CSC 481 – Machine Learning Lab

Decision Trees

The goal of this lab is to introduce you to decision trees and to a machine learning tool suite called Weka. The objectives:

(a) learn how to load data into Weka and build a decision tree
(b) under stand the structure of generated model
(c) evaluate the decision tree on the training data and understand the quality of the tree – accuracy and confusion matrix

I. Weka

You need to download and install Weka. You can download Weka from this site:
http://www.cs.waikato.ac.nz/~ml/weka/downloading.html
You do NOT need to install it in the VM you can install it on your system directly. It is highly recommended that you download and install the package appropriate for you system that INCLUDES the JRE (Java runtime environment). For MacOS systems this is called the Oracle JVM.

II. Data sets

Once you have downloaded and installed Weka you should download the data sets from the course website.

III. Decision Trees

We will use the J48 decision trees available in Weka. These are more advanced decision tree than covered in class, they allow for numeric attributes and missing values. Try building models for all four data sets: tennis, mushroom, tic-tac-toe, zoo. The ARFF format is a text based format which you can read with just an ordinary program editor such as notepad++. There is a lot of additional information in the comment area of these files. Evaluate the decision trees:

(a) does the model make sense – does it summarize the data so you can understand it?
(b) does the model make mistakes on its training data? If so, how many? Compute the error rate: (number of mistakes)/(total number of observations). Compute the accuracy: 1 – (error rate)

IV. Try out some additional data sets: take a look at

http://repository.seasr.org/Datasets/UCI/arff/