Homework 6

IE 7275 Data Mining in Engineering

Before you start: Read textbook Chapters Decision Trees, Logistic Regression and Neural Networks.

Submission Requirement: You should submit two answer sheets for this homework. One for non-coding problems 1 and 3 and the other for coding problems 2 and 4. Please type your steps and answers for the non-coding problems. Hand-written solutions will not be accepted.

Problem 1

We plan to build a decision tree using 7 records in the file Problem1.xlsx. The task in this problem is to find the first split using both Gini index and entropy as the impurity measure. Calculate the purity improvement after the first split respectively.

Problem 2

Answer the following short answer questions and back up your answer with explanations and/or examples.

TODO 1

- What type of input and response variables can a decision tree model handle?
- What kind of dataset is ideal for applying the decision tree model?
- Discuss the classification tree and regression tree separately if necessary.

TODO 2

- What are the pros and cons of a decision tree model compared to other models we learned in class?

TODO 3

- Between the Naive Bayes classifier and the classification tree, which one is more prone to overfitting the training data?

Problem 3 - 6

Please refer to Google Colab file Homework 6 - Coding Problems.