Homework 8 – Due 11:59 pm CST, 1 May 2021

The total points on this assignment is 125.

1. Gamma Ray Bursts (GRBs) are the brightest known electromagnetic events known to occur in space and have been studied extensively ever since their discovery in the late sixties. While the cosmological origin of GRBs is well-established, questions on their source and nature remain unresolved. The Burst and Transient Source Experiment (BATSE) Catalog of the National Aeronautics and Space Administration (NASA) provides temporal and spectral information for many GRBs. Of interest to us are the parameters:

   T_{50}: The time by which 50% of the flux arrive.

   T_{90}: The time by which 90% of the flux arrive.

   P_{64}, P_{256}, P_{1024}: The peak fluxes measured in bins of 64, 256 and 1024 milliseconds, respectively.

   F_1, F_2, F_3, F_4: The four time-integrated fluences in the 20-50, 50-100, 100-300, and $>$ 300 keV spectral channels, respectively.

   The current BATSE catalog, that is, the BATSE 4Br contains bursts from 1973 Gamma Ray Bursts but only 1599 of them are observed in all nine coordinates.

   Chattopadhyayand Maitra (2017) have recently established that there are five distinct kinds of gamma ray bursts that are ellipsoidally-shaped in the logarithmic scale. The R data object `GRB-5groups.rda` contains this classification and the observations (in the logarithmic scale) for each of the 1599 GRBs.

   We will use this dataset to obtain discriminant rules. Answer the following questions:

   (a) Evaluate if the five groups support multivariate normality distributional assumptions. [*10 points*]

   (b) Assuming equal prior probabilities and equal costs of misclassifcation, construct Fisher's linear discriminant function. [*10 points*]

      i. Display the first two linear discriminant coordinates. Do all the variables in the discriminant function appear to be important? [*10 points*]

      ii. Calculate the misclassification rates using the AER and the leave-one-out cross-validation method. [*10 points*]

   (c) Repeat the same exercise as in (b) but using quadratic discriminant analyss, CART and k-nearest neighbors. Choose the best tree or the number of nearest neighbors by cross-validation. Summarize the performance of the results for all four cases (LDA, QDA, CART and k-NN). [*10 + 15+10+5 points*]

   (d) For each of the groups, test the hypothesis that a fewer number of factors is adequate to express the variables in the dataset. [*15 points*]

2. Compare the results of clustering obtained using the Women's and Men's track records results. Use hierarchical clustering with the correlation similarity matrix, k-means and model-based clustering. Display the results using appropriate graphical aids. Comment. [*30 points*]