sim N(x_i^T\beta, \sigma_i^2). \tag{6}$$

The log-likelihood function of β is now

$$\begin{aligned} L_n(\beta) = \sum_{i=1}^n \log p(y_i) &= \sum_{i=1}^n \log \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left\{-\frac{1}{2\sigma_i^2}(x_i^T\beta - y_i)^2\right\} \\ &= \sum_{i=1}^n \log \frac{1}{\sqrt{2\pi\sigma_i^2}} - \frac{1}{2\sigma_i^2}(x_i^T\beta - y_i)^2 \\ &= \sum_{i=1}^n \log \frac{1}{\sqrt{2\pi\sigma_i^2}} - \sum_{i=1}^n \frac{1}{2\sigma_i^2}(x_i^T\beta - y_i)^2 \\ &= C_n' - \sum_{i=1}^n \frac{1}{2\sigma_i^2}(x_i^T\beta - y_i)^2, \end{aligned} \tag{7}$$

where $C_n' = \sum_{i=1}^n \log \frac{1}{\sqrt{2\pi\sigma_i^2}}$ is independent of β. The MLE of β is then

$$\begin{aligned} \hat{\beta}_{mle} = \arg\max_\beta L_n'(\beta) &= \arg\min_\beta \sum_{i=1}^n \frac{1}{2\sigma_i^2}(x_i^T\beta - y_i)^2 \\ &= \arg\min_\beta \sum_{i=1}^n w_i \cdot (x_i^T\beta - y_i)^2, \end{aligned} \tag{8}$$

where $w_i = \frac{1}{2\sigma_i^2}$ is the weight of (x_i, y_i). It is higher if noise ϵ_i is small (or, σ_i is small). The right-most side of (8) suggests the MLE of β is equivalent to the weighted least square of β.

[*Discussion*] What additional assumption(s) did we make to derive (8)?

3. MAP = Ridge Regression

If there is a prior $p(\beta)$, we can show MAP estimate is equivalent to ridge regression.

Assume prior $\beta \sim \mathcal{N}(0, \Sigma)$. Assume β_j's are i.i.d. from $\mathcal{N}(0, \sigma_2^2)$ and thus $\Sigma = \text{diag}(\sigma_2^2)$.

Let $C_\pi = \log \frac{1}{\sqrt{(2\pi\sigma_2^2)^{p+1}}}$. The log prior of β is

$$\begin{aligned} \log p(\beta) &= \log \frac{1}{\sqrt{(2\pi)^{p+1}|\Sigma|}} \exp\left\{-\frac{1}{2}\beta^T\Sigma^{-1}\beta\right\} = \log \frac{1}{\sqrt{(2\pi\sigma_2^2)^{p+1}}} \exp\left\{-\frac{1}{2\sigma_2^2}\beta^T\beta\right\} \\ &= \log \frac{1}{\sqrt{(2\pi\sigma_2^2)^{p+1}}} - \frac{1}{2\sigma_2^2}\beta^T\beta = C_\pi - \frac{1}{2\sigma_2^2}\beta^T\beta. \end{aligned} \tag{9}$$

The MAP estimate of β is the one that maximizes $\log p(\beta) + L_n(\beta)$, i.e.,

$$\begin{aligned} \hat{\beta}_{map} &= \arg\max_\beta \log p(\beta) + L_n(\beta) \\ &= \arg\max_\beta C_\pi - \frac{1}{2\sigma_2^2}\beta^T\beta + C_n - \frac{1}{2\sigma^2}\sum_{i=1}^n (x_i^T\beta - y_i)^2 \\ &= \arg\min_\beta \lambda\,\beta^T\beta + \sum_{i=1}^n (x_i^T\beta - y_i)^2 \end{aligned} \tag{10}$$

where $\lambda = (\sigma/\sigma_2)^2$ can be interpreted as the regularization coefficient. If σ_2^2 is small (strong belief $\beta_j = 0$), then λ is big and β_j will be small. Note the current analysis also shrinks β_0.

[*Exercise*] How to modify the assumption(s) so we do not shrink β_0?