

# Introduction to Machine Learning, Spring 2022

## Homework 1

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

He Haoyu

March 9, 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

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

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

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

Please derive the optimal solution of the model parameters  $\beta$ . [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)

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

Let

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

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

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

Thus,

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

- (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:

(c)

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  $\epsilon$  is an independent variable, called a noise term, which is drawn from a Gaussian distribution with mean 0, and variance  $\sigma^2$ . 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  $(x_i, y_i)$ ,  $i = 1, 2, \dots, n$ , 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  $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_N, y_N)$  from which to estimate the parameters  $\boldsymbol{\beta}$ , where each  $\mathbf{x}_i = [x_{i1}, \dots, x_{ip}]^T$  denotes a vector of feature measurements for the  $i$ th sample. Consider a linear regression problem in which we want to “weight” different training examples differently. Specifically, suppose we aim at minimizing

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

- (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  $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_N]^T$  and  $\mathbf{y} = [y_1, \dots, y_N]^T$ . Please state clearly what  $\mathbf{W}$  is. [2 points]
- (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]
- (c) Is there any way to control the model complexity in (3)? If yes, please formulate the  $\text{RSS}(\boldsymbol{\beta})$  and estimate its closed-form solution of  $\boldsymbol{\beta}$ . [5 points]