MSiA 421
Professor Edward Malthouse

Homework 2

Due: January 23, Start of class

Submit one homework per group. Put all names on the homework.

1. This data uses data from `tradeshow.csv`. You have the following variables:

   - **Buy**. Evaluate and compare specific equipment for purchase, place orders, find new suppliers and solutions.
   - **Social**. Spend time with others, network with colleagues, extend professional network.
   - **Education**. Keep up-to-date on industry trends, attend continuing education sessions, attend keynotes.

   (a) Estimate K-means on the three variables and find the 3-cluster solution. Give the cluster sizes, means and RMSE values. Describe each of the three clusters

   (b) Estimate a Gaussian mixture using the three variables in R with the options `G=3` for three clusters and `modelNames="VII"` for unequal variance, round clusters (`spherical` in Python). Submit a classification plot. Compare the solution to K-means.

      i. Do the cluster means tell the same story, or are there differences?
      ii. Comment on the K-means vs. GMM cluster sizes.
      iii. Comment on the within cluster standard deviations (vs. RMSE for K-means).
      iv. How many variance parameters are estimated in total?

   (c) Estimate Gaussian mixtures using three variables only with the `G=3` option (use `tied` in Python).

      i. Do the cluster means tell the same story, or are there differences?
      ii. Generate a classification plot.
      iii. Which variance model did Mclust pick (it should be EEE)? Describe in words the shape of the class-conditional distributions.
      iv. How many variance parameters are estimated in total?

   (d) Which of the three solutions do you prefer?

2. This problem uses a data set from the Nuoqi retailer in China. You have five factors measuring attitudes toward fashion: Cross, fashion enthusiast, functional, impressive, self-expression. See the Powerpoint for the actual questions that were asked of consumers the alpha values.

   (a) Use K-means to find the five-cluster solution using the first five variables in the data frame. Give the usual sizes, means, and RMSE values. Comment on the solution.

   (b) Suppose that there are individual differences in the way that different respondents use the sales, where some are systematically more positive and others are more negative. The variable `xbar` is the average response *for the given respondent* to all 5-point scales on the survey. Compute five new variables equal to the original variable minus xbar, e.g., `nuoqi$impressI = nuoqi$impress-nuoqi$xbar`. This is called *ipsatization*, and it will be important to us with recommender systems.

(c) Use K-means to find the five-cluster solution using the ipsatized versions of the first five variables in the data frame. Give the usual sizes, means, and RMSE values. Comment on the solution. Is there improvement?

(d) Run K-means solutions for the K =2–6 solutions and examine the fit statistics (SSE, R-Squared, Pseudo F). Try both the raw and ipsatized data. Which do you suggest?

(e) Try Gaussian mixture models and look at the plots. What is the underlying problem when trying to cluster this data set?

3. Write a function to generate data for this problem with parameter μ. There are $K = 2$ equal-sized clusters with one cluster sampled from $\mathcal{N}(-\mu, \sigma^2)$ and the other from $\mathcal{N}(\mu, \sigma^2)$, where $\sigma^2 = 1$. Assume $n_1 = n_2 = 3000$ observations from each (`Mclus` will start to have problems for larger n, but Python should be able to handle somewhat larger sample sizes, and K-means can easily handle much larger n). Estimate GMM and K-means models for $\mu = 0.5, 1$ and 2. Report the estimated means and variances. Discuss the results, especially the biases discussed in class and how they are affected by the separation of the means.