University of Washington
Department of Computer Science & Engineering

CSE 446 - Machine Learning - Winter 2011

CSE 446 Problem Set 1

Due: Friday, January 14 (in class)

This problem set will give you a chance to explore manual data modeling, giving you experience with the task machine learning algorithms are designed to automate.

First, get the data sets from the homework Web page. For this assignment, we've taken the UCI Car Evaluation Database, and created a training and test set. The training set (car.train) contains 1000 records, each on a separate line; the test set (car.test) contains 200 records. Each record contains a number of car attributes (buying, maint, doors, persons, lug_boot, safety), and your models will predict the last field in each record (how acceptable is a car with those features).

The attributes are as follows:

attribute	*values*	*description*
buying	vhigh, high, med, low	buying price
maint	vhigh, high, med, low	price of the maintenance
doors	2, 3, 4, 5more	number of doors
persons	2, 4, more	capacity in terms of persons to carry
lug_boot	small, med, big	the size of luggage boot
safety	low, med, high	estimated safety of the car

The potential values for *acceptability* include: unacc, acc, good, vgood

Using only the training set and your intuitions about the domain, attempt to build a model for the data by hand. For this task, you can use any statistics/data visualization software you choose. This could include Excel (available on departmental Windows machines), MATLAB (available on departmental linux machines -- see http://www.cs.washington.edu/lab/support/faq.html#67), R, etc. As many of these programs come with built-in machine learning algorithms, we stipulate for this assignment that you do not use any sophisticated techniques. Instead, try plotting slices of the data or performing simple calculations to gain intuitions. Any machinery more sophisticated than linear regression should be avoided.

When you are finished, you should have a model that predicts, given a set of attributes, how acceptable a car with those features is. The classification accuracy you report should be the fraction of times your model correctly predicts the car acceptability. After creating your model using the training set, test it against the test set.

Please turn a hard-copy typed write-up (no more than 2 pages) including the following:

1. A description of the steps you took in building your model, including any relevant graphs. Be sure to address the following questions:
    a. What form does your model take? Include a mathematical or pseudo-code specification of your model.
    b. Which of the data set attributes are the most important to the predictions of your model? What led you to choose these?
    c. Were you surprised that some attributes were more or less predictive of car acceptability than you expected?
    d. Based on your answer to (b) and your knowledge of the domain, can you think of other attributes that would be useful for predicting car acceptability and would be good additions to the data set?
2. The classification accuracy of your model on the training and test sets.

Department of Computer Science & Engineering
University of Washington
Box 352350
Seattle, WA 98195-2350
(206) 543-1695 voice, (206) 543-2969 FAX
[comments to *tlin*]