Clara Cullum

Econ 8600

Final Exam Responses

E.) Based on the silhouette scores, all three models predict that 2 groups lead to the most well defined clusters. Because the data are not spherical or well defined, Gaussian mixture model may be the most suitable. At a large number of groups, however, GMM receives lower silhouette score (around .21-.25) compared with KMeans, whose silhouette scores are around .4 for groups of 6 or higher. Therefore, if you are looking for a large number of groups, KMeans may be best. Further, if 4 groups are desired (to fit with the 4 factors), KMeans returns the highest silhouette score of about .5 compared to KMedoids (~.4) and GMM (~.45). Therefore, the best algorithm will depend on your use of the data.

G.) Regressing each factor on math ability yields very small R^2 scores, which would indicate that the variation in math scores is not well predicted by the factors in a linear model. Considering the plots of each factor on math ability, neither linear nor logistic regression would work well to explain the data as most factors are highly heteroskedastic.

H.) First, you would run 40 regression models, where y is math ability and x is a question from the original questionnaire. You would also want to obtain p-values and R^2 scores for each. Then, you would select questions that are statistically significant ($p < .05$) and have high R^2 which indicates it has predictive value on math ability. Then, you will administer the test to recruits. The original data sample we have will act as the testing dataset. The new test responses (from recruits) will act as the test data. You would create a new dataset from the old sample with

just each individual's responses to the 20 questions selected. Then, using KFold, you would split it into two to use half as the training data and half as the training target. You would create a machine that regresses the 20 question responses on math ability. If it has a sufficient R^2, you would then use it to predict half of the new recruits' math abilities. If this test proves sufficient, you use ALL of the data (training data, training target, and test data) to predict the math abilities of the remaining recruits. Finally, the 30 recruits with the highest math ability are selected.

I.) For this problem, you would choose 5 questions that correspond to each of the 4 factors of personality (have high eigenvalues for one of the 4 factors). This will create an abridged version of the original survey with only the most poignant questions. Then, after administering the survey to new recruits, you would use the eigenvalues from the original sample to sort these people along the 4 factors. Finally, you will use one of the clustering methods to obtain 30 groups. I would use KMedoids as the central point of these groups is always one of the individuals, and I would select each of the center points as participants in the project.