Imperial College London
MSc Business Analytics

Assignment 3

Machine Learning
January 24, 2025

1 Individual Assignment

Instructions: *This exercise should be done "by hand", that is, not using Python. All necessary calculations should be included in the submission, as well as brief explanations of what you do.*

The training data set in Table 1 on page 3 provides a summary of congestion experienced along Cromwell road across different days, times and weather conditions.

1. Create a full classification tree that estimates the traffic based on the day of the week, time of the day and the weather condition. Use the entropy as a purity measure to guide your splitting choices.
2. Construct a confusion matrix for the performance of your tree on the training data. What is the misclassification rate, and what is the sensitivity and specificity (assuming that 'yes' is the 'important' class)?
3. Construct the confusion matrix that results if your algorithm is applied to the test set in Table 2 on page 4.

2 Group Assignment:

Instructions: *This exercise should be done using Python. The source codes as well as all relevant outputs should be included in the submission, as well as brief explanations of the code.*

In this exercise, we try to predict defaults in student loan applications. To this end:

1. Load the data set loandata.csv into Python.
2. The data set contains some categorical predictors. Sklearn, which you should use for this exercise, can only handle numerical predictors. Translate the categorical predictors into numerical predictors. (You may want to look into the pandas function **get_dummies**.)
3. Shuffle the data set and split it into 60% training data, 30% validation data and 10% test data.
4. Calculate the accuracy of the naive benchmark (majority predictor) on the validation set.
5. Train a decision tree and calculate the accuracy of this tree on the training and validation set. Choose an appropriate maximum depth and justify your choice. (Look at the **max_depth** parameter). All other settings should be kept at default values. What do you think of this classifier?

6. Try a random forest algorithm instead. Use different number of estimators and plot the accuracy (on training and validation data) as a function of the number of estimators. (Look at the **n_estimators** parameter).

7. Also plot the training time (vs the number of estimators) for the random forest models in the previous step. What appropriate number of estimators would you choose? Why?

8. Compare and explain the performance, interpretability, training time and generalisability of your decision tree in part 5 to your chosen random forest estimator in part 7.

Table 1: Training set for individual assignment

day	weather	time	congestion
weekday	rainy	8am	yes
weekend	sunny	8am	yes
weekday	sunny	8am	no
weekday	sunny	1pm	no
weekday	rainy	1pm	yes
weekend	sunny	8am	yes
weekend	rainy	8am	no
weekend	sunny	1pm	no
weekday	sunny	1pm	no
weekday	sunny	8am	no
weekday	rainy	1pm	no
weekend	rainy	8am	yes
weekday	rainy	1pm	yes
weekend	sunny	8am	yes
weekday	sunny	1pm	yes
weekend	sunny	1pm	yes
weekday	rainy	8am	yes
weekday	sunny	1pm	no
weekday	sunny	8am	no
weekday	sunny	1pm	no
weekday	rainy	8am	no
weekend	rainy	8am	no
weekend	sunny	1pm	yes
weekday	rainy	1pm	yes
weekend	rainy	1pm	yes

Table 2: Test set for individual assignment

day	weather	time	traffic
weekday	sunny	1pm	yes
weekday	sunny	8am	no
weekend	sunny	1pm	yes
weekend	sunny	8am	no
weekend	rainy	1pm	yes
weekday	rainy	8am	no
weekday	sunny	8am	no
weekday	rainy	1pm	yes
weekday	sunny	1pm	no
weekday	sunny	1pm	yes
weekend	rainy	1pm	yes
weekday	rainy	8am	no
weekday	sunny	1pm	no
weekend	rainy	1pm	yes
weekend	sunny	1pm	yes