Lab Exercise 1 – Supervised Learning

Please undertake the following exercises:

* Load play tennis dataset into weka
  + <http://axon.cs.byu.edu/~martinez/classes/478/stuff/labhints/tennis.arff>
  + Train k-nearest neighbor on the data (classifiers > lazy > lbk)
    - Try various values of k. Which has the lowest error?
    - Is this a good classifier for this dataset?
  + Train a j48 Decision Tree on the data set (classifier > trees > j48)
    - What does the resulting tree look like?
      * (right click on the tree in the results list, and select ‘visualize tree’
    - Does it make sense?
    - What’s the classification error?
    - Why do you think that is?
  + Train a pruned DT on the dataset
    - (Click on j48, set reducedErrorPruning -> true)
    - What does the tree look like? Why?
    - What’s the classification error?
    - Look at the confusion matrix. What do you see?
* Load iris dataset into weka
  + <http://mlr.cs.umass.edu/ml/machine-learning-databases/iris/iris.data>
    - (Save as a .arff file)
    - Classification but inputs are continuous
  + Train k-nearest neighbor on the data
    - Change distance weighting from none to “Weight by 1/distance”.
    - How does the mean absolute error change with that setting on vs off? Why?
  + Train a j48 Decision Tree on it
    - What’s the classification error?
    - What’s the tree look like?
  + Train a pruned j48 Tree
    - What’s the classification error?
    - How has the tree changed shape?
* Larger Dataset - Kaggle Titanic Data
  + <https://github.com/birchsport/titanic/blob/master/test.arff>
    - Data is **imbalanced** (290 survived, 1911 died)
  + Train knn on this dataset for a few different values of k (1,5,10,20)
    - Does upping k affect the classification accuracy much?
    - Why do you think this is?
    - Change distance weighting from none to “Weight by 1/distance”.
    - Does this improve the accuracy?
    - Why \ Why not?