# Intro to Big Data Science: Assignment 2

Due Date: March 25, 2025

✏ **Exercise 1: (Maximum Likelihood Estimate)**

Suppose that the samples $\{x_i\}_{i=1}^n$ are drawn from Normal distribution $\mathcal{N}(\mu, \sigma^2)$ with p.d.f. $f_\theta(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{1}{2\sigma^2}(x-\mu)^2)$, where $\theta = (\mu, \sigma^2)$. The Maximum likelihood estimator (MLE) of $\theta$ is the one that maximize the likelihood function

$$L(\theta) = \prod_{i=1}^n f_\theta(x_i)$$

1. Show that the MLE estimator of the parameters $(\mu, \sigma^2)$ is

$$\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^n x_i, \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2.$$

2. Show that

$$\mathrm{E}\hat{\mu} = \mu, \qquad \mathrm{E}\Big(\frac{n}{n-1}\hat{\sigma}^2\Big) = \sigma^2,$$

where E is the expectation. This means that $\hat{\mu}$ is an unbiased estimator of $\mu$, but $\hat{\sigma}^2$ is a biased estimator of $\sigma^2$.

✏ **Exercise 2 (Linear regression)** Consider fitting the linear regression model for the data

| x | -1 | 0 | 2 |
|---|---|---|---|
| y | 1 | -1 | 1 |

1. Fit $y_i = w_0 + \epsilon_i$ (degenerated linear regression), find $w_0$.

2. Fit $y_i = w_1 x_i + \epsilon_i$ (linear regression without constant term), find $w_1$.
3. Fit $y_i = w_0 + w_1 x_i + \epsilon_i$ (full linear regression), find $w_0$ and $w_1$.
4. Repeat 3 by using ridge regression with hyperparameter $\lambda = 1$.

✏ **Exercise 3 (Properties of Linear regression)**

Consider a multivariate liner model $\mathbf{y} = \mathbf{X}\mathbf{w} + \epsilon$ with $\mathbf{y} \in \mathbb{R}^{n\times 1}$, $\mathbf{X} \in \mathbb{R}^{n\times(d+1)}$, $\mathbf{w} \in \mathbb{R}^{(d+1)\times 1}$, and $\epsilon \in \mathbb{R}^{n\times 1}$, where $\epsilon \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$, follows the normal distribution, where $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n)$, follows the normal distribution, where $\mathbf{I}_n$ is the $n \times n$ identity matrix. Suppose $\mathbf{X}$ and $\mathbf{y}$ are given.

1. Show that the least square linear regression predictor is given by $\hat{\mathbf{y}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$.
2. Show that $\hat{\mathbf{w}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ is an unbiased estimator of $\mathbf{w}$, i.e., $\mathbf{E}(\hat{\mathbf{w}}) = \mathbf{w}$. Also show that $\mathrm{Var}(\hat{\mathbf{w}}) = (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2$. (Note that by definition, $\mathrm{Var}(\hat{\mathbf{w}}) = \mathbf{E}[(\hat{\mathbf{w}} - \mathbf{E}(\hat{\mathbf{w}}))(\hat{\mathbf{w}} - \mathbf{E}(\hat{\mathbf{w}}))^T]$).
3. Let $\mathbf{P} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$, show that $\mathbf{P}$ has only 0 and 1 eigenvalues.
4. Recall the definition of $R^2$ score: $R^2 := 1 - \frac{SS_{res}}{SS_{tot}}$, where $SS_{tot} = \sum_{i=1}^n (y_i - \bar{y})^2$, $SS_{reg} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$, and $SS_{res} = \sum_{i=1}^n (y_i - \hat{y}_i)^2$. Prove that for linear regression, $SS_{tot} = SS_{reg} + SS_{res}$. (So that $R^2$ score can also be defined as $R^2 = \frac{SS_{reg}}{SS_{tot}}$)
5. Now we want to use ridge regression with a tuning parameter $\lambda > 0$ to estimate $\mathbf{w}$. The ridge regression estimator is given by $\hat{\mathbf{w}} = (\mathbf{X}^T\mathbf{X} + \lambda \mathbf{I}_d)^{-1}\mathbf{X}^T\mathbf{y}$. Is $\hat{\mathbf{w}}$ unbiased, i.e., $\mathrm{E}\hat{\mathbf{w}} = \mathbf{w}$? Prove your result.
6. Show that the ridge regression predictor is given by $\hat{\mathbf{y}} = \mathbf{Q}\mathbf{y}$, where $\mathbf{Q} = \mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda \mathbf{I}_d)^{-1}\mathbf{X}^T$. Please also compute the limit $\lim_{k\to\infty} \mathbf{Q}^k$.
7. Discuss the influence of the hyper-parameter $\lambda$: what happens to the bias and the variance of the estimator $\hat{\mathbf{y}}$ as $\lambda \to 0$ or $\infty$?

✏ **Exercise 4 (Generalized Cross-Validation, Optional)** Consider ridge regression:

$$\min_{\mathbf{w}} \left[ (\mathbf{y} - \mathbf{X}\mathbf{w})^T (\mathbf{y} - \mathbf{X}\mathbf{w}) + \lambda \|\mathbf{w}\|_2^2 \right]$$

It has the solution $\hat{\mathbf{w}} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$ and prediction $\hat{\mathbf{y}} = \mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{P}\mathbf{y}$ with $\mathbf{P} = \mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T$ be the projection matrix.

1. Define the leave-one-out cross validation estimator as

   $$\hat{\mathbf{w}}^{[k]} = \arg\min_{\mathbf{w}} \Big[ \sum_{i=1, i\neq k}^{n} (y_i - \mathbf{x}_i^T \mathbf{w})^2 + \lambda \|\mathbf{w}\|_2^2 \Big].$$

   Show that $\hat{\mathbf{w}}^{[k]} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I} - \mathbf{x}_k\mathbf{x}_k^T)^{-1}(\mathbf{X}^T\mathbf{y} - \mathbf{x}_k y_k)$

2. Define the ordinary cross-validation (OCV) mean squared error as $V_0(\lambda) = \frac{1}{n}\sum_{k=1}^{n} (\mathbf{x}_k^T \hat{\mathbf{w}}^{[k]} - y_k)^2$. Show that $V_0(\lambda)$ can be rewritten as $V_0(\lambda) = \frac{1}{n}\sum_{k=1}^{n} \left(\frac{\hat{y}_k - y_k}{1 - p_{kk}}\right)^2$, where $\hat{y}_k = \sum_{j=1}^n p_{kj} y_j$ and $p_{kj}$ is the $(k, j)$-entry of $\mathbf{P}$.

(Hint: You may need to use the Sherman-Morrison Formula for nonsingualar matrix $\mathbf{A}$ and vectors $\mathbf{x}$ and $\mathbf{y}$ with $\mathbf{y}^T\mathbf{A}^{-1}\mathbf{x} \neq -1$: $(\mathbf{A}+\mathbf{x}\mathbf{y}^T)^{-1} = \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1}\mathbf{x}\mathbf{y}^T\mathbf{A}^{-1}}{1+\mathbf{y}^T\mathbf{A}^{-1}\mathbf{x}}$)

3. Define weights as $w_k = \left(\frac{1-p_{kk}}{\frac{1}{n}tr(\mathbf{I}-\mathbf{P})}\right)^2$ and weighted OCV as $V(\lambda) = \frac{1}{n}\sum_{k=1}^{n} w_k(\mathbf{x}_k^T\hat{\mathbf{w}}^{[k]} - y_k)^2$. Show that $V(\lambda)$ can be written as

$$V(\lambda) = \frac{\frac{1}{n}\|(\mathbf{I}-\mathbf{A})\mathbf{y}\|^2}{\left[1 - tr(\mathbf{P})/n\right]^2}$$

✏ **Exercise 5** Online study and exercises.