IE30301-Datamining Assignment 5 (70 Points)

Prof. Junghye Lee

May, 2022

Exercise 1

Determine the True/False (T/F) of the following statements and describe the reason(s). No point will be given without appropriate reason. You must also give your reasoning if you think the statement is true. [total 20 pts, 2 pts per each.]

1. For Multiple Linear Regression, We can construct T-test for testing overall model.
2. The parameters of logistic regression obtained by gradient descent algorithm and Newton-Raphson algorithm are same.
3. Among Linear Regression, PCA, Logistic Regression, LDA, Decision Tree, and KNN, only decision tree is non-parametric model.
4. The objective of Principal Component Analysis (PCA) is to derive uncorrelated features having the variance of the original data as maximum.
5. Linear Discriminant Analysis (LDA) will fail when the discriminatory information is not in the variance but rather in the mean of data.
6. In classification task, if there are imbalance of classes, we use accuracy rather than AUC of ROC Curve for evaluation method.
7. To construct the contingency table for the result (output) of the soft classifier, a certain threshold needs to be predefined.
8. In Bagging trees, individual weak learners are dependent of each other.
9. K-Nearest Neighborhoods algorithm does more computation on classify time rather than train time.
10. For Support Vector Machine (SVM), deleting the support vectors will change the position of the hyperplane.

Exercise 2

Describe the "**main difference**" between the following two concepts of each question "**in one sentence.**" (note : you should answer only with the content discussed in class. Otherwise no points even though the answer is correct. You will also get no points when you answer differences that are not main.) [total 21 pts, 3 pts per each.]

1. **Predictive Data-Mining Task vs. Descriptive Data-Mining Task**
2. **Principal Component Analysis vs. Linear Discriminant Analysis**
3. **Over-Fitting vs. Under-Fitting**
4. **Support Vector(s) vs. Non-Support Vector(s)**
5. **Hard Margin vs. Soft Margin**
6. **Euclidean Distance vs. Pearson Linear Correlation**
7. **Single Linkage vs. Complete Linkage**

Exercise 3

Suppose we have a kernel function for arbitrary vectors $\mathbf{x}, \mathbf{z} \in \mathbb{R}^2$, which is defined as $K(\mathbf{x}, \mathbf{z}) = (\mathbf{x}^T\mathbf{z})^2$. [14 pts]

3.1

We know that the given kernel function can be decomposed into the inner product of some transformation function $\phi(\cdot)$; $K(\mathbf{x}, \mathbf{z}) = \phi(\mathbf{x})^T\phi(\mathbf{z})$. For $\mathbf{x} = (x_1, x_2), \mathbf{z} = (z_1, z_2)$, find the transformation function $\phi(\cdot)$ [4 pts]

3.2

For given vectors $\mathbf{x}_1 = (1, 2), \mathbf{x}_2 = (4, 3), \mathbf{x}_3 = (2, 0)$, Find the transformation of each vector using the transformation function you derived in **3.1** [3 pts]

3.3

Construct a kernel matrix K using $\mathbf{x}_1 = (1, 2), \mathbf{x}_2 = (4, 3), \mathbf{x}_3 = (2, 0)$. [4 pts]

3.4

Check if the kernel matrix you get in **3.3** is positive semi-definite or not. [3 pts] (To check positive semi-definiteness : you should verify it based on the fact that all eigenvalues of the given matrix are non-negative. https://www.symbolab.com/solver/matrix-eigenvalues-calculator/eigenvectors)

Exercise 4

Suppose you are the teaching assistant(TA) of the Data Mining Course. The final exam is coming soon, and the professor asked you to make one question for the Final exam. As the TA, you wanted to check if the students have a solid understanding of the topic "SVM." Make a question that checks the concepts and theory you have learned on "SVM" with the below considerations. Provide the solution to your question as well. [15 pts]

1. The total score of the final exam is 100 points. The portion of this question you are making is 10 points.
2. The question should not be so easy nor so hard. You want the answer rate to be around 4 ~ 60%
3. Students who have studied the lecture note should be able to answer this question without much difficulty

To evaluate this question, all 5 TAs will grade this question together, and the median high score will be given as your score to this question. The grading will be based on

1. Completeness of the question. Is the question understandable and doable?
2. Does the question well satisfy the above considerations?
3. Does it well tests the understanding of the topic?