Homework 6 – Due 11:59 pm CT, 7 April 2021

The total points on this assignment is 100. The dataset is posted on the class website. All problems are worth the points indicated in the parenthesis.

1. Consider the crabs dataset in R used in Exam 1.

   (a) Use principal components analysis to reduce the dimensionality of the crabs dataset into two dimensions. Display the results. Is there any distinctiveness in the four species/sex combinations? [*10 points*]

   (b) Perform a kernel principal components analysis with two features and display the results. You may use the Gaussian radial basis funcion. But display the results for different values of $\sigma = 0.2, 0.4, 0.8, 1.0, 1.5, 3$. [*20 points*]

2. The United States Postal Service has had a long-term project to automating the recognition of handwritten digits for zip codes. The file available in the usual places has data on different numbers of specimens for each digit. Each observation is in the form of a 256-dimensional vector of pixel intensities. These form a 16x16 image of pixel intensities for each letter. The objective is to distinguish one digit from another.

   (a) We will see whether the digits are distinguishable. To do so, we will first prepare the dataset by rooting out those pixels (coordinates) which do not contribute to categorization. Do so, using appropriate univariate but simultaneous methods, at the 5% level. For the remainder of this question, we will be focused on the 100 most significant coordinates (in terms of the p-value for the above test). [*10 points*]

      i. Evaluate whether the variance-covariance matrices are all equal across all digits. [*10 points*]

      ii. If these are not equal, we will assume that they are the true values of the individual group dispersion matrices. Derive a likelihood ratio test statistic for testing differences in mean effects across different digits. [*10 points*]

   (b) We will now use principal components to reduce dimensionality of the *original dataset*. Note that the images for the different digits have different means and chanracteristics, therefore, it would be preferred to remove the effect of the digit-specific means before performing the principal components analysis. (Transformed versions of these means need to be added back before proceeding much further.) Use the principal components and determine the number of components needed to explain at least 80% of the total variation in the data, at the 5% level of significance. [*20 points*]

      i. Display the components (using color or characters for each digit) using appropriate methods. Compare with the displays obtained using the reduced dataset. [*10 points*]