Data Mining Assignment 1

Nischal Regmi
Everest Engineering College

2021

Part 1: Basics

1. How would you define data mining?
2. Describe in brief how
   - Presence of noise could lead to overfitting
   - Lack of representative sample could lead to overfitting

   Note: Refer to the textbook by Pang-Ning-Tan et al.
3. Your are given twenty numbers

   ```
   9 16 25  8 11 11 15 48 64 7 4 47 17 37 83 21 25 1 88  4
   ```

   Do you think the distribution is uniform? Left-skewed? Right-skewed?
4. Give at least two examples of situations where data imbalance would be a problem. Assure that the examples you give are not in my lecture slides.
5. What is statistical significance? What statistical parameters can be used to assess statistical significance. Explain with an example.
6. What is a classification problem? Give the formal definition. How does the classification problem differs with the regression problem?
7. Give at least one examples for the following variants of real-world classification problems.
   (a) Classification problem where all the predictor/explanatory variables are numeric
   (b) Classification problem where all the predictor/explanatory variables are categorical
   (c) Classification problem where some of the variables are numeric and some are categorical.

   In each of your examples, you should describe the class labels and the variables in detail.
8. Define the indicators of classifier accuracy.
9. Suppose that given a person's name, x, a classifier predicts the gender as following:

$$\hat{f}(x) = \begin{cases} \text{female} & \text{if the last letter of } x \text{ is a vowel} \\ \text{male} & \text{if the last letter of } x \text{ is a consonant} \end{cases}$$

   Calculate the different accuracy parameters for the classifier in context of the below data.

x	Ram	Suman	Shiva	Geeta	Ranjan	Gunjan
Gender	Male	Female	Male	Female	Male	Male

10. Why the Hunt's algorithm for building decision tree is not applicable to realistic datasets? Explain.

11. Define the purity measures viz. entropy, Gini and classification error. What will be the values of these measures after applying the classifier in problem 8 to the data in the same problem?

12. Consider for a binary class problem, a node corresponds to the following data

A	B	Class Label
T	F	+
T	T	+
T	T	+
T	F	-
T	T	+
F	F	-
F	F	-
F	F	-
T	T	-
T	F	-

    (a) Calculate the information gain when the node is splitted using attribute A? Calculate the same using attribute B? Which attribute will the decision tree induction algorithm chose for splitting the node?

    (b) Which attribute the algorithm will chose if we use the gain in Gini instead of information gain?

13. What is overfitting? Describe how cross-validation can be used to find the appropriate values of model parameters and avoid overfitting.

14. Formally define the regression problem. Differentiate between

    (a) Linear and non-linear regression

    (b) Parametric and non-parametric model

15. You are given the annual values of Gross Domestic Product for Nepal.

x (Year)	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021
y GDP PC	791	794	809	828	882	880	1028	1162	1186	1139	1208

    (a) Fit the linear regression model $y = ax + b$ and calculate the total least square error.

    (b) Fit the non-linear regression model $y = ae^{bx}$ and calculate the total least square error.

    (c) Which model is better?

16. The linear Support Vector Regression model estimates a function of the form

$$y = \mathbf{w}^T\mathbf{x} + b$$

    which is exactly same as the linear regression model. Then what is the difference between the difference between the basic linear model, that is usually estimated using least squares method, and the SVR model? (Hint: Refer to the definition of OLS estimation for linear regression, and the convex optimization problem for the SVR)

17. Illustrate Cover's theorem with any other example that is not given in my lecture slides.

18. How does Mercer's theorem allows to convert linear SVR model to non-linear SVR. Explain referring to the final solution to the non-linear SVR model (equation 11 in my slides for chapter4, regression)

Note: The internal exam will also ask question related to the algorithms discussed in the lectures. Students are informed to understand the major idea behind the algorithms.