Analyses of daily COVID-19 cases across nations

Group11: Sibei Liu, Xue Jin, Yuchen Qi, Xinru Wang

05/01/2020

Objective

Statistical Methods

Adam Algorithm

Guassian mixture model (with EM algorithm)

Cluster analysis is a method for finding clusters with similar characters within a dataset. And clustering methods can be divided into probability model-based approaches and nonparametric approaches[1]. The probability model-based approach contains Gussian Mixture Method, which assumes that the dataset follows a gussian mixture mixture distributions.

Given that $\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\} \in \mathbb{R}^p$ be a collection of p dimensional data points. Assuming the following equation:

$$x_i \sim \begin{cases} N(\boldsymbol{\mu}_1, \Sigma_1), \text{with probability } p_1 \\ N(\boldsymbol{\mu}_2, \Sigma_2), \text{with probability } p_2 \\ \quad \vdots \qquad , \qquad \vdots \\ N(\boldsymbol{\mu}_k, \Sigma_k), \text{with probability } p_k \end{cases}$$

$\sum_{j=1}^k p_j = 1$

Let $\mathbf{r}_i = (r_{i,1}, ..., r_{i,k}) \in \mathbb{R}^k$ as the cluster indicator of $\mathbf{x}_i$, which takes form $(0, 0, ..., 0, 1, 0, 0)$ with $r_{i,j} = I\{\mathbf{x}_i \text{ belongs to cluster } j\}$. The cluster indicator $\mathbf{r}_i$ is a latent variable that cannot be observed. What is complete likelihood of $(\mathbf{x}_i, \mathbf{r}_i)$.

The distribution of $\mathbf{r}_i$ is

$$f(\mathbf{r}_i) = \prod_{j=1}^k p_j^{r_i, j}$$

The complete log-likelihood is

$$\ell(\theta; \mathbf{x}, \mathbf{r}) = \sum_{i=1}^n \sum_{j=1}^k r_{i,j}[\log p_i + \log f(\mathbf{x}_i; \boldsymbol{\mu}_j, \Sigma_j)] = \sum_{i=1}^n \sum_{j=1}^k r_{i,j}[\log p_i - 1/2 \log|\Sigma| - 1/2(\mathbf{x}_i - \boldsymbol{\mu}_j)^\top \Sigma(\mathbf{x}_i - \boldsymbol{\mu}_j)]$$

E-step Evaluate the responsibilities using the current parameter values

$$\gamma_{i,k}^{(t)} = P(r_{i,k} = 1|\mathbf{x}_i, \theta^{(t)}) = \frac{p_k^{(t)} f(\mathbf{x}_i|\boldsymbol{\mu}_k^{(t)}, \Sigma_k^{(t)})}{\sum_{j=1}^{K} f(\mathbf{x}_i|\boldsymbol{\mu}_j^{(t)}, \Sigma_j^{(t)})}$$

M-step

$\theta^{(t+1)} = \arg\max \ell(\mathbf{x}, \gamma^{(t)}, \theta)$.

Let $n_k = \sum_{i=1}^{n} \gamma_{i,k}$, we have

$$\boldsymbol{\mu}_k^{(t+1)} = \frac{1}{n_k} \sum_{i=1}^{n} \gamma_{i,k} \mathbf{x}_i$$

$$\Sigma_k^{(t+1)} = \frac{1}{n_k} \sum_{i=1}^{n} \gamma_{i,k} (\mathbf{x}_i - \boldsymbol{\mu}_k^{(t+1)})(\mathbf{x}_i - \boldsymbol{\mu}_k^{(t+1)})^T$$

$$p_k^{(t+1)} = \frac{n_k}{n}$$

K-mean

The K-means algorithm partitions data into k clusters (k is predetermined). We denote $\{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, ..., \boldsymbol{\mu}_k\}$ as the centers of the k (unknown) clusters, and denote $\mathbf{r}_i = (r_{i,1}, ..., r_{i,k}) \in \mathbb{R}^k$ as the "hard" cluster assignment of $\mathbf{x}_i$.

k-means finds cluster centers and cluster assignments that minimize the objective function

$$J(\mathbf{r}, \boldsymbol{\mu}) = \sum_{i=1}^{n} \sum_{j=1}^{k} r_{i,j} \|\mathbf{x}_i - \mu_k\|^2$$

K-means is a special case for Gussian Mixture.

Result

Discussion

Task 1:

Task 2

Conclusions

Figures

References

1 Miin-ShenYang, Chien-YoLai, Chih-YingLin. ”A robust EM clustering algorithm for Gaussian mixture models.” Pattern Recognition (2012).

2 Friedman, Jerome, et al. ”Pathwise coordinate optimization.” The annals of applied statistics 1.2 (2007): 302-332.