Introduction to Machine Learning, Spring 2023
Homework 1

(Due Friday, Mar. 7 at 11:59pm (CST))

February 21, 2023

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]

   (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 $f(x) = \mathrm{median}(Y|X = x)$ minimizes $\mathrm{EPE}(f)$ w.r.t. f. [4 points]

2. [10 points]

(a) Ridge regression can be considered as an unconstrained optimization problem

$$\min_{\mathbf{w}} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 - \lambda\|\mathbf{w}\|_2^2. \tag{3}$$

where $\mathbf{X} \in \mathbb{R}^{n\times d}$ is a data matrix, and $\mathbf{y} \in \mathbb{R}^n$ is the target vector. Consider the following augmented target vector $\hat{\mathbf{y}}$ and data matrix $\hat{\mathbf{X}}$

$$\hat{\mathbf{y}} = \begin{bmatrix} \mathbf{y} \\ \mathbf{0}_d \end{bmatrix} \hat{\mathbf{X}} = \begin{bmatrix} \mathbf{X} \\ \sqrt{\lambda}\mathbf{I}_d \end{bmatrix}$$

where $\mathbf{0}_d$ is the zero vector in $\mathbb{R}^d$ and $\mathbf{I}_d \in \mathbb{R}^{d\times d}$ is an identity matrix. Please derive the optimal solution of the optimization problem $\min_\omega \|\hat{\mathbf{y}} - \hat{\mathbf{X}}\mathbf{w}\|_2^2$ only use $\mathbf{X}, \mathbf{y}$. [3 points]

(b) Let's consider another situation by constructing an augmented matrix in the following way

$$\hat{\mathbf{X}} = \begin{bmatrix} \mathbf{X} & \alpha\mathbf{I}_n \end{bmatrix}$$

where α is a scalar multiplier. Then consider the following problem

$$\min_{\beta} \|\beta\|_2^2 \quad \text{s.t. } \hat{\mathbf{X}}\beta = \mathbf{y} \tag{4}$$

If β^* is the optimal solution of (4), show that the first d coordinates of β^* form the optimal solution of (3) for a specific α, and find the α. And What the final n coordinates of β^* represent? [3 points]

(c) As we all know, the standard formula for Ridge Regression is the optimal solution of (3).
Suppose the SVD of $\mathbf{X}$ is $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{T}}$, then we can make some changes on coordinates in the feature space, so that $\mathbf{V}$ becomes identity, where $\mathbf{X}' = \mathbf{X}\mathbf{V}$ and $\mathbf{w}' = \mathbf{V}^{\mathrm{T}}\mathbf{w}$, and denote $\hat{\mathbf{w}}'$ as the solution of the ridge regression in new coordinates. Please write down the i-th coordinate of $\hat{\mathbf{w}}'$. (Hints: try to use σ_i to represent the i-th singular value of $\mathbf{X}$) [4 points]

3. [10 points] A random variable $\mathbf{X}$ has unknown mean and variance: μ, σ^2. n iid realizations $\mathbf{X}_1 = \mathbf{x}_1, \mathbf{X}_2 = \mathbf{x}_2, \cdots, \mathbf{X}_n = \mathbf{x}_n$from the random variable $\mathbf{X}$ are used to estimate the mean of $\mathbf{X}$. We will call our estimate of μ the random variable $\hat{\mathbf{X}}$, which has mean $\hat{\mu}$. There are two possible ways to estimate μ with the realizations of n samples:

   1. Average the n samples: $\frac{\mathbf{x}_1+\mathbf{x}_2+\cdots+\mathbf{x}_n}{n}$
   2. Average the n samples and n_0 samples of 0: $\frac{\mathbf{x}_1+\mathbf{x}_2+\cdots+\mathbf{x}_n}{n+n_0}$

   The bias is defined as $\mathbb{E}[\hat{\mathbf{X}} - \mu]$ and the variance of $Var[\hat{\mathbf{X}}]$

   (a) What are the bias and the variance of each of the two estimators above? [2 points]

   (b) Now we denote a new independent sample of $\mathbf{X}$ as $\mathbf{X}'$, in order to test how well $\hat{\mathbf{X}}$ estimates a new sample of $\mathbf{X}$. Please derive an expression for $\mathbb{E}[(\hat{\mathbf{X}} - \mu)^2]$ and $\mathbb{E}[(\hat{\mathbf{X}} - \mathbf{X}')^2]$, and then make some comments on the differences between them. (Hints: Using the Bias-Variance Tradeoff) [6 points]

   (c) Compute $\mathbb{E}[(\hat{\mathbf{X}} - \mu)^2]$ for each of the estimators above. [2 points]