**ECE 8527: Introduction to
Machine Learning and Pattern Recognition**

# HW No. 09: Bootstrapping, Bagging and Combining CLasssifiers

Using the data set no. 10, create two types of partitions:

1. Conventional: Use the entire training set;
2. Bagging: create 10 randomly selected subsets that are 75% of the size of the training set (put an equal number of tokens from each class in each subset);
3. Cross-validation: split the training set into 10 subsets that are mutually exclusive.

Use your class-dependent PCA system as the representative system that we will test on these datasets.

Next, construct these systems:

1. Conventional: Your baseline PCA system (previously reported) trained on (1) (system A).
2. Conventional with bootstrapping: train a system on 9 of the subsets from (3) and evaluate on the held-out set. Do this for all permutations of the 10 sets (e.g., rotate the held-out set so that each subset serves once as the held-out set. Choose the best performing system (label it system B.1), and also the worst performing system (label it system B.2).
3. Conventional with cross-validation: Train your PCA system on the 9 subsets of (2) and evaluate on the remaining subset. Rotate through all the subsets. Compute the error rates as the average across these 10 partitions (system C). Compare this to the eval error rate for (a).
4. Bootstrap: Compute an averaged transformation matrix for each class by averaging the transformation matrices that you computed on each of the 10 partitions (system D). Report the error rate (not the averaged transformation matrix).
5. Bagging: Using (2), train 10 systems (one on each subset), and evaluate using a majority vote (system E).
6. Boosting: Implement the boosting approach discussed in lecture 20 (system F).

Present the data in a single table and comment on your findings. Include the error rate on the training set (estimated in an algorithm-dependent manner), the dev set and the eval set (estimated using systems A-F).

Finally, using the SKLearn tools to generate an ROC curve, generate and plot an ROC curve comparing systems A, E and F.