CSE 40647/60647 Data Science (Spring 2018)
Lecture 10: Classification: Ensembles

Goals:

- Describe ensemble methods
    - Bagging: Random Forest (Bagged Decision Trees)
    - Boosting: AdaBoost (Adaptive Boosting)

Bagging:

- Given a data set D of *d* instances, a classifier model M_i is learned for a training set D_i of *d* instances that is *sampled with replacement* from D (i = 1…k)
- As a result of the *sampling-with-replacement* procedure, each classifier is trained on approximately *63.2%* of the training examples

Boosting:

- *Weights* are assigned to each training instance
- A series of *k* classifiers is *iteratively* learned
- After a classifier M_i is learned, the weights are updated to allow the subsequent classifier, M_{i+1}, to pay more attention to the training instances that were *misclassified* by M_i
- The final M* *combines the votes* of each individual classifier, where the *weight* of each classifier's vote is a function of its *accuracy* on classifying training instances

AdaBoost:

- Given a set of *d* class-labeled instances, $(\mathbf{X_1}, y_1), \ldots, (\mathbf{X_d}, y_d)$
- Initially, all the *weights* of instances are set the same (1/*d*)
- Generate *k* classifiers in *k* rounds. At round *i*,
    - Instances from D are *sampled with replacement* to form a training set D_i of the same size
    - Each instance's chance of being selected is based on its *weight*
    - A classification model M_i is derived from D_i
    - Its *error rate* is calculated *using D_i as a "test set"*
    - If an instance is misclassified, its *weight* is increased, otherwise it is decreased