Introduction to Machine Learning, Spring 2022
Homework 1

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

He Haoyu

March 16, 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

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

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) 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]

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

(c) 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. [4 points]

Solution:

(a)

$$\frac{\partial E[(Y - X^T\beta)^2]}{\partial \beta} = -2X^T(Y - X^T\beta)E[(Y - X^T\beta)]$$

Let

$$\frac{\partial E[(Y - X^T\beta)^2]}{\partial \beta} = 0$$

$E[(Y - X^T\beta)^2] > 0$, X^T is a $1 \times p$ matrix, so

$$Y - X^T\beta = 0$$

Thus,

$$\beta = (X^T)^{-1}Y$$

(b) The nearest neighbors: $\hat{Y}(x) = \frac{1}{k}\sum_{x_i \in N_k(x)} y_i$
Firstly, it uses the neighbours information to approximate the current point. Also, it uses the average value instead of the approximate expectation.
The least squares:
Firstly, it replaces the theoretical expection by averaging over the obversed data. By EPE, we know $\beta = E(XX^T)^{-1}E(XY)$, which can be approximate by average $\beta = (XX^T)^{-1}Xy$.

(c) By $L(Y, f(X)) = |Y - f(X)|$, we know

$$\begin{aligned}\hat{f}(x) &= \underset{f}{\text{argmin}} E_{Y|X}[|Y - f(x)||X = x] \\ &= \underset{f}{\text{argmin}} \int_y |y - f(x)| P_r(y|x)\text{d}y\end{aligned}$$

By Law of large numbers, we know

$$\begin{aligned}
\underset{f}{\operatorname{argmin}} E_{Y|X}[|Y - f(x)||X = x] &= \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{N} |y_i - f(x_i)| \\
&\approx \frac{1}{n} \sum_{i=1}^{N} |y_i - f(x_i)| (\text{when n is large})
\end{aligned}$$

Thus,

$$\begin{aligned}
\underset{f}{\operatorname{argmin}} E_{Y|X}[|Y - f(x)||X = x] &= \underset{f}{\operatorname{argmin}} \int_y |y - f(x)| P_r(y|x) \mathrm{d}y \\
&= \frac{1}{n} \sum_{i=1}^{N} |y_i - f(x_i)|
\end{aligned}$$

Then, use partial to get optimal f

$$\begin{aligned}
&\frac{\partial \underset{f}{\operatorname{argmin}} \int_y |y - f(x)| P_r(y|x) \mathrm{d}y}{\partial f} = 0 \\
&\Rightarrow \frac{\partial \frac{1}{n} \sum_{i=1}^{N} |y_i - f(x_i)|}{\partial f} = 0 \\
&\Rightarrow \sum_{i=0}^{N} \mathbf{sign}(y_i - f(x_i)) = 0
\end{aligned}$$

Thus, we know

$$f(x) = \operatorname{median}(Y|X = x)$$

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]

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

(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]

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

$$\mathrm{RSS}(\boldsymbol{\beta}) = \frac{1}{2}\sum_{i=1}^{N} w_i (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2. \tag{3}$$

(a) Show that $\mathrm{RSS}(\boldsymbol{\beta}) = (\mathbf{X}\boldsymbol{\beta} - \mathbf{y})^T \mathbf{W} (\mathbf{X}\boldsymbol{\beta} - \mathbf{y})$ 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]

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

(c) Is there any way to control the model complexity in (3)? If yes, please formulate the $\mathrm{RSS}(\boldsymbol{\beta})$ and estimate its closed-form solution of $\boldsymbol{\beta}$. [5 points]

Solution:

(a) First,

$$W = \begin{bmatrix} \frac{1}{2}w_1 & 0 & \cdots & 0 \\ 0 & \frac{1}{2}w_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{2}w_N \end{bmatrix}$$

The proof is as following:

$$\begin{aligned} RSS(\beta) &= (X\beta - y)^T W (X\beta - y) \\ &= \begin{bmatrix} y_1 - x_1^T\beta & y_2 - x_2^T\beta & \cdots & y_N - x_N^T\beta \end{bmatrix} \begin{bmatrix} \frac{1}{2}w_1 & 0 & \cdots & 0 \\ 0 & \frac{1}{2}w_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{2}w_N \end{bmatrix} \begin{bmatrix} y_1 - x_1^T\beta \\ y_2 - x_2^T\beta \\ \vdots \\ y_N - x_N^T\beta \end{bmatrix} \\ &= \frac{1}{2}\sum_{i=1}^{N} w_i (y_i - x_i^T\beta)^2 \end{aligned}$$

Thus,

$$W = \begin{bmatrix} \frac{1}{2}w_1 & 0 & \cdots & 0 \\ 0 & \frac{1}{2}w_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{2}w_N \end{bmatrix}$$

(b)

$$\begin{aligned} \nabla_\beta RSS(\beta) &= \nabla_\beta (X\beta - y)^T W (X\beta - y) \\ &= 2X^T W (X\beta - y) \end{aligned}$$

Setting that to zero,

$$\nabla_\beta RSS(\beta) = 0$$

Then,

$$\begin{aligned} &X^T W (X\beta - y) = (X^T W X\beta - X^T W y) = 0 \\ &\Rightarrow \hat{\beta} = (X^T W X)^{-1} X^T W y \end{aligned}$$

(c) Like Shrinkage methods-Ridge Regression, we can impose a penalty on the size.

$$\hat{\beta} = \underset{\beta}{\mathrm{argmin}} \left\{ \frac{1}{2}\sum_{i=1}^{N} w_i \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

Using the same way in (b), we get

$$(X^T W X\beta - X^T W y) + \lambda\beta = 0$$

Thus,

$$\hat{\beta} = (X^T W X + \lambda I_p)^{-1} X^T W y$$