Data Mining - EX8

Deadline: Friday, Dey 7 , 1403 - December 28, 2024

Question 1: K-Nearest Neighbors Classification

The table below provides a training data set containing six observations, three predictors, and one qualitative response variable.

Obs.	X	X	X	Y
1	0	3	0	Red
2	2	0	0	Red
3	0	1	3	Red
4	0	1	2	Green
5	-1	0	1	Green
6	1	1	1	Red

Suppose we wish to use this data set to make a prediction for Y when X = X = X = 0 using K-nearest neighbors.

(a) Compute the Euclidean distance between each observation and the test point, X = X = X = 0.

(b) What is our prediction with K = 1? Why?

(c) What is our prediction with K = 3? Why?

(d) If the Bayes decision boundary in this problem is highly non-linear, then would we expect the *best* value for K to be large or small? Why?

Question 2: Decision Trees (CART and C4.5)

In this exercise, you will use two well-known decision tree algorithms (CART and C4.5) to analyze the given data and evaluate the results. First, you will fit a regression tree, and then you will convert the target variable into a categorical variable to compare the two decision tree algorithms.

The table below contains 14 rows of data with 4 columns. The first column represents the target variable (numerical), while the other three columns are predictor variables (with categorical values only).

Part 1: Regression Tree

1. **Fit a regression tree** using the data in the table above.

Predictor 1 (Categorical)	Predictor 2 (Categorical)	Target Variable (Numerical)
A	X	30
B	Y	50
A	X	40
A	Z	60
B	Y	55
A	Y	45
B	Z	70
A	X	50
B	Y	60
A	Z	65
B	Y	80
A	X	35
B	Z	50
A	Y	40

2. Analyze the regression tree results. Draw the tree diagram and explain the results.

Part 2: Convert Target Variable to Categorical

1. **Convert the target variable to a categorical variable**:
   - if the target variable is greater than 51, classify it as High; otherwise, classify it as Low.

Part 3: Decision Tree Algorithms

1. **Apply the C4.5 algorithm**: After converting the target variable into a categorical variable, apply the C4.5 algorithm to build the decision tree.
   - The splitting criterion in C4.5 is based on reducing entropy (Information Gain).
2. **Apply the CART algorithm**: After converting the target variable into a categorical variable, apply the CART algorithm to build the decision tree.
   - The splitting criterion in CART is based on reducing Gini impurity.

Part 4: Model Evaluation

1. **Calculate the Missclassification Error**:
   - For both C4.5 and CART, calculate the missclassification error.

- The formula for Missclassification Error is:

$$\text{Missclassification Error} = \frac{\text{Number of Incorrect Predictions}}{\text{Total Number of Samples}}$$

Questions:

1. What differences did you observe between the decision trees generated by the C4.5 and CART algorithms?
2. What were the Missclassification Error results for each algorithm? Which algorithm had the better accuracy?
3. How can decision tree algorithms be optimized for different datasets? (Bonus)

Question 3: KNN And Tree Model Implementation on Boston Dataset

In this problem, we aim to implement the KNN (K-Nearest Neighbors) model on the famous Boston dataset. The Boston dataset contains various information about house prices in the city of Boston and can be used to predict house prices.

Let's first see how we can load the Boston dataset in **R** and **Python**.

Loading the Boston Dataset in R:

To load the Boston dataset in R, you need to install and load the `MASS` package first:

```
install.packages("MASS")
library(MASS)
data(Boston)
```

Loading the Boston Dataset in Python:

In Python, we use the `sklearn.datasets` library to load the Boston dataset:

```
from sklearn.datasets import load_boston
import pandas as pd

boston = load_boston()
df = pd.DataFrame(boston.data, columns=boston.feature_names)
df['target'] = boston.target
```

KNN Implementation

1. **Select a Variable from the Dataset:**

   Choose the "CRIM" variable (crime rate in each area) and convert it to categorical.

   - If "CRIM" is less than 3.5, classify it as "Low Crime".
   - If "CRIM" is greater than 3.5, classify it as "High Crime".

2. **Split the Data into Training and Testing Sets:**

   After selecting and categorizing the variable, split the data into two parts: training and testing datasets.

3. **Normalize the Training Data:**

   Normalize the training data to standardize the feature scales.

4. **Train the KNN Model:**

   Now that the data is prepared, train the KNN model and make predictions using the test data. Experiment with different values of **k** and discuss the results.

5. **Confusion Matrix Calculation:**

   Finally, calculate the confusion matrix to evaluate the accuracy of the model.

6. **Explain Your Results:**

   Discuss the results and explain the performance of the model.

7. This time, the **Decision tree** is fitted to the data.

8. **Confusion Matrix Calculation:**

   Finally, calculate the confusion matrix to evaluate the accuracy of the model.

9. **Explain Your Results:**
   Discuss the results and explain the performance of the model.