Data Mining Homework

1. Goal: observe the performance of classification on difficult vs easy datasets

* Newsgroup data has 6 main classes with multiple subclasses for each class. The main classes include Computer, recreation, science, politics, miscellaneous and other.

1. Perform Naiive bayes classification on an easy data set that includes the classes (recreation, computers)
2. Perform Naiive Bayes classification on a difficult dataset that includes the classes rec.motorcycles and rec.autos.
3. Repeat a and b using the decision tree classifier
4. Discuss the results

* Repeat the classification of the difficult data set above using:

1. Bagging
2. AdaBoost
3. RandomForest
4. Compare performance of the ensemble classifiers.
5. Experiment with (10,20,30)-fold cross validation and discuss whether increasing the folds affects the stability of the ensembles performance.

1. Multi-label classification:

* Form a sample of the following classes: comp.graphics, rec.autos, talk.politics.gun, soc.religion.christian. Include at least 200 documents from each subclass.
* Perform one-vs-all classification using Naïve Bayes, decision trees and support vector machine SVC classifiers.
* Repeat classification using all-vs-all.
* Discuss the results for both approaches and rank classifiers based on performance

1. Assuming a decision stump classifier with the AdaBoost ensemble classifier. How does the classifier process the weights of the data points to focus on misclassified data points?