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]

Solution:

$$\beta = \mathbb{E}^{-1}[(\mathbf{X}\mathbf{X}^{\mathrm{T}})]\mathbb{E}[(\mathbf{X}\mathbf{Y})]$$

(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 $\hat{f}(x) = \frac{1}{k}\sum_{x_i \in N_k(x)} y_i$ as two approximations. The first one is averaging over sample data to approximate 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 $\beta = \mathbb{E}^{-1}[(\mathbf{X}\mathbf{X}^{\mathrm{T}})]\mathbb{E}[(\mathbf{X}\mathbf{Y})]$, and the actual solution for least square is $\beta = (\mathbf{X}\mathbf{X}^{\mathrm{T}})^{-1}\mathbf{X}\mathbf{Y}$ 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

$$\begin{aligned}\hat{f}(x) &= \arg\min_f \mathbb{E}_{Y|X}[|Y - f(x)||X = x] \\ &= \arg\min_f \int_y |y - f(x)| Pr(y|x) dy\end{aligned}$$

where we can obtain the optimal solution according to

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

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

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

Then, the following equations hold

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

Therefore, we reach the conclusion (x) = median(Y$|X = x$)

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]

Solution:

For $\arg\min_{\mathbf{w}} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \lambda\|\mathbf{w}\|_2^2$

$$\begin{aligned}
\frac{\partial}{\partial \mathbf{w}}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 - \lambda\|\mathbf{w}\|_2^2 &= -2\mathbf{X}^{\mathrm{T}}\mathbf{y} + 2\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{w} + 2\lambda\mathbf{w} = 0 \\
&\Rightarrow (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I}_d)\mathbf{w} = \mathbf{X}^{\mathrm{T}}\mathbf{y} \\
&\Rightarrow \mathbf{w} = (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I}_d)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}
\end{aligned}$$

For $\arg\min_{\mathbf{w}} \|\hat{\mathbf{y}} - \hat{\mathbf{X}}\mathbf{w}\|_2^2$

$$\begin{aligned}
\frac{\partial}{\partial \omega}\|\hat{\mathbf{y}} - \hat{\mathbf{X}}\mathbf{w}\|^2 &= -2\hat{\mathbf{X}}^{\mathrm{T}}\hat{\mathbf{y}} + 2\hat{\mathbf{X}}^{\mathrm{T}}\hat{\mathbf{X}}\mathbf{w} = 0 \\
&\Rightarrow \mathbf{w} = (\hat{\mathbf{X}}^{\mathrm{T}}\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}^{\mathrm{T}}\hat{\mathbf{y}}
\end{aligned}$$

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

$$\begin{aligned}
\hat{\mathbf{X}}^{\mathrm{T}}\hat{\mathbf{X}} &= \begin{bmatrix} \mathbf{X}^{\mathrm{T}} & \sqrt{\lambda}\mathbf{I}_d \end{bmatrix} \begin{bmatrix} \mathbf{X} \\ \sqrt{\lambda}\mathbf{I}_d \end{bmatrix} = \mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I}_d \\
\hat{\mathbf{X}}^{\mathrm{T}}\hat{\mathbf{y}} &= \begin{bmatrix} \mathbf{X}^{\mathrm{T}} & \sqrt{\lambda}\mathbf{I}_d \end{bmatrix} \begin{bmatrix} \mathbf{y} \\ \mathbf{0}_d \end{bmatrix} = \mathbf{X}^{\mathrm{T}}\mathbf{y}
\end{aligned}$$

Therefore,

$$\mathbf{w} = (\hat{\mathbf{X}}^{\mathrm{T}}\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}^{\mathrm{T}}\hat{\mathbf{y}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I}_d)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

(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]

Solution:

Solve the optimal solution β^*

$$\begin{aligned}
\mathcal{L}(\beta, \lambda) &= \|\beta\|_2^2 + \lambda^{\mathrm{T}}(\hat{\mathbf{X}}\beta - \mathbf{y}) \\
\frac{\partial \mathcal{L}(\beta, \lambda)}{\partial \beta} &= 2\beta + \hat{\mathbf{X}}^{\mathrm{T}}\lambda = 0 \\
\beta &= -\frac{1}{2}\hat{\mathbf{X}}^{\mathrm{T}}\lambda \\
g(\lambda) = \inf_{\beta} \mathcal{L}(\beta, \lambda) &= \frac{1}{4}\|\hat{\mathbf{X}}^{\mathrm{T}}\lambda\|_2^2 + \lambda^{\mathrm{T}}(-\frac{\hat{\mathbf{X}}\hat{\mathbf{X}}^{\mathrm{T}}\lambda}{2} - \mathbf{y})
\end{aligned}$$

Then we can change the original optimization problem into $\min_{\lambda} \frac{1}{4}\|\hat{\mathbf{X}}^{\mathrm{T}}\lambda\|_2^2 + \lambda^{\mathrm{T}}\mathbf{y}$

$$
\begin{aligned}
\frac{\partial}{\partial \lambda}\frac{1}{4}\|\hat{\mathbf{X}}^{\mathrm{T}}\lambda\|_2^2 + \lambda^{\mathrm{T}}\mathbf{y} = \frac{1}{2}\hat{\mathbf{X}}\hat{\mathbf{X}}^{\mathrm{T}}\lambda + \mathbf{y} = 0 \\
\Rightarrow \lambda^* = -2(\hat{\mathbf{X}}\hat{\mathbf{X}}^{\mathrm{T}})^{-1}\mathbf{y} \\
\Rightarrow \beta^* = \hat{\mathbf{X}}^{\mathrm{T}}(\hat{\mathbf{X}}\hat{\mathbf{X}}^{\mathrm{T}})^{-1}\mathbf{y}
\end{aligned}
$$

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

$$
\begin{aligned}
\beta^* &= \begin{bmatrix}\mathbf{X}^{\mathrm{T}} \\ \alpha\mathbf{I}_n\end{bmatrix} \begin{bmatrix}\mathbf{X} & \alpha\mathbf{I}_n\end{bmatrix} \begin{bmatrix}\mathbf{X}^{\mathrm{T}} \\ \alpha\mathbf{I}_n\end{bmatrix} \mathbf{y} \\
&= \begin{bmatrix}\mathbf{X}^{\mathrm{T}} \\ \alpha\mathbf{I}_n\end{bmatrix} (\mathbf{X}\mathbf{X}^{\mathrm{T}} + \alpha^2\mathbf{I}_n)^{-1}\mathbf{y}
\end{aligned}
$$

The first d coordinates of β^* is $\mathbf{X}^{\mathrm{T}}(\mathbf{X}\mathbf{X}^{\mathrm{T}} + \alpha^2\mathbf{I}_n)^{-1}\mathbf{y}$. Therefore, when $\alpha = \sqrt{\lambda}$, the first d coordinates of β^* form the optimal solution of (3). The final n coordinates of β^* represent the parameters of some fake features.

(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]

Solution:

$$
\begin{aligned}
\hat{\mathbf{w}} &= (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} \\
&= (\mathbf{V}\mathbf{\Sigma}^{\mathrm{T}}\mathbf{U}^{\mathrm{T}}\mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{T}} + \lambda\mathbf{I})^{-1}\mathbf{V}\mathbf{\Sigma}^{\mathrm{T}}\mathbf{U}^{\mathrm{T}}\mathbf{y} \\
&= (\mathbf{V}\mathbf{\Sigma}^{\mathrm{T}}\mathbf{\Sigma}\mathbf{V}^{\mathrm{T}} + \lambda\mathbf{I})^{-1}\mathbf{V}\mathbf{\Sigma}^{\mathrm{T}}\mathbf{U}^{\mathrm{T}}\mathbf{y} \\
&= \mathbf{V}(\mathbf{\Sigma}^{\mathrm{T}}\mathbf{\Sigma} + \lambda\mathbf{I})^{-1}\mathbf{V}^{\mathrm{T}}\mathbf{V}\mathbf{\Sigma}^{\mathrm{T}}\mathbf{U}^{\mathrm{T}}\mathbf{y} \\
&= \mathbf{V}(\mathbf{\Sigma}^{\mathrm{T}}\mathbf{\Sigma} + \lambda\mathbf{I})^{-1}\mathbf{\Sigma}^{\mathrm{T}}\mathbf{U}^{\mathrm{T}}\mathbf{y} \\
\Rightarrow \mathbf{V}^{\mathrm{T}}\hat{\mathbf{w}} &= (\mathbf{\Sigma}^{\mathrm{T}}\mathbf{\Sigma} + \lambda\mathbf{I})^{-1}\mathbf{\Sigma}^{\mathrm{T}}(\mathbf{U}^{\mathrm{T}}\mathbf{y}) \\
\hat{\mathbf{w}}' &= (\mathbf{\Sigma}^{\mathrm{T}}\mathbf{\Sigma} + \lambda\mathbf{I})^{-1}\mathbf{\Sigma}^{\mathrm{T}}(\mathbf{U}^{\mathrm{T}}\mathbf{y}) \\
\hat{\mathbf{w}}'[i] &= \frac{\sigma_i}{\sigma_i^2 + \lambda}(\mathbf{U}^{\mathrm{T}}\mathbf{y})[i]
\end{aligned}
$$

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]

Solution:

$\mathbb{E}[\hat{\mathbf{X}} - \mu] = \mathbb{E}[\hat{\mathbf{X}}] - \mu$, so we have the following biases:

- $\mathbb{E}[\hat{\mathbf{X}}] = \mathbb{E}[\frac{\mathbf{X}_1+\mathbf{X}_2+\cdots+\mathbf{X}_n}{n}] = \frac{n\mu}{n} \Rightarrow bias = 0$
- $\mathbb{E}[\hat{\mathbf{X}}] = \mathbb{E}[\frac{\mathbf{X}_1+\mathbf{X}_2+\cdots+\mathbf{X}_n}{n+n_0}] = \frac{n\mu}{n+n_0} \Rightarrow bias = -\frac{n_0}{n+n_0}\mu$

Variances:

- $Var[\hat{\mathbf{X}}] = Var[\frac{\mathbf{X}_1+\mathbf{X}_2+\cdots+\mathbf{X}_n}{n}] = \frac{1}{n^2}Var[\mathbf{X}_1 + \mathbf{X}_2 + \cdots + \mathbf{X}_n] = \frac{1}{n^2}(n\sigma^2) = \frac{\sigma^2}{n}$
- $Var[\hat{\mathbf{X}}] = Var[\frac{\mathbf{X}_1+\mathbf{X}_2+\cdots+\mathbf{X}_n}{n+n_0}] = \frac{1}{(n+n_0)^2}Var[\mathbf{X}_1 + \mathbf{X}_2 + \cdots + \mathbf{X}_n] = \frac{1}{(n+n_0)^2}(n\sigma^2) = \frac{n\sigma^2}{(n+n_0)^2}$

(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]

Solution:

$$
\begin{aligned}
\mathbb{E}[(\hat{\mathbf{X}} - \mathbf{X}')^2] &= \mathbb{E}[(\hat{\mathbf{X}} - \mu + \mu - \mathbf{X}')^2] \\
&= \mathbb{E}[(\hat{\mathbf{X}} - \mu)^2] +^2 \\
&= \mathbb{E}[(\hat{\mathbf{X}} - \mathbb{E}[(\hat{\mathbf{X}}] + \mathbb{E}[(\hat{\mathbf{X}}] - \mu)^2] + \sigma^2 \\
&= \mathbb{E}[(\hat{\mathbf{X}} - \mathbb{E}[(\hat{\mathbf{X}}])^2] + \mathbb{E}[(\mathbb{E}[(\hat{\mathbf{X}}] - \mu)^2] + 2\mathbb{E}[\hat{\mathbf{X}} - \mathbb{E}[\hat{\mathbf{X}}]] \cdot \mathbb{E}[\mathbb{E}[\hat{\mathbf{X}}] - \mu] + \sigma^2 \\
&= Var[\hat{\mathbf{X}}] + bias^2 + \sigma^2 \\
\mathbb{E}[(\hat{\mathbf{X}} - \mu)^2] &= \mathbb{E}[\hat{\mathbf{X}}^2] + \mathbb{E}[\mu^2] - 2\mathbb{E}[\hat{\mathbf{X}}\mu] \\
&= (Var[\hat{\mathbf{X}}] + \mathbb{E}[\hat{\mathbf{X}}]^2) + (Var[\mu] + \mathbb{E}[\mu]^2) - 2\mathbb{E}[\hat{\mathbf{X}}\mu] \\
&= (\mathbb{E}[\hat{\mathbf{X}}]^2 + \mathbb{E}[\mu] - 2\mathbb{E}[\hat{\mathbf{X}}\mu]) + Var[\hat{\mathbf{X}}] + Var[\mu] \\
&= Var[\hat{\mathbf{X}}] + bias^2
\end{aligned}
$$

Notice that these two expected squared errors resulted in the same expressions except for the σ^2 in $\mathbb{E}[(\hat{\mathbf{X}} - \mathbf{X}')^2]$. The error σ^2 is considered "irreducible error" because it is associated with the noise that comes from sampling from the distribution of $\mathbf{X}$. This term is not present in the second derivation because μ is a fixed value that we are trying to estimate.

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

Solution:

$\mathbb{E}[(\hat{\mathbf{X}} - \mu)^2] = (\mathbb{E}[\hat{\mathbf{X}} - \mu])^2 + Var[\hat{\mathbf{X}} - \mu]$, so we have the following biases:

- $\mathbb{E}[(\hat{\mathbf{X}} - \mu)^2] = \frac{\sigma^2}{n}$
- $\mathbb{E}[(\hat{\mathbf{X}} - \mu)^2] = \frac{1}{(n+n_0)^2}(n_0^2\mu^2 + n\sigma^2)$