Homework 5

Due on 05/05/2022

1. In this problem, we will apply support vector machines to predict whether a given car gets high or low gas mileage based on the dataset "auto.csv" (used in Homework 3; see Homework 3 for more details of the dataset). The response variable is `mpg_cat`. The predictors are `cylinders`, `displacement`, `horsepower`, `weight`, `acceleration`, `year`, and `origin`. Split the dataset into two parts: training data (70%) and test data (30%).

(a) Fit a support vector classifier (linear kernel) to the training data. What are the training and test error rates?

(b) Fit a support vector machine with a radial kernel to the training data. What are the training and test error rates?

2. In this problem, we perform hierarchical clustering on the states using the `USArrests` data in the `ISLR` package. For each of the 50 states in the United States, the dataset contains the number of arrests per 100,000 residents for each of three crimes: `Assault`, `Murder`, and `Rape`. The dataset also contains the percent of the population in each state living in urban areas, `UrbanPop`. The four variables will be used as features for clustering.

(a) Using hierarchical clustering with complete linkage and Euclidean distance, cluster the states. Cut the dendrogram at a height that results in three distinct clusters. Which states belong to which clusters?

(b) Hierarchically cluster the states using complete linkage and Euclidean distance, after scaling the variables to have standard deviation one.

(c) Does scaling the variables change the clustering results? Why? In your opinion, should the variables be scaled before the inter-observation dissimilarities are computed?