?? Mixture Model

1 Introduction

hard clustering: several overly strong assumptions in k-means

soft clustering: large variance induced by way too many non-zeros in the probability simplex.

variance-bias tradeoff? reducing variability on the weight

2 ?? Mixture Model

2.1 General Framework

On a separable metrics space $\{\mathcal{X}, \mathcal{F}\}$, we assume each data y_i has the following probability density as the likelihood:

$$
\begin{aligned}
&\pi(y_i) \sim G(\theta_i) \\
&\pi(\theta_i) = \sum_{k=1}^{\kappa} w_{k,i} \delta_{\theta_k}(\theta_i) \\
&w_{k,i} = \frac{u_{k,i} v_k}{\sum_{k=1}^{\kappa} u_{k,i} v_k} \\
&u_{k,i} \in \{0, 1\}
\end{aligned}
\tag{1}
$$

where $\delta_{\theta_k}(y_i)$ is a probabilty density corresponds to a component distribution with parameter θ_k, and $w_{k,i}$ is the component weight that varies by index i and $\sum_{k=1}^{\kappa} w_{k,i} = 1$. The weight varying is due to the randomness in $u_{k,i}$. Due to the 0's $u_{k,i}$ introduces, this induces sparsity in the membership probabilty $w_{k,i}$.

2.2 Estimation

We now use the latent variable $\{z_{1,i}, z_{2,i}, \dots z_{\kappa,i}\} \sim MultiNomial(\{w_{1,i}, w_{2,i}, \dots w_{\kappa,i}\})$ that takes value of $\{0, \dots, 1, \dots 0\}$ that assign randomly one 1 to one of κ vertices in the simplex. The likelihood can be rewritten as:

$$[y_i] = \sum_{k=1}^{\kappa} \int z_{k,i} \delta_{\theta_k}(y_i) P(dz_{k,i}) = \int \sum_{k=1}^{\kappa} z_{k,i} \delta_{\theta_k}(y_i) P(dz_{k,i}) \tag{2}$$

where $P(dz_{k,i})$ is the measure of the multinomial distribution aforementioned, the last equation is due to Fubini theorem. Note due to the $z_{k,i}$ takes only one 1 out of K, this allows a simple log-density augmented with $z_{k,i}$:

$$\log[\{y_i\}_i, \{z_{k,i}\}_{k,i}] = \sum_i \sum_{k=1}^{\kappa} z_{k,i} \log\{\delta_{\theta_k}(y_i) v_k\} \tag{3}$$

This allows us to utilize Expectation-Maximization algorithm for parameter estimation:

Algorithm 1 EM algorithm

1: **while** $||\theta^{(s)} - \theta^{(s+1)}|| > \epsilon$ **do**
2: Maximize over $\{u_{i,k}\}_k$ by $\underset{\{u_{i,k}\}_k}{\text{argmax}} \{u_{i,k} v_k \delta_{\theta_k}(y_i)\}$ for all i; ▷ pick top J out K based on (1)
3: Compute $\mathbb{E}(z_{k,i}) = \frac{w_{k,i}\delta_{\theta_k}(y_i)}{\sum_k w_{k,i}\delta_{\theta_k}(y_i)}$ with $w_{k,i} = u_{i,k} v_k$; ▷ E step based on (3)
4: Set $\theta_k^{(s+1)} = \text{argmax} \sum_{k=1}^{\kappa} \mathbb{E}z_{k,i} \log \delta_{\theta_k}(y_i)$ for all k ▷ M step: obtain MLE based on (3)
5: Set $v_k^{(s+1)} = \sum_i \mathbb{E}z_{k,i} / \sum_i \sum_k \mathbb{E}z_{k,i}$ ▷ M step: obtain MLE based on (3)

3 Theory

variance reduction in the weight parameter

1. k-means is asymptotic biased 2. large k reducing variance 3. small sample theory

4 Simulation

large K with small J in 1d

high p setting

5 Application

6 Discussion

References