Homework Assignment 2

To learn the programming and analytics skills, master students are required to complete Option 1: hands-on exercises. PhD students have two options: (1) hands-on exercises and (2) paper reading.

Option 1: Hands-on exercises

1. Use the same data matrix in the homework assignment 1
2. Run kNN and Decision Tree with 5-folder cross validation. (label each new article with the category). You can find the introduction of article categories from http://qwone.com/~jason/20Newsgroups/. If you are a python user, you can find:

- Decision tree classifier: http://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html
- KNN classifier: http://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html

To perform model training and model testing, write your own training and testing codes, for example:
training.py:
%read training set (a CSV file)
%train a decision tree with training set
%save the trained decision tree model to a txt file

testing.py:
%load the trained decision tree model from the txt file
%read testing set (a CSV file)
%compute the accuracy
%save the results to a txt file

With your codes, you can train the classifier (decisontree/knn) with training set and test the learned model (decisontree/knn) on test set. I recommend you to implement a simple five fold cross validation by yourself. In this way, you control all your codes. For instance, we can create a main.py:
% randomly split the data matrix into five exclusive partitions (namely part1, part2, part3, part4, par5) in terms of rows. Don't split data matrix in terms of features.
%use part1 as testing set, part2+3+4+5 as training set, feed the training set to training.py and and get a trained decision tree model, thend feed the testing set to testing.py and get the result
%similarly, you can apply the above step to part2, or part3, or part4, or part5 for testing
%aggregate all the five results into an average performance measurement

3. Compare the overall accuracies and f-measures of kNN and decision tress before feature selection and after feature selection. Can feature selection help improve the accuracies of classifiers and why?

4. Evaluate how K impacts the overall accuracy and f-measure of kNN on the dataset. Use histogram plots to visualize the results and identify the best K.
5. Compare the overall accuracies and f-measures of kNN with the best K and decision trees using histograms. Which classifier is better and why?
6. Compare the accuracies of each article category of kNN with the best K and decision trees using histograms. For a particular article category, which classifier is better?
7. In general, which classifier is better?

Deadline: please submit your results by 02/21. You may compress your data matrix, codes, and report into a zip file to submit.

Alternative option for PhD students: Paper reading

Select a classification paper that has more than 200 citations, read, and present in the class on 02/21. You are encouraged to include motivation, problem formulation, methodology overview, technical details, and experiment interpretations in your presentation. The presentation should take no more than 10 minutes.