Statistical Methods for Data Science

DATA7202

Semester 1, 2021

Lab 3

Objectives

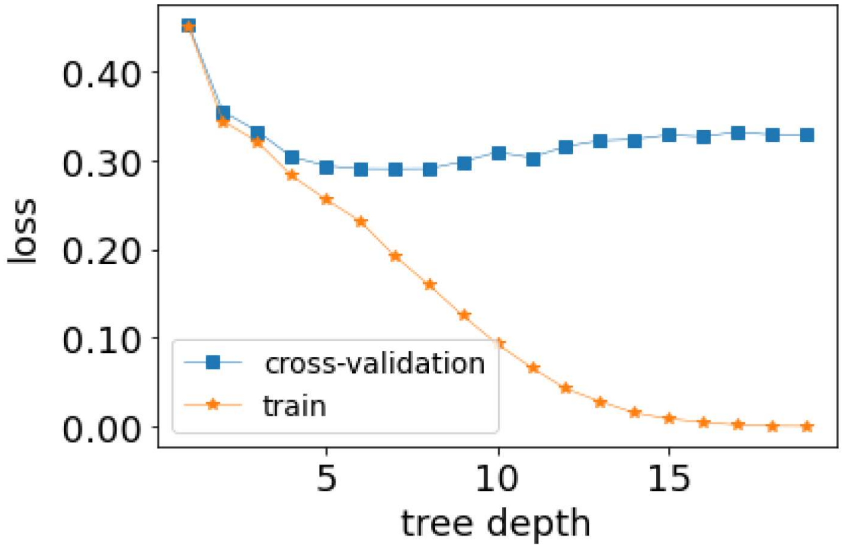
On completion of this laboratory session you should be able to understand and implement decision trees.

1. Using the program from the Lecture, write a basic implementation of a decision tree for a binary classification problem. Implement the misclassification, the Gini index, and the entropy score criterion. Compare the Gini index and the entropy with the misclassification score. What metric is more sensitive to node's impurity?

2. Consider the following data generation process.

```
X, y =  make_blobs(n_samples=5000, n_features=10, centers=3,
                                random_state=10, cluster_std=10)
```

We are going to find the best decision tree depth using cross-validation procedure. Write a code to reproduce the following Figure.



3. Explain why bagging decision trees is a special case of random forest.

4. Consider the `mnist` dataset.

   (a) Plot several images from the dataset.

   (b) Split the dataset to train and test sets (75% train and 25% test).

   (c) Fit logistic regression model and evaluate the miss-classification rate.

   (d) Fit a random forest classifier, evaluate the miss-classification rate, and compare to the results obtained in (c).