Introduction to Machine Learning, Spring 2022
Reference Solutions of Homework 1

(Due Friday, Mar. 18 at 11:59pm (CST))

February 28, 2022

1. [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

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

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 $\min_f \mathrm{EPE}(f)$ in the pointwise manner.

(a) In Least Squares, a linear model $X^\top \beta$ is used to approximate $f(X)$ according to

$$\min_\beta \mathbb{E}[(Y - X^\top \beta)^2]. \tag{2}$$

Please derive the optimal solution of the model parameters β. [3 points]

Solution:

$$\beta = \mathbb{E}^{-1}[(XX^\top)]\mathbb{E}[(XY)]. \tag{3}$$

(b) Please explain how the nearest neighbors and least squares approximate the regression function, and discuss their difference. [3 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.

(c) Given absolute error loss $L(Y, f(X)) = |Y - f(X)|$, please prove that $f(x) = \mathrm{median}(Y|X = x)$ minimizes $\mathrm{EPE}(f)$ w.r.t. f. [4 points]

Solution: The optimization problem is

$$\hat{f}(x) = \operatorname*{argmin}_f \mathbb{E}_{Y|X}[|Y - f(x)| \mid X = x], \tag{4}$$

$$= \operatorname*{argmin}_f \int_y |y - f(x)|\Pr(y|x)dy. \tag{5}$$

where we can obtain the optimal solution according to

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

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

$$\int_y |y - f(x)|\Pr(y|x)dy = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^n |y_i - f(x_i)| \approx \frac{1}{n}\sum_{i=1}^n |y_i - f(x_i)|. \tag{7}$$

Then, the following equations hold.

$$
\begin{aligned}
& \frac{\partial}{\partial f} \int_y |y - f(x)| \Pr(y|x) dy = 0 && (8) \\
\Rightarrow \quad & \frac{\partial}{\partial f} \frac{1}{n} \sum_{i=1}^{n} |y_i - f(x_i)| = 0 && (9) \\
\Rightarrow \quad & -\frac{1}{n} \sum_{i=1}^{n} \text{sign}(y_i - f(x_i)) = 0 && (10) \\
\Rightarrow \quad & \sum_{i=1}^{n} \text{sign}(y_i - f(x_i)) = 0. && (11)
\end{aligned}
$$

Therefore, we reach the conclusion

$$\hat{f}(x) = \text{median}(Y|X = x). \tag{12}$$

2. [10 points] Consider real-valued variables X and Y, in which Y is generated conditional on X according to

$$Y = aX + b + \epsilon, \text{ where } \epsilon \sim \mathcal{N}(0, \sigma^2).$$

Here ϵ is an independent variable, called a noise term, which is drawn from a Gaussian distribution with mean 0, and variance σ^2. This is a single variable linear regression model, where a is the only weight parameter and b denotes the intercept. The conditional probability of Y has a distribution $p(Y|X, a, b) \sim \mathcal{N}(aX + b, \sigma^2)$, so it can be written as:

$$p(Y|X, a, b) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(Y - aX - b)^2\right).$$

(a) Assume we have a training dataset of n i.i.d. pairs (x_i, y_i), $i = 1, 2, ..., n$, and the likelihood function is defined by $L(a, b) = \prod_{i=1}^{n} p(y_i|x_i, a, b)$. Please write the Maximum Likelihood Estimation (MLE) problem for estimating a and b. [3 points]

Solution:

(a) $\arg\max_a \prod_i \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(y_i - ax_i)^2\right)$

(b) $\arg\max_a \prod_i \exp\left(-\frac{1}{2\sigma^2}(y_i - ax_i)^2\right)$

(c) $\arg\min_a \frac{1}{2} \sum_i (Y_i - aX_i)^2$

All of the above answers are correct.

(b) Estimate the optimal solution of a and b by solving the MLE problem in (a). [4 points]

Solution:

$$\begin{aligned} \hat{a} &= \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2}, \\ \hat{b} &= \bar{y} - \hat{a}\bar{x}, \end{aligned} \tag{13}$$

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.

(c) Based on the result in (b), argue that the learned linear model $f(X) = aX + b$, always passes through the point $(\bar{x}, \bar{y})$, 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. [3 points]

Solution: We can plug $(\bar{x}, \bar{y})$ into the equation $\hat{y} = \hat{a}x_i + \hat{b}$, and we find $\bar{y} = \hat{a}\bar{x} + \bar{y} - \hat{a}\bar{x} = \bar{y}$ satisfies. So the least squares line always passes through the point $(\bar{x}, \bar{y})$.

3. [10 points] Given a set of training data $(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_N, y_N)$ from which to estimate the parameters $\boldsymbol{\beta}$, where each $\mathbf{x}_i = [x_{i1}, \cdots, x_{ip}]^T$ denotes a vector of feature measurements for the ith sample. Consider a linear regression problem in which we want to "weight" different training examples differently. Specifically, suppose we aim at minimizing

$$\text{RSS}(\boldsymbol{\beta}) = \frac{1}{2}\sum_{i=1}^{N} w_i (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2. \tag{14}$$

(a) Show that $\text{RSS}(\boldsymbol{\beta}) = (\mathbf{X}\boldsymbol{\beta} - \mathbf{y})^T \mathbf{W} (\mathbf{X}\boldsymbol{\beta} - \mathbf{y})$ for an appropriate diagonal matrix $\mathbf{W}$, and where $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_N]^T$ and $\mathbf{y} = [y_1, \cdots, y_N]^T$. Please state clearly what $\mathbf{W}$ is. [2 points]
Solution: $\mathbf{W}$ is a diagonal matrix with its i-th diagonal element being $\frac{1}{2}w_i$. Suppose we have the predictions $\hat{\boldsymbol{y}} = \mathbf{X}\beta$, $\text{RSS}(\beta)$ is rewritten by

$$\text{RSS}(\beta) = (\mathbf{X}\beta - \mathbf{y})^T \mathbf{W} (\mathbf{X}\beta - \mathbf{y}) = (\hat{\mathbf{y}} - \mathbf{y})^T \mathbf{W} (\hat{\mathbf{y}} - \mathbf{y})$$

$$\begin{bmatrix} \hat{y}_1 - y_1 \\ \vdots \\ \hat{y}_N - y_N \end{bmatrix}^T \begin{pmatrix} \frac{1}{2}w_1 & 0 & \cdots & 0 & 0 \\ 0 & \frac{1}{2}w_2 & \cdots & 0 & 0 \\ 0 & 0 & \ddots & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{2}w_{N-1} & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{2}w_N \end{pmatrix} \begin{bmatrix} \hat{y}_1 - y_1 \\ \vdots \\ \hat{y}_N - y_N \end{bmatrix}$$

$$= \begin{bmatrix} \frac{1}{2}w_1(\hat{y}_1 - y_1) \\ \vdots \\ \frac{1}{2}w_N(\hat{y}_N - y_N) \end{bmatrix}^T \begin{bmatrix} \hat{y}_1 - y_1 \\ \vdots \\ \hat{y}_N - y_N \end{bmatrix} = \frac{1}{2}\sum_{i=1}^{N} w_i \left(y_i - x_i^T \beta\right)^2.$$

(b) By finding the derivative $\nabla_{\boldsymbol{\beta}}\text{RSS}(\boldsymbol{\beta})$ w.r.t. $\boldsymbol{\beta}$ and setting that to zero, derive the closed-form solution of $\boldsymbol{\beta}$ that minimizes $\text{RSS}(\boldsymbol{\beta})$. [3 points]
Solution:

$$\begin{aligned} \nabla_{\beta}\text{RSS}(\beta) &= \frac{\partial\, \text{RSS}(\beta)}{\partial \beta} \\ &= \frac{\partial}{\partial \beta}(\mathbf{X}\beta - \mathbf{y})^T \mathbf{W} (\mathbf{X}\beta - \mathbf{y}) \\ &= 2\mathbf{X}^T \mathbf{W} (\mathbf{X}\beta - \mathbf{y}) \\ &= 0 \\ \Rightarrow \hat{\boldsymbol{\beta}} &= (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{y}. \end{aligned}$$

(c) Is there any way to control the model complexity in (14)? If yes, please formulate its $\text{RSS}(\boldsymbol{\beta})$ and estimate its closed-form solution of $\boldsymbol{\beta}$. [5 points]
Solution:

$$\text{RSS}(\boldsymbol{\beta}) = \frac{1}{2}\sum_{i=1}^{N} w_i (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 + \lambda \boldsymbol{\beta}^\top \boldsymbol{\beta}. \tag{15}$$

Its optimal solution is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{W} \mathbf{X} + \lambda \mathbf{I}_p)^{-1} \mathbf{X}^T \mathbf{W} \mathbf{y}, \tag{16}$$

where $\mathbf{I}_p$ denotes an identity matrix in size of $p \times p$.