IS 733 Homework 4, Due 4/20/2021

Points

1. (a) Briefly describe the *boosting* algorithm. State why it may improve classification accuracy. (10)
   (b) What is the *bias-variance trade-off* for machine learning methods? Explain. (5)
   (c) Briefly describe the *bagging* procedure. Discuss why it may improve the accuracy of *decision tree* classifiers, in terms of the bias-variance trade-off. (10)

   For questions 2 and 3 of this homework, we will use the Old Faithful Geyser dataset, which you can download at `http://www.stat.cmu.edu/~larry/all-of-statistics/=data/faithful.dat` . This dataset describes the properties of eruptions of the Old Faithful geyser, located in Yellowstone National Park, Wyoming, USA. There are two numeric attributes per instance: the length of time of the eruption, in minutes, and the waiting time until the next eruption, also in minutes. The geyser was named "Old Faithful" because its eruption patterns are very reliable. See `https://en.wikipedia.org/wiki/Old_Faithful` for more information, if you are interested.

2. (a) Using any software tool or programming language of your choice, create and print out a scatter plot of this dataset, eruption time versus waiting time. Note that for many tools, before the data can be loaded you will need to make a copy of the file and delete the header information. You will need to ignore the first column, which contains ID numbers for each instance. (10)
   (b) How many clusters do you see based on your scatter plot? For the purposes of this question, a cluster is a "blob" of many data points that are close together, with regions of fewer data points between it and other "blobs"/clusters. (5)
   (c) Describe the steps of a hierarchical clustering algorithm. Based on your scatter plot, would this method be appropriate for this dataset? (10)

3. Implement the k-means algorithm in the programming language of your choice, and use it to perform clustering on the Old Faithful dataset. Use the number of clusters that you identified in Question 2b.

   I recommend using a high-level data-friendly programming language such as matlab, R, or python. Be sure to ignore the first column, which contains instance ID numbers. Report the following items:

   - Your source code for the k-means algorithm. You do not need to report code for loading the data, or for drawing a scatter plot. **You need to implement the algorithm from scratch.**
   - A scatter plot of your final clustering, with the data points in each cluster color-coded, or plotted with different symbols. Include the cluster centers in your plot.
   - A plot of the k-means objective function versus iterations of the algorithm. Recall that the objective function is

     $$E = \sum_{i=1}^{k} \sum_{p \in C_i} \|p - c_i\|^2 \ ,$$

     where k is the number of clusters, C_i is the set of instances assigned to the ith cluster, and c_i is the cluster center for the ith cluster. Note that the objective function should always decrease. If this is not the case, look for a bug in your code.
   - Did the method manage to find the clusters that you identified in Question 2b? If not, did it help to run the method again with another random initialization?

   (50)

Total: 100