MthStat 568/768 – Multivariate Statistical Analysis – Spring 2024

Homework 5

Due Wednesday, April 10

1. The pendigit dataset contains digitalized handwritten digits. The variables $x1, y1, \ldots, x8, y8$ are the coordinates of a pen on a writing pad at eight different time points (so, if you want to visualize the written digit, you have to do plot(c(x1,...,x8),c(y1,...,y8),type="l").) The variable digit identifies the digit that was written. The goal is to construct a classifier that will identify the handwritten digits as accurately as possible.

   Split the data into training and test sets (roughly an 80/20 split). Fit single-layer neural networks to the training data, with one, two and three hidden nodes (or more if necessary). Compute the respective misclassification rates on the test set.

   What's the lowest misclassification rate attained? From the cross-classification tables, which digits have the largest misclassifiaction rates?

2. The spambase dataset contains data for 4,601 emails which are classified as spam or not spam (as indicated by the variable class); 58 feature variables are measured on each email. A more detailed description of the data is given on p. 259 of the book.

   Split the data into training and test sets (roughly an 80/20 split). Fit a linear support vector machine classifier to the training data, starting with a very large ("infinite") cost, in the event the groups are separable, and progressively lowering the cost if they aren't. Compute the respective misclassification rates on the test set.

   What's the lowest misclassification rate attained?