Homework IV

STA 629, Fall 2024

Due: November 15, Friday

1. (80 pts) For this problem, start by obtaining the authorship dataset from the assignment webpage on Canvas. Within the authorship data, the first 69 columns represent the frequency of the 69 most frequently used words in an article, while the final column corresponds to the article's authorship. Implement the classification task that involves forecasting the authorship of an article based on the frequencies of these common words, using the provided training and testing splits of the authorship data. Compare and contrast the following methods for predicting authorship:

   (a) Classification Trees. (Which error measure did you use? Why?)

   (b) Bagging.

   (c) Boosting. (Which boosting method did you use? Why?)

   (d) Random Forests. (Which parameter settings did you use? Why?)

   Reflect upon your results. Which method yields the best error rate? Which method yields the most interpretable results? Which words are most important for authorship attribution?

2. (20 pts) Textbook Problem, ESL, Exercise 10.1.

- Bonus question (20 pts): Textbook Problem, ESL, Exercise 10.2.