Weka Exercises

KNN classifier is implemented with the name IBk
The tree classifier to use is the J48
The SVM are implemented with the name SMO (package functions)
iris.arff, vote.arff, diabetes.arff and glass.arff, you can find these datasets on internet just typing the name of the dataset with the suffix on google.

In Preprocess:
a) load a dataset (iris.arff) and look at it
b) use the Data Set Editor
c) apply a filter (to remove attributes and instances).

Load a dataset (iris.arff) and classify it with the J48 decision tree learner (test on training set):
a) examine the tree in the Classifier output panel
b) visualize the tree (by right-clicking the entry in the result list) c) interpret classification accuracy and confusion matrix

Experiment with the IBk classifier for nearest neighbour learning:
a) load glass.arff data; list attribute names and identify the class attribute
b) classify using IBk, testing with cross-validation
c) repeat using 10 and then 20 nearest neighbours
d) interpret the results and draw conclusions about IBk.

Experiment with the IBk classifier for nearest neighbour Learning :
a) load diabètes.arff data ; list attribute names and identify the class attribute
b) classify using IBk (3NN), testing with Hold-out (Training 70% - Test 30%)
c) classify using IBk (3NN), testing with 10 fold cross-validation
d) Note the différence in classification between hold-out and cross validation

Experiment with the SMO classifier:
a) load iris.arff data ; list attribute names and identify the class attribute
b) classify using SMO using 10 fold cross validation
c) note the results obtained in term of accuracy
d) discretize the iris dataset and re-apply the poin b) and c)
e) note the différence in classification between the non discretize and the discretize version

Investigate linear and non-linear support vector machines:
a) apply SMO to iris.arff dataset, again evaluating on the training set
b) change the “exponent” option of the kernel “PolyKernel” from 1 to 2 and repeat
c) try to explain the differences in the test results

Apply feature selection using CfsSubsetEval:
a) load the vote.arff dataset and apply J48, IBk and NB, evaluating with cross-validation
b) select attributes using CfsSubsetEval and GreedyStepwise search
c) interpret the results
d) use AttributeSelectedClassifier (with CfsSubsetEval and GreedyStepwise search) for classifiers J48, IBk and NB, evaluating with cross-validation f) interpret the results.

Create an “arff”-file containing the datapoints
t1 = (4,2,3,5,2,2,2,1) t2 = (3,2,5,4,3,2,1,4) t3 = (1,3,3,5,2,3,2,1) t4 = (4,2,0,5,2,2,2,1) t5 = (3,2,3,4,3,2,1,4) t6 = (2,5,3,5,2,2,2,1) t7 = (4,1,3,7,2,1,2,1) t8 = (3,1,5,4,3,2,1,4) t9 = (2,5,2,5,2,5,2,1)
Cluster the data file using EM with k=2 and k=3 clusters.

Create an “arff”-file containing the following document-word representation (sparse arff file)

t1 = {machine, learning, classifier}
t2 = {data, mining, associative, classifier} t3 = {mining, decision, tree}
t4 = {association, mining, data}
t5 = {decision, tree, classifier}