Homework IV

Statistics 640 / 444 Fall 2015

Assigned: November 3

Due: November 24

Stat 640 & Stat 444:
For your homework, download the authorship data set from the course webpage.

1. Visualization & Clustering.
   For this problem use the combined training AND testing data from the course website. Your goal is to use the word counts to cluster the data (without the author labels) and see if your groups coincide with the true author attribution.
   (a) Visualize the data. (You may choose to use one or more methods to visually summarize the data and use for exploratory analysis).
   (b) Compare are contrast the following clustering methods:
       i. K-means.
       ii. Hierarchical Clustering (Try at least 4 linkages and at least 3 distances. Which ones did you choose? Why?)
       iii. Biclustering. (You may try your choice(s) of biclustering method(s). Hint: Try the NMF and the cluster heatmap.)

   Reflect upon your results. Which is the best method to visualize the data? Which distance metric is best? Which clustering method yielded groups that closely coincide with true authorship? Why? Are there any words that are more important for clustering? Which ones?

   For this problem use 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?