ECE 532, Spring 2019
Homework #7
Due: Thurs, May 9 @ 2:30 pm

Problem 1. The file `faces.csv` contains 33 images of the same person from different angles (each image has dimension 112×92). You can visualize the i^{th} image using the commands `colormap(gray); imagesc(reshape(X(i,:),112,92))`, as in previous homework exercises.

(a) Use MDS with squared euclidean distances to embed the 33 images into 2D space. Plot the result, labeling each data point by its row index. Also perform PCA with $q = 2$ and check that the results are the same (up to rescaling).

(b) Now perform the Laplacian eigenmaps method on the 5-nearest neighbor graph (where two points are connected as long as *either* point is among the 5 nearest neighbors of the other). Use $q = 2$ and provide a scatter plot of the result, again labeling each data point by its row index.

(c) Now perform the Isomap method on the same 5-nearest neighbor graph from (b). (You can use the `distances` command in Matlab to compare shortest path distances.) Use $q = 2$ and provide a scatter plot of the result, again labeling each data point by its row index.

(d) Visualize all 33 face images using the `imagesc` command. Examine your results for (a), (b), and (c). How might you label the axes in your plot according to some properties of the images? Which visualization (MDS, Laplacian eigenmaps, or Isomap) do you think works best?

Problem 2. We return to the breast cancer data set `bc_wisc.csv` studied in HW 3 and HW 5. First separate the data matrix into the first 400 rows (training data) and last 160 rows (test data). Recall that the first column lists the ID number of the individual, the second column records benign (0) vs. malignant (1), and the remaining 30 columns list measurements related to the tumor. (For more information on the dataset, see the description here: `https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)`.)

(a) Use the Matlab command `fitctree` (with default settings) to train a decision tree on the training data. Report the fraction of test points that are classified correctly using the decision tree (the command `predict` might be helpful).

(b) Visualize the trained decision tree using the Matlab command `view`. Suppose we consider a feature to be "important" if it appears in one of the splits in the top three layers of the decision tree. Which features are considered to be important, according to the trained tree? (You can just list the indices of the features—these should be contained in the first 7 components of the `CutPredictor` array in the output.)

(c) Now use bagging: Create an ensemble of 100 decision trees, where each tree is trained on a bootstrapped sample of 300 data points (randomly sampled, with replacement, from the training data set). The prediction for a test point is computed using a majority vote among the 100 trained trees. Report the fraction of test points that are classified correctly using the ensemble, and compare it to (a).

(d) Use the same criterion as in (b) to determine which features are important for each of the 100 trees in the ensemble. What are the 5 most important features ovreall, ranked according to the number of times they appear as important features among the 100 trees in the ensemble?