Optimization and Machine Learning, Spring 2021
Homework 1
(Due Thursday, Mar. 18 at 11:59pm (CST))

March 19, 2021

1. [10 points] Suppose that we have N training data points, which are sampled independently and identically from an unknown joint distribution of p input variables and one continuous output variable. Here the linearity assumption holds between the input variables and the output variable. (Note: you have to consider the intercept in the linear model.)

   (a) Please define the input and output variables, and show a linear relationship between them. [2 points]
   Solution: Define the input variables as a vector $X \in \mathbb{R}^{p+1}$ with an additional constant $X_0 = 1$, and the output label as $Y \in \mathbb{R}$ or $Y \in \{0,\ 1\}$. The linear relationship is $Y = \beta^T X$, where $\beta \in \mathbb{R}^{p+1}$ is the linear coefficients with β_0 denoting the intercept.

   (b) Please define a data matrix and corresponding response/label vector, and find your i-th ($i = 1, ..., N$) sample with its response/label. [2 points]
   Solution: Define a data matrix $\mathbf{X} = \begin{bmatrix} -x_1^T- \\ \vdots \\ -x_N^T- \end{bmatrix}$, and corresponding response vector $\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}$.
   The i-th sample is the i-th row of $\mathbf{X}$ and its response is the i-th element of $\mathbf{y}$.

   (c) Based on the notations defined in (b), please use the least squares method to estimate the parameters of the linear model in (a), and explain in which case the solution is unique. [6 points]
   Solution: Using the least squares, our target is to minimize

   $$RSS(\beta) = (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta).$$

   Differentiating with β and set the derivative to 0, we have

   $$\mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta) = 0.$$

   The unique solution $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ is reachable if and only if $\mathbf{X}^T\mathbf{X}$ is invertible (nonsingular, full-ranked) [Note: other reasonable explanations are acceptable].

2. [10 points] Given the input variables $X \in \mathbb{R}^p$ and output variable $Y \in \mathbb{R}$, the Expected Prediction Error (EPE) is defined by

$$\text{EPE}(\hat{f}) = \mathbb{E}[L(Y, f(X))], \tag{1}$$

where $\mathbb{E}(\cdot)$ denotes the expectation over the joint distribution $\Pr(X, Y)$, and $L(Y, f(X))$ is a loss function measuring the difference between the estimated $f(X)$ and observed Y. We have shown in our course that for the squared error loss $L(Y, f(X)) = (Y - f(X))^2$, the regression function $f(x) = \mathbb{E}(Y|X = x)$ is the optimal solution of $\min_f \text{EPE}(f)$ in the pointwise manner.

(a) Please explain how the nearest neighbors and least squares approximate the regression function, and discuss their difference. [4 points]
Solution:

- The nearest neighbors method $\hat{f}(x) = \frac{1}{k}\sum_{x_i \in N_k(x)} y_i$ has two approximations. The first one is averaging over sample data to appoximate expectation, and the second one is conditioning on neighborhood to approximate conditioning on a point.

- The least square method approximates the theoretical expectation by averaging over the observed data. Using EPE in least squares, we can find the theoretical solution $\beta = \mathbb{E}(XX^T)^{-1}\mathbb{E}(XY)$, and the actual solution for least square is $\beta = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$ which is an approximation for theoretical value.

(b) Given absolute error loss $L(Y, f(X)) = |Y - f(X)|$, please prove that $f(x) = \text{median}(Y|X = x)$ minimizes $\text{EPE}(f)$ w.r.t. f. [6 points]
Solution: The optimization problem is

$$\hat{f}(x) = \underset{f}{\text{argmin}}\, \mathbb{E}_{Y|X}[|Y - f(x)| \mid X = x], \tag{2}$$

$$= \underset{f}{\text{argmin}} \int_y |y - f(x)| \Pr(y|x) dy. \tag{3}$$

where we can obtain the optimal solution according to

$$\frac{\partial}{\partial f} \int_y |y - f(x)| \Pr(y|x) dy = 0. \tag{4}$$

Based on the Law of Large Numbers (LLN), we have

$$\int_y |y - f(x)| \Pr(y|x) dy = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} |y_i - f(x_i)| \approx \frac{1}{n} \sum_{i=1}^{n} |y_i - f(x_i)|. \tag{5}$$

Then, the following equations hold.

$$\frac{\partial}{\partial f} \int_y |y - f(x)| \Pr(y|x) dy = 0 \tag{6}$$

$$\Rightarrow \quad \frac{\partial}{\partial f} \frac{1}{n} \sum_{i=1}^{n} |y_i - f(x_i)| = 0 \tag{7}$$

$$\Rightarrow \quad -\frac{1}{n} \sum_{i=1}^{n} \text{sign}(y_i - f(x_i)) = 0 \tag{8}$$

$$\Rightarrow \quad \sum_{i=1}^{n} \text{sign}(y_i - f(x_i)) = 0. \tag{9}$$

Therefore, we reach the conclusion

$$\hat{f}(x) = \text{median}(Y|X = x). \tag{10}$$

3. [10 points] Given a set of observation pairs $\{(x_1, y_1) \cdots (x_N, y_N)\}$, where $x_i, y_i \in \mathbb{R}$, $i = 1, 2, ..., N$. By assuming the linear model is a reasonable approximation, we consider to fit the model via the least squares method. Thus, our goal is to estimate the coefficients β_0 and β to minimize the Residual Sum of Squares (RSS),

$$\hat{\beta}_0,\ \hat{\beta} = \underset{\beta_0,\ \beta}{\operatorname{argmin}} \sum_{i=1}^{N} (y_i - \beta_0 - \beta x_i)^2.$$

(a) Please show that

$$\begin{aligned} \hat{\beta} &= \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2}, \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}\bar{x}, \end{aligned} \tag{11}$$

where $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$ and $\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$ denote the sample means. [7 points]
Solution: Firstly, we compute β_0

$$\begin{aligned} &\frac{\partial}{\partial \beta_0} \sum_{i=1}^{N} (y_1 - \beta_0 - \beta x_i)^2 = \sum_{i=1}^{N} -2(y_i - \beta_0 - \beta x_i) = 0 \\ &\Rightarrow \sum_{i=1}^{N} (y_i - \beta x_i) = \sum_{i=1}^{N} \beta_0 = N\beta_0 \\ &\Rightarrow \beta_0 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \beta x_i) = \frac{1}{N}\sum_{i=1}^{N} y_i - \beta \frac{1}{N}\sum_{i=1}^{N} x_i = \bar{y} - \beta\bar{x} \\ &\Rightarrow \hat{\beta}_0 = \bar{y} - \hat{\beta}\bar{x}. \end{aligned}$$

Plug β_0 into $\sum_{i=1}^{N}(y_1 - \beta_0 - \beta x_i)^2$ and differentiate with β,

$$\begin{aligned} &\frac{\partial}{\partial \beta} \sum_{i=1}^{N} (y_1 - \beta_0 - \beta x_i)^2 = \frac{\partial}{\partial \beta} \sum_{i=1}^{N} (y_1 - \bar{y} + \beta\bar{x} - \beta x_i)^2 = \sum_{i=1}^{N} 2[y_i - \bar{y} + \beta(\bar{x} - x_i)](\bar{x} - x_i) = 0 \\ &\Rightarrow \sum_{i=1}^{N} (y_i - \bar{y})(\bar{x} - x_i) = -\beta \sum_{i=1}^{N} (\bar{x} - x_i)^2 \\ &\Rightarrow \hat{\beta} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2}. \end{aligned}$$

In conclusion,

$$\hat{\beta} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}\bar{x}.$$

(b) Based on the result in (11), argue that in the case of simple linear regression, the least squares line always passes through the point $(\bar{x}, \bar{y})$. [3 points]
Solution: We can plug $(\bar{x}, \bar{y})$ into the equation $\hat{y} = \hat{\beta}x_i + \beta_0$, and we find $\bar{y} = \hat{\beta}\bar{x} + \bar{y} - \hat{\beta}\bar{x} = \bar{y}$ satisfies. So the least squares line always passes through the point $(\bar{x}, \bar{y})$.