STAT440 Final Project Proposal

Group 39: Model-Based Classification

Chenxi Zhang Ziqing Chen Jiabo Liu

Introduction

A novel model-based classification technique is introduced based on **Parsimonious Gaussian Mixture Models (PGMMs)**, which were introduced recently as a model-based clustering technique, arise from a generalization of the mixtures of factor analyzers model and are based on a latent Gaussian mixture model, such that

$$z_i \overset{iid}{\sim} Multinomial(K, \rho)$$

$$x_i|z_i \overset{ind}{\sim} N(\mu_{z_i}, \Sigma_{z_i})$$

where $z_i \in \{1.....K\}$ is drawn from a multinomial distribution, and subsequently x_i is drawn from the cluster distribution corresponding to z_i. Therefore, classification reduces to estimate the unobserved for each observation. In addition, according to the **MAP** decision rule, we have

$$\pi_i = \pi(x_i, \Psi) = \arg\max_{1 \le k \le K} Pr(z_i = k|x_i, \Psi)$$

where

$$Pr(z_i = k|x_i, \Psi) = \frac{\rho_k \cdot p(x_i|z_i = k)}{\sum_{m=1}^{K} \rho_m \cdot p(x_i|z_i = m)} = \frac{\rho_k \cdot \varphi(x_i|\mu_k, \Sigma_k)}{\sum_{m=1}^{K} \rho_m \cdot \varphi(x_i|\mu_k, \Sigma_k)}$$

where Ψ is unknown but could be computed via the EM algorithm.

Importance of the Question

The importance of studying this problem is that for some problem it is hard to do real sampling. For example, doing a sampling about the probability of windy, rainy and afternoon respectively is easy. However, if we want a joint sampling to combine probability of windy, rainy and afternoon is hard. Then we can use Gibbs sampler.

Understanding

The mission is to build an R package that includes the essential implantation of Gibbs Sampler and the inference of the parameters of each conditional distribution to achieve classifying multivariate observation $x_1.....x_n$ into K clusters. The most basic idea is to use the mixture-normal model from the introduction for x_i, such that

$$x_i \overset{iid}{\sim} \sum_{k=1}^{K} \rho_k \cdot N(\mu_k, \Sigma_k)$$

With the property of mixture model, the classification process can be treated as

$$z_i \overset{iid}{\sim} Multinomial(K, \rho)$$

$$x_i|z_i \overset{ind}{\sim} N(\mu_{z_i}, \Sigma_{z_i})$$

which means that first use a multinomial distribution to draw K zs act as cluster membership indicator, and then use z as the condition in the distribution to draw xs. So the quantity of interest is

$$\pi_i = \pi(x_i, \Psi) = \arg\max_{1 \le k \le K} Pr(z_i = k|x_i, \Psi)$$

where

$$Pr(z_i = k|x_i, \Psi) = \frac{\rho_k \cdot p(x_i|z_i = k)}{\sum_{m=1}^{K} \rho_m \cdot p(x_i|z_i = m)} = \frac{\rho_k \cdot \varphi(x_i|\mu_k, \Sigma_k)}{\sum_{m=1}^{K} \rho_m \cdot \varphi(x_i|\mu_k, \Sigma_k)}$$

However, this idea is too idealistic to accomplish since the process will break cause the dimension of x_i will eventually become very large. So the approach we choose is to eliminate the not-so-useful x_i first to reduce the dimension of x_i, where only the meaningful x_i remains, such that

$$\theta_i \overset{iid}{\sim} \sum_{k=1}^{K} \rho_k \cdot N(\mu_k, \Sigma_k)$$

$$x_i|\theta_i \overset{ind}{\sim} f(x_i|\theta_i)$$

where $f(x)$ is a family of models. Then, consider the following approximation which dramatically reduces the complexity of the problem above. Instead, in many settings, the hierarchical clustering model above can be approximated and rewrite as

$$z_i \overset{iid}{\sim} Multinomial(K, \rho)$$

$$\theta_i|z_i \overset{iid}{\sim} N(\mu_{z_i}, \Sigma_{z_i})$$

$$y_i|\theta_i \overset{iid}{\sim} N(\theta_i, V_i)$$

where $Y = (y_1....y_N)$ is the observed data, and $Z = (z_1....z_N)$ and $\theta = (\theta_1....\theta_N)$ are the missing data, such that the loglikelihood for the model parameters Ψ given the complete data z, θ, Y.

Planning

In this project, we aim to create an R package containing an implementation of the Gibbs sampler above. In this case, we will need to consider more directions and confirm the result we get above. However, it is simpler to use the Gibbs sampler within a Bayesian approach which we will do more researches on. The sampling model will be built the MCMC sampler **Stan**. Specifically, we would all learn the knowledge of how to compute the flat prior on Ψ and the corresponding posterior distribution. Moreover, all of us will also do more researches on how to do the conditional update for each parameter here, and the conditional update of each parameter will be done in R (or Rcpp, will do more experiments later). The implentation works will be evently divided into three part and be assigned to each team member, the R package will include the functions and the parameters inference based on the file **nnm-functions.R**. After that, we will provide unit tests for the parameters of each conditional distribution by comparing the normalized conditional log-PDF to the unnormalized complete data loglikelihood, and we may experiment with the cluster model in a real dataset. To be more specific, our team will try to finish the coding part before March 31, and start the report right after the implentation.