MthStat 568/768 – Multivariate Statistical Analysis – Spring 2024

Final Work

Due Saturday, May 18

1. Consider the vehicle dataset. The goal is to identify 3D objects from 2D images captured by cameras at different angles. The objects in this case are four types of vehicles (identified by the variables class and classdigit), and the other 18 numerical variables are measurements extracted from these 2D images.

   Split the data into training and test sets (80/20). On the training set compute:

   - a multinomial logistic classifier
   - a single-hidden-layer neural network classifier (with the number of hidden nodes to be determined)

   On the test data, find the cross-classification tables and the misclassification rates. Which of the above methods is better? Is there any specific type of vehicle that is harder to classify than the others?

2. The pendigits dataset consists of discretized handwritten digits (for a full description, see Section 7.2.10 in the book). From this set, extract the subset corresponding to digits 0, 6, 8 and 9, and scale the variables so that the variances are 1.

   (a) Compute ordinary principal components and draw a scatterplot of the first two component scores, using different colors (or symbols) for different digits. Are the digits well separated?

   (b) Compute kernel principal components using Gaussian kernels with various scales, and draw scatterplots of the first two component scores as in (a). Are the digits now better separated than in (a)?