Intelligent Data Analysis – Fall 2015

Homework #3

Due Date: October 6th, 2015, 11:59PM

Submit only one pdf file in response to this homework. Do not upload a zip file. Include the code you as text in a file and make it a part of the submitted pdf file.

For this homework you have to use a well-known data set available at the following website:

<http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29>

This is known as the Wisconsin Breast-Cancer Dataset. It has 699 records and 10 attributes, describing various features of a tumor. The eleventh attribute is the class label: benign or malignant. Your tasks for this homework are the following:

1. Download this dataset. It doesn’t need to be normalized because each attribute is valued between 1 and 10.
2. Select 500 records randomly from these dataset for training and keep the remaining 199 for the test set.
3. Create a decision tree with this dataset with the constraint that every leaf node has at least 25 data records. (**Submit This**) From this decision tree report those rules whose leaf nodes have at least 75% class purity. Include rule purity number with each reported rule.
4. Use the test dataset to predict the class labels of its records using the decision tree model created in #3 above. (**Submit This**) Report the precision, recall and F1 metrics of this classifier based on the actual and predicted labels of the test dataset.
5. Use the fitcsvm function of Matlab to train a SVM model for the 500 records of the training set. Use ‘KernelFunction’ parameter with value ‘RBF’ to tell the trainer to use the Radial Basis Function as the non-linear transformation of the data space. Use this model to test the 199 records of the test dataset. (**Submit This**) Report the precision, recall, and F1 metrics of this classifier.
6. (**Submit This**) Compare and contrast the performance metrics obtained in #4 and #5 above. Give reasons for the differences that you observe.
7. (**Submit This**) Assume the cost of making a correct diagnosis is 0; the cost of predicting an actual benign case as malignant is 10 units, and the cost of predicting an actual malignant case as benign is 30 units. Use the results obtained in #4 and #5 above and determine the cost of misclassification for each type of classifier. Show your work for arriving at your answers.
8. (**Submit This**) Pick one record from the test set that is misclassified by the decision tree model. Find its 3 nearest neighbors, using Euclidean distance, in the training dataset. Show the query instance and the nearest neighbors retrieved. Find the class label that this instance should get using the KNN method? Repeat the above steps for 1, 5, and 7 nearest neighbors. Comment on the class labels that you get.