CSC869 Data Mining at San Francisco State University
Project #3: C4.5 classifier and Clustering evaluation
Due: 11:55pm, Friday, May 15, 2015

Part I: Comparing C4.5 and Naïve Bayesian Classifier

You will use the same dataset in your second mini project (the Adult dataset) to build a decision tree classifier by employing an existing implementation of the c4.5 classifier. If you are using Weka, the classifier is named as J48. If you are using Python SciKit Learn, the implementation is called sklearn.tree.DecisionTreeClassifier

What to submit: Report the average classification accuracy of C4.5 using 10-fold cross validation. Then compare this with that from your second mini project.

Part II. Clustering evaluation using the Iris dataset

In this part of the project, you will

1. First cluster the Iris dataset using three different clustering algorithms: (i) k-means, (ii) x-means, and (iii) a density-based clustering algorithm. You can find all three clustering algorithms in the Weka machine learning software. The Iris dataset is included in the Weka machine learning software as one of its sample datasets. You can locate it under the .\data folder. You can also download this dataset online at: http://archive.ics.uci.edu/ml/machine-learning-databases/iris/
2. You will then evaluate the clustering results based on four external evaluation measures (i) purity, (ii) the normalized mutual information, (iii) the Rand index, and (iv) the F measure. A detailed explanation of the 4 external evaluation measures can be found at http://nlp.stanford.edu/IR-book/html/htmledition/evaluation-of-clustering-1.html#fig:clustfg3;
3. Finally, you will evaluate the clustering results using three internal evaluation measures: (i) the diameter of a cluster, (ii) the average link between two clusters, and (iii) the complete link between two clusters. These internal measures can be found on the first set of lecture slides on clustering or in your textbook.

What to submit: Report your evaluation results. Furthermore, compare the three clustering algorithms based on these results.

Submission requirement: please submit a single PDF file on iLearn.