36-350: Data Mining

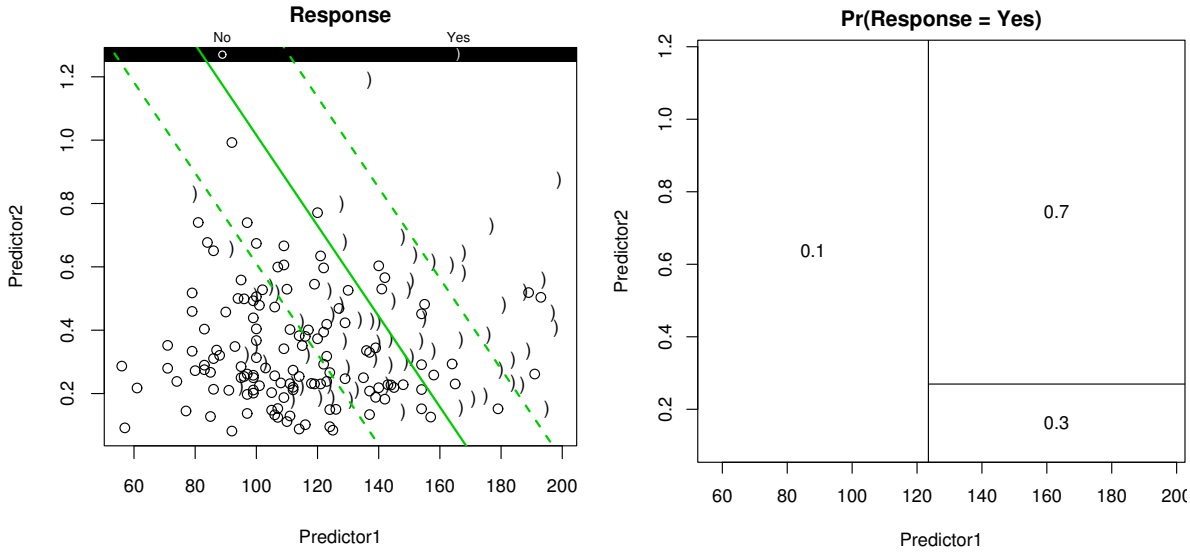
Homework 13
Date: November 17, 2003 **Due: start of class November 24, 2003**

---

1. A logistic regression model and a cross-validated pruned tree are constructed from a dataset with two predictors. The logistic regression has the following coefficients:

```
(Intercept)   Predictor1   Predictor2
   -6.62404      0.03873      2.70761
```

The class probabilities computed by each model are depicted below. Left is a contour plot of the logistic probabilities, with contours at 0.25, 0.5, and 0.75. Right is the probability of "Yes" in each tree partition.



(a) Suppose a test point has `Predictor1=130` and `Predictor2=0.4`. What is the probability that this point has `Response=Yes`, under each model?

(b) If false positive classifications and false negative classifications have very different costs, which classifier should you prefer? Explain.

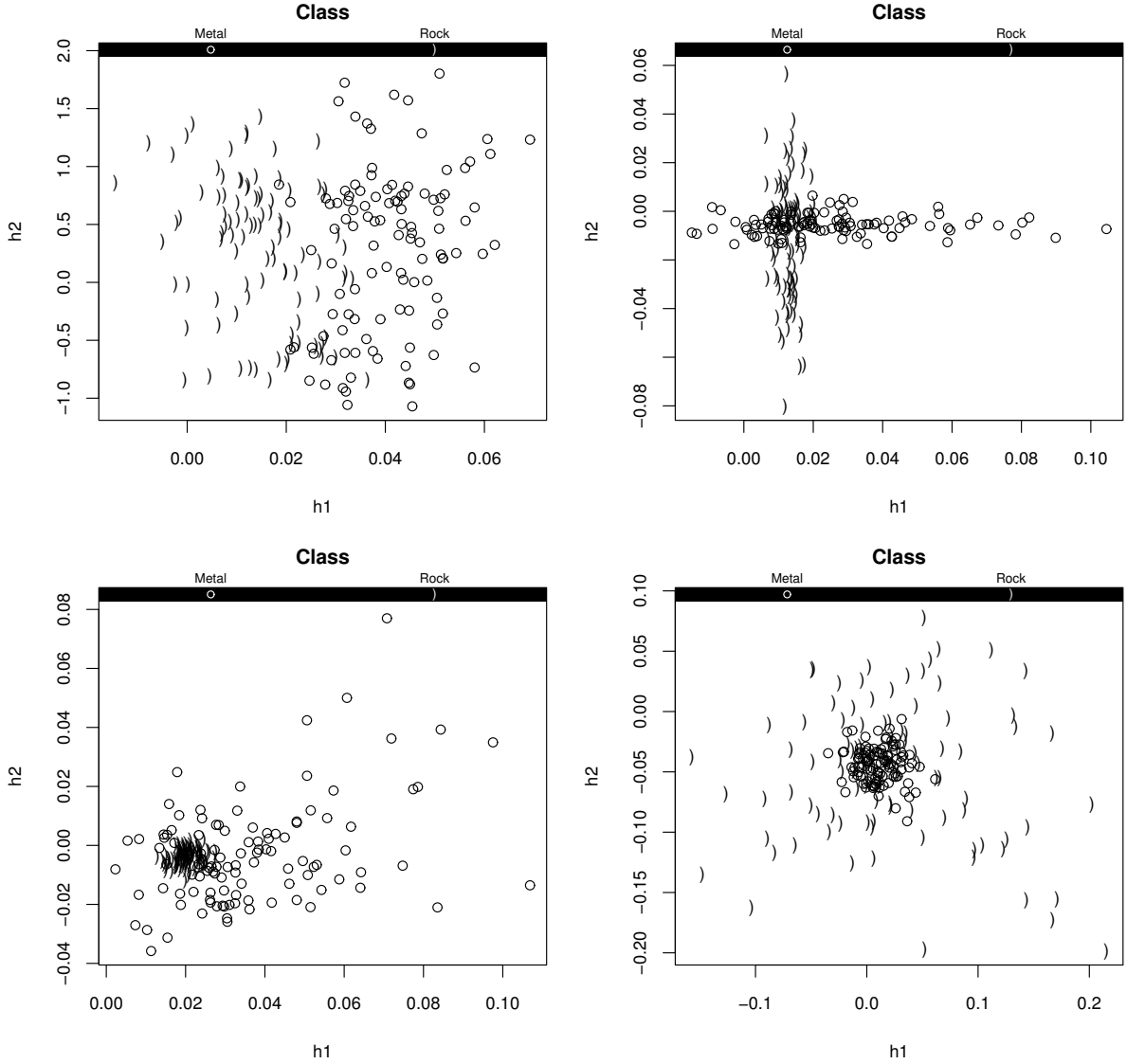
2. In the computer lab, you compared a tree, nearest-neighbor, linear, and quadratic classifier on the credit dataset.

(a) What does the performance of the linear classifier, relative to the tree and k-nn, suggest about the credit data? (More specific answers get more credit.)

(b) What does the performance of the quadratic classifier suggest about the credit data?

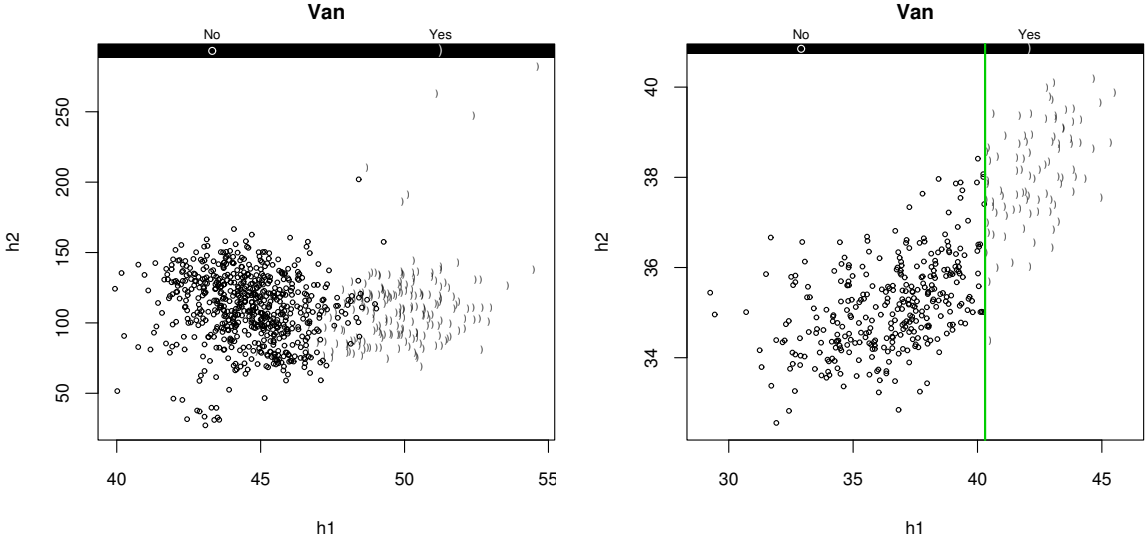
(c) What does the performance of KNN and its best value of k suggest about the credit data?

3. Sonar works by broadcasting a well-chosen sound wave and analyzing the frequency content of the echo. A useful task is to discriminate the echo of metal objects from those of rocks. The dataset contains 208 sonar echoes, some of which came from metal cylinders and others from rocks. Each echo is represented by the energy in 60 different frequency bands. Below are four different discriminative projections of this high-dimensional data (`h1` and `h2` have different meanings in each plot):



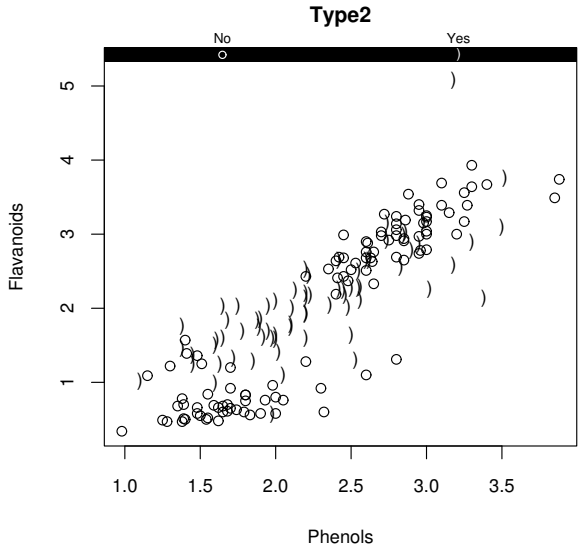
Based on the data description and the plots, which classifier should you expect to perform best on a holdout set: linear, quadratic, or nearest-neighbor? Explain why. (The classifier would be able use all of the original variables, not just the projections above.)

4. Using a camera and a computer, it is possible to automatically identify cars on the road, even at night. A dataset was collected in which 846 vehicles were observed at several different camera angles, and the silhouette of each vehicle was extracted. The goal is to discriminate vans versus other vehicles, based on the silhouette alone. Each silouette is described by 18 variables measuring shape properties like circularity (`Circ`) and elongatedness (`Elong`). Below are two different projections of this data (the projection on the right shows the boundary of a linear classifier):



Based on the data description and the plots, which classifier should you expect to perform best on a holdout set: linear, quadratic, or nearest-neighbor? Explain why.

5. Wines made from different grapes have slightly different chemical properties, which lead to different tastes. It is useful to determine exactly what those distinguishing properties are. This can be done by constructing a classifier to discriminate the wines. A dataset was constructed which describes 178 wines in three categories by two different chemical properties. To make it a Yes/No problem, the wines are grouped into "Type 2" and "Not type 2". Below is a plot:



Based on the data description and the plot, which classifier should you expect to perform best on a holdout set: linear, quadratic, or nearest-neighbor? Explain why.