IE30301-Datamining Assignment 4 (70 Points)

Prof. Junghye Lee

April 29, 2021

Exercise 1

Summarize the following concepts in 3 ~ 4 sentences each. (write it in your own words). If not, there are 2 points deduction per problem. [12 pts, 3 pts for each.]

1. **Linear Discriminant Analysis**
2. **Cross Validation**
3. **AUC for ROC curves**
4. **Decision Tree**

Exercise 2

Consider a multinomial logistic regression model with the dependent variable that has three or more nominal type categories. If we define v_{ij} as the value of category j from the r_i independent trial $\left(\text{instead of the usual binary logistic regression formula } v_{ij} = \begin{cases} 1, & \text{for } y_i = j \\ 0, & \text{for } y_i \neq j \end{cases}\right)$ then the v_{ij} follows a multinomial distribution with probabilities $(P_1, \dots, P_j)$.
Construct the likelihood function for this case. (It is not that complicated. You just need to use the probability mass function of the multinomial distribution) [10 pts]

Exercise 3

Compute the Linear Discriminant projection for the following two-dimensional dataset. [10 pts]

Variable_A	Variable_B	Result
1.84	7.57	1
1.37	9.83	1
2.26	7.82	1
2.18	8.71	1
1.58	4.97	0
1.16	6.31	0
2.27	4.32	0

3.1

Calculate the class statistics: scatter matrices S and mean $\boldsymbol{\mu}$ ($S_1, S_2, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2$) [4 pts]

3.2

Calculate the within- and between-class scatter (S_B, S_W) [3 pts]

3.3

Based on the results 3.1 and 3.2, calculate the optimal $\mathbf{w}^\star$. [3 pts]

Exercise 4

The following tables are confusion matrices of the test dataset from the two methods. (Logistic Regression and Decision Tree). [12 pts]

Table 1: Logistic Regression

Predicted	Disorder	No Disorder
Actual		
Disorder	8	18
No Disorder	45	929

Table 2: Decision Tree

Predicted	Disorder	No Disorder
Actual		
Disorder	12	14
No Disorder	60	914

4.1

Explain how we can interpret the accuracy, sensitivity, and specificity, respectively.
Calculate accuracy rate, sensitivity, and specificity for each method. (Positive class = 'Disorder') [3 pts]

4.2

Compare the accuracy obtained in (4.1) with that of the naïve rule.
(naïve rule: classify all records as belonging to the most prevalent class) [3 pts]

4.3

Which method do you prefer for further implementation in terms of accuracy, sensitivity, and specificity? Explain your reasons. (You should note that the class is imbalanced.) [3 pts]

4.4

If the accuracy rates of those data mining methods were no better than the naïve rule, what would you do to improve accuracy? (Write your own opinion.) [3 pts]

Exercise 5

The following are training samples of 12 objects. Each object is represented as variable X and divided into two classes. (Positive : Class 1, Negative : Class 0) [11 pts]

Object	1	2	3	4	5	6	7	8	9	10	11	12
X	24	30	35	37	42	49	54	56	60	68	72	73
Class	0	0	0	0	1	0	1	1	0	1	1	1

5.1

For the data above, compute sensitivity and specificity according to the change of classification criterion(C) value.

- You should fill in the table below
- Use a classification criterion that if X < C, then classify it as class 0. [6 pts]

Classification criterion	Sensitivity	1-Specificity
X < 24	1	1

5.2

Generate ROC curve based on the computed sensitivity and specificity. And explain how to interpret the ROC curve. [5 pts]
(Use Python or R to plot the ROC curve, but you should provide a screenshot of the code for generating the plot)

Exercise 6

The following data set was collected from the survey, consisting of four attributes: Age, Health Concern, Exercise, Health Status, and one target variable: Health Checkup. [15 pts] **(For only exercise 6, Handwriting is allowed. Illegible handwriting will not be graded)**

Age	Health Concern	Exercise	Health Status	Health Checkup
senior	low	frequent	fair	yes
middle-aged	high	seldom	fair	yes
youth	medium	frequent	excellent	yes
middle-aged	medium	seldom	excellent	yes
youth	high	seldom	excellent	no
youth	medium	seldom	fair	no
middle-aged	low	frequent	excellent	yes
middle-aged	high	frequent	fair	yes
senior	medium	seldom	excellent	no
youth	high	seldom	fair	no
senior	low	frequent	excellent	no
senior	medium	seldom	fair	yes
youth	low	frequent	fair	yes
senior	medium	frequent	fair	yes

6.1

Compute the Gain Ratio for each attribute. Which variable will be a splitting criterion at the root node in terms of Gain Ratio? (Take the multi-split approach for splitting and the binary logarithm (i.e., base 2) for calculating the Gain Ratio. Write down the calculation process) [10 pts]

6.2

What is the classification error right after splitting the root node according to the result of 5.1? [5 pts]