Final Project (Part 2)

Problem 1 and Problem 2 are based on the "heart.csv" data. The "heart.csv" data is massive, you can randomly choose a few hundred rows and focus on these data only.

Problem 1

Use spectral clustering to separate the data into two clusters (without using the "Heartdiseaseorattack" column). How well do the clusters agree with the "Heartdiseaseorattack" label? (For this question, choosing a proper dissimilarity measure is crucial, make sure you explain and justify your choice).

Problem 2

Use the same dissimilarity measure in problem 1 to design a nonconformity measure. Randomly select half of the data as training data, and construct 90% prediction region for the rest testing data. How well does your prediction result agree with the clustering result in Problem 1?

Problem 3

Read the paper "LAWS: A Locally Adaptive Weighting and Screening Approach To Spatial Multiple Testing". Replicate the numerical result in Section 5.2 (You only need to do this for the data-driven LAWS procedure. There are many ways to choose the bandwidth h, you do not have to use the kedd package.)