IE30301-Datamining Assignment 1 (70 Points)

Prof. Junghye Lee

March 13, 2021

Exercise 1

What is the goal of learning in data mining? Please explain the following three keywords: **parameter, model, generalization**. The answer does not need to contain the exact terminology, but it should explain the key concepts. [2.5pt]

Exercise 2

Summarize the following concepts in 3 ~ 4 sentences each (write it in your own words). [7.5pt]

1. **Supervised Learning**
2. **Unsupervised Learning**
3. **Regression**
4. **Classification**
5. **Clustering**

Exercise 3

When you derive $SST = SSR + SSE$, you will need to use the Equation (3.1). Please prove that Equation (3.1) holds. [10pt]

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = 0 \tag{3.1}$$

Exercise 4

4.1

Derive the following equations, which are related to simple linear regression. [7pt]

$$\hat{\beta}_0 \sim \mathcal{N}\left(\beta_0, \sigma^2 \frac{\sum x_i^2}{n S_{xx}}\right) \tag{4.1}$$

$$\hat{\beta}_1 \sim \mathcal{N}\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right) \tag{4.2}$$

4.2

Equation (4.1) and (4.2) both follow a normal distribution. However, why do we have to perform a t-test to test the significance of parameters? [3pt]

Exercise 5

In multiple linear regression, you have seen the following Equation (5.1). What does V_{ii} mean? Please describe in detail. [5pt]

$$\hat{\beta}_i \sim \mathcal{N}\left(\beta_i, \sigma^2 V_{ii}\right) \tag{5.1}$$

Exercise 6

Given a set of independent and identically distributed data points $\{x_1, \ldots, x_N\}$, x follows a distribution of $\mathcal{N}(\mu, \sigma^2)$. Derive the **Maximum Likelihood Estimation(MLE)** process that results in parameter estimate in Equation (6.1) and (6.2). Assume that both μ and σ are known. Please show the whole process of your derivation. [20pt]

$$\mu_{ML} = \frac{1}{N} \sum_{t=1}^{N} x_t \tag{6.1}$$

$$\sigma^2_{ML} = \frac{1}{N} \sum_{t=1}^{N} (x_t - \mu_{ML})^2 \tag{6.2}$$

Exercise 7

Given the following data $\mathbf{X} \in \mathbb{R}^{(4\times2)}$ and label $\mathbf{y} \in \mathbb{R}^{(4\times1)}$, solve the below questions. You may use a matrix multiplication calculator. **Explicitly state** all the intermediate results of the calculation you used to find your answer. Otherwise, no points will be given.

$$\mathbf{X} = \begin{bmatrix} 1 & 2 \\ -2 & -2 \\ 0 & 1 \\ 3 & 1 \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} 0 \\ 1 \\ 2 \\ 3 \end{bmatrix}$$

7.1

Fit a multiple linear regression without intercept. Find the parameters of each feature β_1, β_2. Write the answer in **fractional form** (* You do not need to consider the intercept β_0) [5pt]

7.2

Fit a multiple linear regression with intercept. Find the parameters $\beta_0, \beta_1, \beta_2$. Write the answer in **fractional form** (Hint. You may need to manipulate $\mathbf{X}$ to find the intercept) [5pt]

7.3

Compare your results using *Python*. Use the skeleton code given below. Paste your codes and screen capture the results of intercepts and coefficients for both question A and B. [5pt]

```
# Import necessary libraries. Only the following will be needed
import numpy as np
from sklearn import linear_model

X = np.array([[1,2],[-2,-2],[0,1],[3,1]])
y = np.array([0,1,2,3])

```