Pawel Micun  
CISH-6180  
Assignment 2  
4/14/14

After removing records with missing attributes from the original data set, we’re left with 683 entries. We randomly split that into two roughly-equal sets: 334-record train set, and 349-record test set.

Tree construction proceeds as follows:

* For a given record set, we perform splits on all the remaining attributes and pick one with the lowest GINI
* For convenience, attributes are treated as discrete rather than continuous. This results in a shallower, but wider tree
* After splitting, process continues recursively until all the records have the same class

No pruning if performed on the final tree.

Classification is performed by walking down the tree to a leaf node. If no leaf node is found (ie a record with this sequence of attributes was not present in the test set), the result is based on neighboring leaf nodes.

SVM construction proceeds as follows:

* Training set attributes/class are used to set up a quadratic problem of the following form  
    minimizex (1/2)xTPx+qTx+r  s.t.   
      Gx ≤ h   
      Ax = b,   
  where P is constructed using the Gaussian Radial Basis kernel
* JOptimizer (<http://www.joptimizer.com>) is used to solve this problem

Classification is performed using: 

|  |  |  |
| --- | --- | --- |
| Classifier | Train accuracy | Test accuracy |
| Tree | 1.0 | 0.93 |
| SVM | 1.0 | 0.97 |

Train accuracy is 1.0 in both cases, since:

* Tree is not pruned
* SVM kernel is flexible enough to cleanly separate the data

