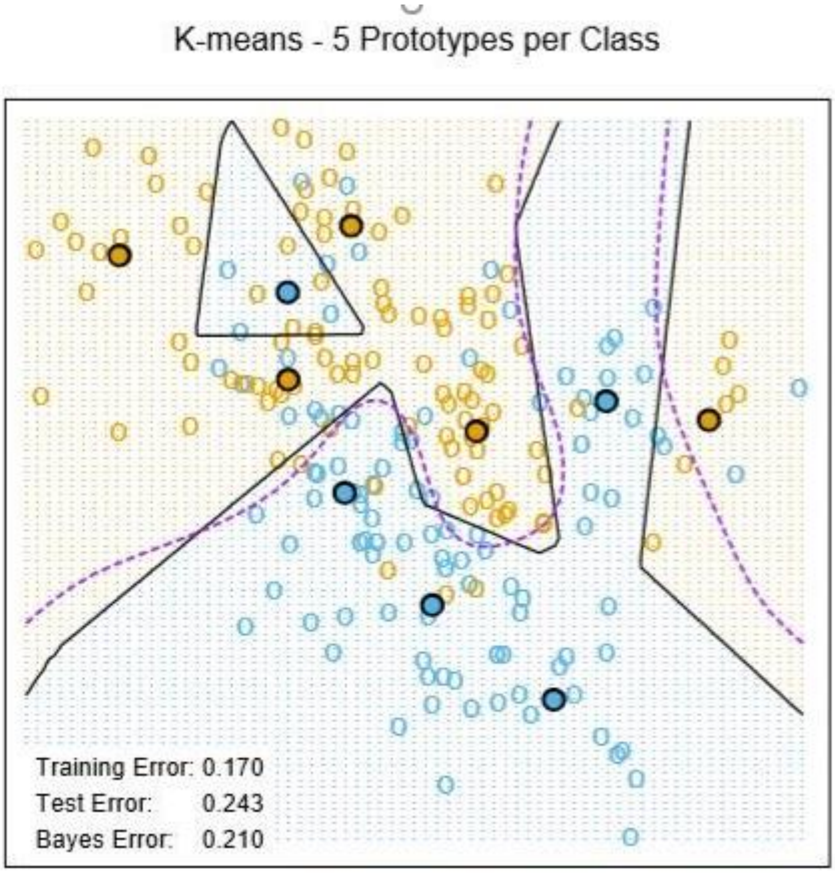
CS599 Project1 report

Qingkun Liu

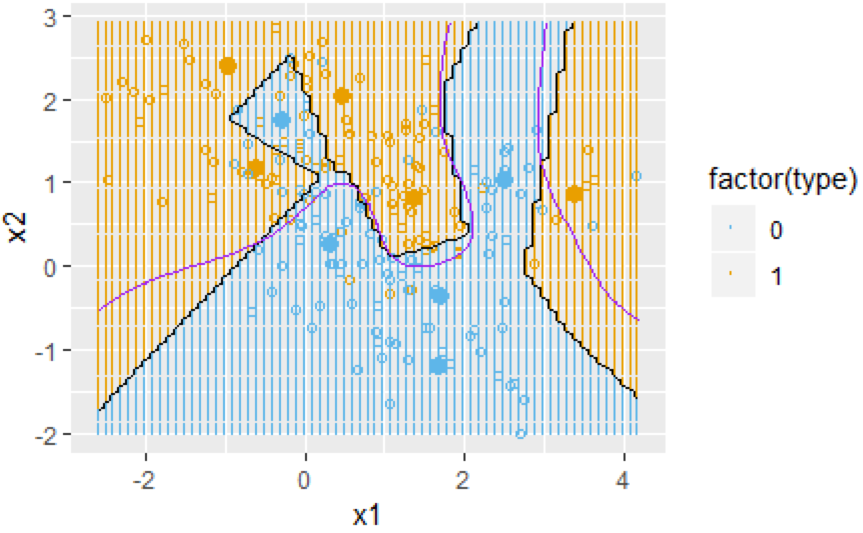
1. Figure and Citation:



Citation:

Hastie, et al. Elements of Statistical Learning, Figure 13.2

2. Reproduced figure:



3. Difference

   From the figure that I create, you can find it is not absolutely same with original figure. The reason is the result of k-mean clustering algorithm depend on the initial center points, which means that is we choose different initial centers, the result will be different. However, we cannot find the initial points that the original figure use in the citation. We just know the initial center points are generated randomly. So, I set a seed and try different value. Finally, I only can generate a figure that similar with the original figure.

4. Software

   In this project, I use python for calculating and Rstudio for draw figure.

   In python part, I use three libraries. First, I use csv package to read the csv file. Then, I use numpy package to do some basic calculation (just like sqrt, mean, np.array). Finally, I use random package to generate initial center.

   In Rstudio part, I use ggplot2 library to draw the scatter figure

   I write k-mean algorithm by myself, instead of using the kmeans package that in a library of python, I come true the algorithm by myself.

5. Issue and problem

   The first problem is choosing seed. As I have talked about, the initial centers will affect the result. To generate a figure that similar with original figure, I need choose fit seed. So, I try the seed from 0 to 200 and use the best one. Although the figure that I generate seem different with original figure. It is the best one that I have checked.

   Another problem is how to draw the decision boundary. I try to use python draw figure. However, I only can draw the points and I cannot draw decision boundary. Then, I must use Rstudio to finish the figure.