# 36-350: Data Mining

**Homework 11**
**Date: November 19, 2001** **Due: start of class November 30, 2001**

1. The files `wine-tr.dat` and `wine-te.dat` contain the results of a chemical analysis of Italian wines derived from three different types of grape. The task is to detect wines made from a certain grape, as indicated by the variable `Type`.

   (a) Fit a linear logistic regression model to the training set (`wine-tr.dat`) to predict `Type` from the other variables. Evaluate it on the test set.

   (b) Train a nearest-neighbor classifier on the training set and evaluate on the test set. Use cross-validation on the training set to determine the number of neighbors to use. Does it perform better than logistic regression?

   (c) Make a `predict.plot` for `Type`. Does it give reasons for the relative performance of the classifiers?

2. (a) Repeat parts (a) and (b) of the last problem, using only the two predictors `Phenols` and `Flavanoids`.

   (b) Make a `cplot` of the two classifiers and give reasons for their relative performance.

   (c) Would you expect a classification tree to do better or worse than logistic regression on this problem? Explain.

   (d) A new wine is observed to have `Phenols=2.5` and `Flavanoids=3`. Using the logistic regression model found in part (a), what is the probability that `Type = Yes`?

3. Four vehicles were observed at several different camera angles, and the silhouette of each vehicle was extracted. We want to discriminate vans versus other vehicles, based on the silhouette alone. The files `Vehicle-tr.dat` and `Vehicle-te.dat` contain a training set and test set of labeled silhouettes. Describing each silhouette are 18 variables measuring properties like circularity (`Circ`) and elongatedness (`Elong`). The variable `Class` indicates a van.

   (a) Train a logistic regression classifier and evaluate it on the test set.

   (b) Do the same for a cross-validated nearest-neighbor classifier.

   (c) Using arguments given in class, give possible reasons for the relative performance of the two classifiers on this task. If one of them is bad, why isn't the other bad for the same reason?