MDS5210 · Homework 1

Due: (23:59), March 3

Instructions:

- Homework problems must be carefully and clearly answered to receive full credit. Complete sentences that establish a clear logical progression are highly recommended.
- You must submit your assignment in Blackboard. Please upload a file or a zip file. The file name should be in the format **last name-first name-hw1**.
- The homework must be written in English.
- Late submission will not be graded.
- Each student **must not copy** homework solutions from another student or from any other source.

Problem 1 (30pts). Fundamental Knowledge

(1) Clearly state the difference between supervised learning and unsupervised learning.

(2) Explain the usage of training set, validation set, and test set in a learning task. Also explain why we need a validation set.

(3) Suppose we have a dataset of people in which we record their heights h_i, as well as length of left arms l_i, and right arms r_i. Suppose $h_i \sim \mathcal{N}(10, 2)$ (in unspecified units), and $l_i \sim \mathcal{N}(\rho_i h_i, 0.02)$ and $r_i \sim \mathcal{N}(\rho_i h_i, \sigma^2)$, with $\rho_i \sim \text{Unif}(0.6, 0.7)$. Is using both arms necessarily a better choice than using only one arm to approximate h_i? What if $\sigma^2 = 0.02$? Explain the intuition.

(4) Let $\mathbf{X} \in \mathbb{R}^{n\times d}$ be a full column rank matrix. Explain why $\mathbf{X}^T\mathbf{X}$ is positive definite using SVD. (Hint: The singular matrices are orthonormal)

(5) Consider the problem of

$$\min_{\boldsymbol{\theta}} \|\mathbf{X}\boldsymbol{\theta} - \mathbf{y}\|_2^2 + \lambda\|\boldsymbol{\theta}\|_2^2.$$

Suppose $\mathbf{X}$ is full column rank, write down its optimal solution $\boldsymbol{\theta}^*$.

Problem 2 (30pts). Least Square without Full Column Rank
Consider the problem

$$\min_{\boldsymbol{\theta}} \|\mathbf{X}\boldsymbol{\theta} - \mathbf{y}\|_2^2,$$

where $X \in \mathbb{R}^{n\times d}, \boldsymbol{\theta} \in \mathbb{R}^d, \mathbf{y} \in \mathbb{R}^n$.

(1) Given

$$\mathbf{X} = \begin{bmatrix} 1 & 0 \end{bmatrix}, \mathbf{y} = \begin{bmatrix} 1 \end{bmatrix}.$$

Draw the figure of the objective function using python.

(2) The thin SVD of $\mathbf{X}$ is given by

$$\mathbf{X} = \mathbf{V} \begin{bmatrix} \boldsymbol{\Sigma}_1 & 0 \end{bmatrix} \begin{bmatrix} \mathbf{U}_1^T \\ \mathbf{U}_2^T \end{bmatrix} = \mathbf{V}\boldsymbol{\Sigma}_1\mathbf{U}_1^T.$$

Show that when $n < d$, optimal solutions are non-unique. Derive the expression of the optimal solutions using thin SVD. (Hint: Let $\mathbf{A} := \mathbf{V}\boldsymbol{\Sigma}_1, \mathbf{z} := \mathbf{U}_1^T\boldsymbol{\theta}$. Solve $\|\mathbf{A}\mathbf{z} - \mathbf{y}\|_2^2$ first, then solve $\mathbf{U}_1^T\boldsymbol{\theta} = \mathbf{z}$)

Problem 3 (50pts). A Robust LP Formulation

Suppose we have the generative linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta}^{\star} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\epsilon}$ is the error term and $\boldsymbol{\epsilon} \sim N(0, \Sigma)$. The maximum likelihoog estimator for $\boldsymbol{\theta}$ is:

$$\begin{aligned}\hat{\boldsymbol{\theta}}_{LS} &= \underset{\boldsymbol{\theta}\in\mathbb{R}^d}{\operatorname{argmin}} \ \|\mathbf{X}\boldsymbol{\theta} - \mathbf{y}\|_2^2 \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}.\end{aligned}$$

(a) Suppose the error term, $\boldsymbol{\epsilon} = [\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n]$ follows the Laplace distribution, i.e. $\varepsilon_i \overset{i.i.d}{\sim} L(0, b), i = 1, 2, \cdots, n$ and the probability density function is $P(\varepsilon_i) = \frac{1}{2b}e^{-\frac{|\varepsilon_i - 0|}{b}}$ for some $b > 0$. Under the MLE principle, what is the learning problem? Please write out the derivation process. (15 points)

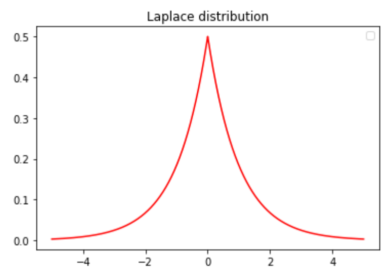

Figure 1: PDF of Laplace distribution

(b) **Huber-smoothing.** $L1$-norm minimization

$$\hat{\boldsymbol{\theta}}_{L1} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \|\mathbf{X}\boldsymbol{\theta} - \mathbf{y}\|_1$$

is one possible solution for robust regression. However, it is nondifferentiable. We utilize smoothing technique for approximately solving the $L1$-norm minimization. Huber function is one possibility. The definition and sketch map are shown as below.

$$h_\mu(z) \begin{cases} |z|, & |z| \geq \mu \\ \frac{z^2}{2\mu} + \frac{\mu}{2}, & |z| \leq \mu \end{cases}$$

Then,

$$H_\mu(\mathbf{Z}) = \sum_{j=1}^{n} h_\mu(z_j).$$

By using Huber smoothing, the approximation of the optimization of $L1-norm$ can be changed to

$$\min_{\boldsymbol{\theta}} \ H_\mu(\mathbf{X}\boldsymbol{\theta} - \mathbf{y}).$$

Let

$$f(\boldsymbol{\theta}) = H_\mu(\mathbf{X}\boldsymbol{\theta} - \mathbf{y}),$$

find the gradient $\nabla f(\boldsymbol{\theta})$.(10 points)

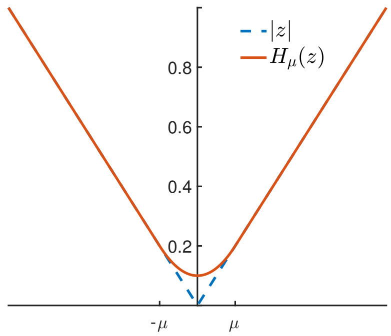

Figure 2: Huber smoothing

(c) Gradient descent for minimizing $f(\boldsymbol{\theta})$. The process of gradient descent algorithm is shown in the following table.

1.	**Input:** observed data $\mathbf{X}$,$\mathbf{y}$ and initialization parameter $\boldsymbol{\theta}_0$ Huber smoothing parameter μ, total iteration number T, learning rate α.
2.	**for** $k = 1, 2, \cdots, T$,**do**
3.	$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \alpha \nabla f(\boldsymbol{\theta}_k)$
4.	**end for**
5.	**return** $\boldsymbol{\theta}_T$

The data set is generated by the linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta}^\star + \boldsymbol{\epsilon}_1 + \boldsymbol{\epsilon}_2,$$

where $\boldsymbol{\epsilon}_1 \in R^n$ follows Gaussian distribution, $\boldsymbol{\epsilon}_2$ are outliers. Given the observed data $(\mathbf{x}, \mathbf{y}) = \{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$ and true value $\boldsymbol{\theta}^\star$,

(1) calculate the estimation $\hat{\boldsymbol{\theta}}_{LS}$ by using linear least squares and compute $\left\|\hat{\boldsymbol{\theta}}_{LS} - \boldsymbol{\theta}^\star\right\|_2$.(5 points)

(2) suppose $n = 1000, d = 50$, use python to implement the gradient descent algorithm to minimize $f(\boldsymbol{\theta})$,the parameters are set as $\mu = 10^{-5}, \alpha = 0.001, T = 1000$, plot the error $\|\boldsymbol{\theta}_k - \boldsymbol{\theta}^\star\|_2$ as a function of iteraction number.You can download the data $\{\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}^\star\}$ from Blackboard.(20 points)