IST707 Applied Machine Learning

HW5: Use Decision Tree to Solve a Mystery in History

In this homework assignment, you are going to use the decision tree algorithm to solve the disputed essay problem. Last week you used clustering techniques to tackle this problem.

Organize your report using the following template:

Section 1: Data preparation

You will need to separate the original data set to training and testing data for classification experiments. Describe what examples in your training and what in your test data.

Section 2: Build and tune decision tree models

First build a DT model using default setting, and then tune the parameters to see if better model can be generated. Compare these models using appropriate evaluation measures. Describe and compare the patterns learned in these models.

Section 3: Prediction

After building the classification model, apply it to the disputed papers to find out the authorship. Does the DT model reach the same conclusion as the clustering algorithms did?

If you are interested in such kind of applications, here are several relevant scientific papers:

1. Y. Yang et al., Estimating the deep replicability of scientific findings using human and artificial intelligence. Proc. Natl. Acad. Sci. U.S.A 117 (20) 10762-10768 (2020).
2. A. Dreber et al., Using prediction markets to estimate the reproducibility of scientific research. Proc. Natl. Acad. Sci. U.S.A 112, 15343–15347 (2015).
3. E. Forsell et al., Predicting replication outcomes in the Many Labs 2 study. J. Econ. Psychol. 75, 102117 (2019).
4. A. Altmejd et al., Predicting the replicability of social science lab experiments. PloS one 14, e0225826 (2019).