Random Forest

- Random forest(RF) is a classifier

- RF is built on a forest of decision trees

- Classify a new data object x : every decision tree assigns a label l. The final over-all class label for x is obtained by majority voting, same as in KNN.
  - Why majority voting is good?

- Random forest typically performs very well, similar to SVM, logistic regression.

- RF is an example of ensemble learning (bagging)

Random Forest

- Building (training) the RF:
  - Repeat for each decision tree
  - Randomly split training data into (X_train, X_test)
  - From this X_train, build a decision tree
    - But in spliting a node, we only consider a limited number of features (instead of all features in standard decision tree construction)
    - This limited number of features are randomly chosen at each node.
    - The number of this feature set is input parameter. This number could be all features
  - For each decision tree, because the training data is different (and also the feature set could differ), the constructed decision tree is different
  - X_test is used to compute the classification error of this decision tree. This is out-of-bag error (oobERROR). No cross-validation is necessary.