Quant III

Lab 9: Mixture Model

Junlong Aaron Zhou

November 17, 2020

Mixture Model

- A latent variable problem
- Two different distribution $N(1,1)$ and $N(3,1)$ mixed together (Same thing as structure zero)
- Heterogeneous treatment effect across two groups and we don't know the group label
- Consequence: wrong model $\rightarrow$ wrong estimates and wrong inference

Some diagnosis

- Predictive checks
- Suppose your model is true, generate statistics from your DGP
- Compared with the true statistics
- If not comparable, something wrong with your model

Gaussian mixture model

- N observations
- K groups, with proportion of π_k
- Each group has its own distribution $N(\mu_k, \sigma_k^2)$
- Let z_i denotes observation i's group label
- Ex ante, what is $Pr(z_i = k)$ for some k?

Gaussian mixture model ctd.

- π_k
- Given $z_i = k$, we know $y_i | z_i = k \sim N(\mu_k, \sigma_k^2)$
- Equivalently, we know $f(y_i | z_i = k) = \phi(\mu_k, \sigma_k^2)$
- $f(y_i) = \int f(y_i | z_i) f(z_i) dz_i$
- $y_i \sim \sum_1^K \pi_k N(\mu_k, \sigma_k^2)$
- Now we know the likelihood!

Likelihood estimation

- For simplicity, let $\theta_k = (\mu_k, \sigma_k^2)$ and $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\sigma^2})$
- For observation i, $L(y_i|\boldsymbol{\theta}) = \sum_1^K \pi_k f(y_i|\theta_k)$
- Observed data likelihood: $l = \sum_{i=1}^K log(\sum_1^K \pi_k f(y_i|\theta_k))$
- This could be enough! But it might be hard to solve.

Data augmentation

- Suppose we can observe the label:

$$\begin{aligned} f(y_i, z_i = k|\boldsymbol{\theta}) &= f(y_i|z_i = k, \boldsymbol{\theta}) f(z_i = k|\boldsymbol{\theta}), \\ &= \phi(y_i|\theta_k) \Pr(z_i = k). \end{aligned}$$

- Write this more compactly (for any k) as

$$f(y_i, z_i|\boldsymbol{\theta}) = \prod_{k=1}^{K} (\phi(y_i|\theta_k)\pi_k)^{1\{z_i=k\}}$$

Data augmentation ctd.

- Complete data likelihood (all data):

$$L^{comp}(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{z}) = \prod_{i=1}^{n}\prod_{k=1}^{K}(\phi(y_i|\theta_k)\pi_k)^{1\{z_i=k\}}.$$

- Complete data log-likelihood (all data):

$$\ln L^{comp}(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{z}) = \sum_{i=1}^{n}\sum_{k=1}^{K} 1\{z_i = k\}(\ln\phi(y_i|\theta_k) + \ln\pi_k).$$

EM algorithm

EM Algorithm

1. Initialize randomly $\boldsymbol{\theta}$
2. Repeat (a) and (b) until convergence:
    1. "E-step": given current estimate of $\boldsymbol{\theta}$, compute $E(\ln L^{comp}(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{z}))$
    2. "M-step": update $\boldsymbol{\theta}$ by maximizing $E(\ln L^{comp}(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{z}))$

- **Intuition 1**: The EM algorithm is a coordinate-wise hill-climbing algorithm with respect to the likelihood function
- **Intuition 2**: If we knew label z_i, we could get MLE directly. Even if we don't, posterior can tell us information

EM algorithm: Sketch of proof

- E-step: conditional on the $\boldsymbol{\theta^t}$ we estimate in iteration t, we calculate

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta^t}) = E_Z(log\ P(\boldsymbol{Y}, \boldsymbol{Z}|\boldsymbol{\theta})|\boldsymbol{Y}, \boldsymbol{\theta^t}) = \sum_Z log\ P(\boldsymbol{Y}, \boldsymbol{Z}|\boldsymbol{\theta})P(\boldsymbol{Z}|\boldsymbol{Y}, \boldsymbol{\theta^t})$$

- M-step: calculate $\boldsymbol{\theta^{t+1}} = \underset{\boldsymbol{\theta}}{argmax}\ Q(\boldsymbol{\theta}, \boldsymbol{\theta^t})$

EM algorithm: Sketch of proof Ctd.

1. Show that log likelihood $l(\boldsymbol{\theta}) \overset{def}{=} log\ P(\boldsymbol{Y}|\boldsymbol{\theta}) = log \sum_Z P(\boldsymbol{Y}|\boldsymbol{Z}, \boldsymbol{\theta})P(\boldsymbol{Z}|\boldsymbol{\theta})$.
2. For a fixed $\boldsymbol{\theta^t}$, show that $l(\boldsymbol{\theta}) \geq B(\boldsymbol{\theta}, \boldsymbol{\theta^t})$, where

$$B(\boldsymbol{\theta}, \boldsymbol{\theta^t}) \overset{def}{=} l(\boldsymbol{\theta^t}) + \sum_Z P(\boldsymbol{Z}|\boldsymbol{\theta^t}, \boldsymbol{Y}) log \frac{P(\boldsymbol{Y}|\boldsymbol{Z}, \boldsymbol{\theta})P(\boldsymbol{Z}|\boldsymbol{\theta})}{P(\boldsymbol{Z}|\boldsymbol{Y}, \boldsymbol{\theta^t})P(\boldsymbol{Y}|\boldsymbol{\theta^t})}$$

3. Show that $\boldsymbol{\theta^{t+1}} \overset{def}{=} \underset{\boldsymbol{\theta}}{argmax} B(\boldsymbol{\theta}, \boldsymbol{\theta^t})$ also maximizes $Q(\boldsymbol{\theta}, \boldsymbol{\theta^t})$.
4. Show that $P(\boldsymbol{Y}|\boldsymbol{\theta^{t+1}}) \geq P(\boldsymbol{Y}|\boldsymbol{\theta^t})$, where $\boldsymbol{\theta^t}$ and $\boldsymbol{\theta^{t+1}}$ are calculated in iterations.
5. It converges if $P(\boldsymbol{Y}|\boldsymbol{\theta})$ is bounded.

Bayesian Approach

- Specify priors for each cluster-specific regression:

$$p(\boldsymbol{\mu}_k) \propto 1$$
$$p(\boldsymbol{\sigma}_k^2) \propto 1/\sigma^2$$

- Specify priors for cluster-assignment probability:

$$(\pi_1, ..., \pi_K) \sim Dirichlet(\alpha_1, ..., \alpha_K)$$

Group label

- How should we label?

$$\Pr(z_i = k | y_i, \boldsymbol{\theta}) = \frac{\phi(y_i|\theta_k)\pi_k}{\sum_{k=1}^{K} \phi(y_i|\theta_k)\pi_k}$$

In a typical classification problem

$$z_i = \underset{k}{argmax} \Pr(z_i = k | y_i, \boldsymbol{\theta})$$