A latent class model for multivariate categorical data

Model

Let $X = \left(X^{(1)}, \ldots, X^{(m)}\right)$ be a vector of categorical variables, where

$$X^{(r)} \in \mathcal{X}_r = \{1, 2, \ldots, C_r\}.$$

The joint space $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_m$ is finite with total number of categories $C = \prod_{r=1}^{m} C_r$.

When m is large, it becomes impossible to learn the distribution of X since the total number of categories explodes exponentially with m. In order to reduce parameters and discover possible patterns, we introduce a latent class variable

$$H \in [K] := \{1, \ldots, K\},$$

and assume that condition on H, the components of X are mutually independent, i.e.,

$$P(X = x) = \sum_{k=1}^{K} P(H = k) \prod_{r=1}^{m} P\left(X^{(r)} = x^{(r)} \mid H = k\right), \quad \forall x \in \mathcal{X}.$$

To see why this model significantly reduces the number of parameters, let's say $C_r = 2$ for all r 's. Then the original model has $2^m - 1$ parameters to estimate, while the latent class model only has $K - 1 + mK$ parameters.

Estimation

Define mixture weights as

$$\pi_k = P(H = k), \quad k = 1, \ldots, K,$$

and component categorical probabilities as

$$\theta_{rkc} = P\left(X^{(r)} = c | H = k\right), \quad k = 1, \ldots, K, \quad r = 1, \ldots, m, \quad c = 1, \ldots, C_r.$$

Let the complete parameter set be

$$\Theta = \{\pi_k, \theta_{rkc} : k = 1, \ldots, K, \quad r = 1, \ldots, m, \quad c = 1, \ldots, C_r\}$$

Given n i.i.d. observations $\{X_i\}_{i=1}^{n}$, the likelihood for the latent class model is given by

$$\mathcal{L}\left(\Theta \mid \{X_i\}_{i=1}^{n}\right) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k \prod_{r=1}^{m} \theta_{rkX_i^{(r)}}.$$

We use the classical EM algorithm to estimate the model parameters.

The E-step computes the posterior probability of latent class k for each sample i:

$$\begin{aligned}\gamma_{ik} &:= P\left(H_i = k | X_i, \Theta^{\text{old}}\right) \\ &= \frac{\pi_k^{\text{old}} \prod_{r=1}^{m} \theta^{\text{old}}_{rkX_i^{(r)}}}{\sum_{j=1}^{K} \pi_j^{\text{old}} \prod_{r=1}^{m} \theta^{\text{old}}_{rjX_i^{(r)}}}.\end{aligned}$$

The M-step maximizes the expected complete log-likelihood via:

$$\begin{aligned}\pi_k^{\text{new}} &= \frac{1}{n} \sum_{i=1}^{n} \gamma_{ik}, \\ \theta_{rkc}^{\text{new}} &= \frac{\sum_{i=1}^{n} \gamma_{ik} \mathbf{1}\left(X_i^{(r)} = c\right)}{\sum_{i=1}^{n} \gamma_{ik}}.\end{aligned}$$

Note that the expression $\gamma_{ik} = \frac{\pi_k^{\text{old}} \prod_{r=1}^{m} \theta_{rkX_i^{(r)}}}{\sum_{j=1}^{K} \pi_j^{\text{old}} \prod_{r=1}^{m} \theta_{rjX_i^{(r)}}}$ is numerically bad because the products can underflow. So we move everything to the log-space.

Define

$$a_{ik} = \log \pi_k + \sum_{r=1}^{m} \log\left(\theta_{rkX_i^{(r)}}\right).$$

Then

$$\begin{aligned}\gamma_{ik} &= \frac{\exp\left(a_{ik}\right)}{\sum_{j=1}^{K} \exp\left(a_{ij}\right)} \\ &= \frac{\exp\left(a_{ik} - M_i\right)}{\sum_{j=1}^{K} \exp\left(a_{ij} - M_i\right)},\end{aligned}$$

where $M_i = \max_j a_{ij}$.

Since each $a_{ij} - M_i \leq 0$, the exponentials don't blow up.

The latent class model has an identifiability issue due to label switching. Any permutation of the latent class labels produces an equivalent model with the same likelihood. This means that:

- Classes $\{0, 1, 2\}$ with weights $\{0.5, 0.3, 0.2\}$ is equivalent to classes $\{2, 0, 1\}$ with weights $\{0.2, 0.5, 0.3\}$.

To this end, we impose an ordering constraint on the mixture weights:

$$\pi_1 \geq \ldots \geq \pi_K.$$

Classification

After fitting the model, we may predict the label of each sample i by

$$\hat{H}_i = \arg\max_{k \in [K]} \gamma_{ik}^{(\text{final})},$$

where

$$\begin{aligned}\gamma_{ik}^{(\text{final})} &= P\left(H_i = k | X_i, \Theta^{(\text{final})}\right) \\ &= \frac{\pi_k^{(\text{final})} \prod_{r=1}^{m} \theta^{(\text{final})}_{rkX_i^{(r)}}}{\sum_{j=1}^{K} \pi_j^{(\text{final})} \prod_{r=1}^{m} \theta^{(\text{final})}_{rjX_i^{(r)}}}.\end{aligned}$$

Again, this expression should be evaluated in the log-space in order to avoid potential numerical issue.

Model selection (choosing K)

If the number of latent classes K is unknown, we use Bayesian information criterion (BIC) to estimate it, i.e.,

$$\hat{K}^{\text{BIC}} = \arg\min_{K} \left\{ \left[(K-1) + \sum_{r=1}^{m} K(C_r - 1) \right] \cdot \log n - 2 \log\left(\hat{\mathcal{L}}_K\right) \right\},$$

where

- $(K-1) + \sum_{r=1}^{m} K(C_r - 1)$ is the total number of parameters,
- n is the sample size,
- $\hat{\mathcal{L}}_K$ is the maximized value of the likelihood function of the model when the number of latent classes is K.

It is generally believed that BIC estimators are consistent under mild conditions. We evaluate the performance of $\hat{K}^{\text{BIC}}$ in our simulation studies.