

STA 360/602L: MODULE 8.2

FINITE MIXTURE MODELS: UNIVARIATE CATEGORICAL DATA

DR. OLANREWAJU MICHAEL AKANDE

MULTINOMIAL MODEL RECAP

- Suppose $y_1, \dots, y_n | \boldsymbol{\theta} \stackrel{iid}{\sim} \text{Categorical}(\boldsymbol{\theta})$, then

$$\Pr[y_i = d | \boldsymbol{\theta}] = \prod_{d=1}^D \theta_d^{1[y_i=d]},$$

- With prior $\pi[\boldsymbol{\theta}] = \text{Dirichlet}(\boldsymbol{\alpha})$, we have

$$\pi[\boldsymbol{\theta}] \propto \prod_{d=1}^D \theta_d^{\alpha_d - 1}, \quad \alpha_d > 0 \text{ for all } d = 1, \dots, D.$$

- So that the posterior is

$$\pi(\boldsymbol{\theta} | Y) = \text{Dirichlet}(\alpha_1 + n_1, \dots, \alpha_D + n_D)$$

- However, what if our data actually comes from K different sub-populations of groups of people?
- For example, if our data comes from men and women, and we don't expect marginal independence across the two groups (vote turnout, income, etc), then we have a mixture of distributions.

FINITE MIXTURE OF MULTINOMIALS

- With our data coming from a "combination" or "mixture" of sub-populations, we no longer have independence across all observations, so that the likelihood $p[Y|\boldsymbol{\theta}] \neq \prod_{i=1}^n \prod_{d=1}^D \theta_j^{1[y_i=d]}$.
- However, we can still have "conditional independence" within each group.
- Unfortunately, we do not always know the indexes for those groups.
- That is, we know our data contains K different groups, but we actually do not know which observations belong to which groups.
- **Solution:** introduce a **latent variable** z_i representing the group/cluster indicator for each observation i , so that each $z_i \in \{1, \dots, K\}$.
- This is a form of **data augmentation**, but we will define that properly later.

FINITE MIXTURE OF MULTINOMIALS

- Given the cluster indicator z_i for observation i , write

- $\Pr(y_i = d | z_i) = \psi_{z_i, d} \equiv \prod_{d=1}^D \psi_{z_i, d}^{1[y_i=d|z_i]}$, and

- $\Pr(z_i = k) = \lambda_k \equiv \prod_{k=1}^K \lambda_k^{1[z_i=k]}$.

- Then, the marginal probabilities we care about will be

$$\begin{aligned}\theta_d &= \Pr(y_i = d) \\ &= \sum_{k=1}^K \Pr(y_i = d | z_i = k) \cdot \Pr(z_i = k) \\ &= \sum_{k=1}^K \lambda_k \cdot \psi_{k, d},\end{aligned}$$

which is a **finite mixture of multinomials**, with the weights given by λ_k .

POSTERIOR INFERENCE

- Write
 - $\lambda = (\lambda_1, \dots, \lambda_K)$, and
 - $\psi = \{\psi_{z_i, d}\}$ to be a $K \times D$ matrix of probabilities, where each k th row is the vector of probabilities for cluster k .
- The observed data likelihood is

$$\begin{aligned} p[Y = (y_1, \dots, y_n) | Z = (z_1, \dots, z_n), \psi, \lambda] &= \prod_{i=1}^n \prod_{d=1}^D \Pr(y_i = d | z_i, \psi_{z_i, d}) \\ &= \prod_{i=1}^n \prod_{d=1}^D \psi_{z_i, d}^{1[y_i=d|z_i]}, \end{aligned}$$

which includes products (and not the sums in the mixture pdf), and as you will see, makes sampling a bit easier.

- Next we need priors.

POSTERIOR INFERENCE

- First, for $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_K)$, the vector of cluster probabilities, we can use a Dirichlet prior. That is,

$$\pi[\boldsymbol{\lambda}] = \text{Dirichlet}(\alpha_1, \dots, \alpha_K) \propto \prod_{k=1}^K \lambda_k^{\alpha_k - 1}.$$

- For $\boldsymbol{\psi}$, we can assume independent Dirichlet priors for each cluster vector $\boldsymbol{\psi}_k = (\psi_{k,1}, \dots, \psi_{k,D})$. That is, for each $k = 1, \dots, K$,

$$\pi[\boldsymbol{\psi}_k] = \text{Dirichlet}(a_1, \dots, a_d) \propto \prod_{d=1}^D \psi_{k,d}^{a_d - 1}.$$

- Finally, from our distribution on the z_i 's, we have

$$p[Z = (z_1, \dots, z_n) | \boldsymbol{\lambda}] = \prod_{i=1}^n \prod_{k=1}^K \lambda_k^{1[z_i=k]}.$$

POSTERIOR INFERENCE

- Note that the unobserved variables and parameters are $Z = (z_1, \dots, z_n)$, ψ , and λ .
- So, the joint posterior is

$$\begin{aligned}\pi(Z, \psi, \lambda | Y) &\propto p[Y | Z, \psi, \lambda] \cdot p(Z | \psi, \lambda) \cdot \pi(\psi, \lambda) \\ &\propto \left[\prod_{i=1}^n \prod_{d=1}^D p(y_i = d | z_i, \psi_{z_i, d}) \right] \cdot p(Z | \lambda) \cdot \pi(\psi) \cdot \pi(\lambda) \\ &\propto \left(\prod_{i=1}^n \prod_{d=1}^D \psi_{z_i, d}^{1_{[y_i = d | z_i]}} \right) \\ &\quad \times \left(\prod_{i=1}^n \prod_{k=1}^K \lambda_k^{1_{[z_i = k]}} \right) \\ &\quad \times \left(\prod_{k=1}^K \prod_{d=1}^D \psi_{k, d}^{a_d - 1} \right) \\ &\quad \times \left(\prod_{k=1}^K \lambda_k^{\alpha_k - 1} \right).\end{aligned}$$

POSTERIOR INFERENCE

- First, we need to sample the z_i 's, one at a time, from their full conditionals.
- For $i = 1, \dots, n$, sample $z_i \in \{1, \dots, K\}$ from a categorical distribution (multinomial distribution with sample size one) with probabilities

$$\begin{aligned}\Pr[z_i = k | \dots] &= \Pr[z_i = k | y_i, \psi_k, \lambda_k] \\ &= \frac{\Pr[y_i, z_i = k | \psi_k, \lambda_k]}{\sum_{l=1}^K \Pr[y_i, z_i = l | \psi_l, \lambda_l]} \\ &= \frac{\Pr[y_i | z_i = k, \psi_k] \cdot \Pr[z_i = k, \lambda_k]}{\sum_{l=1}^K \Pr[y_i | z_i = l, \psi_l] \cdot \Pr[z_i = l, \lambda_l]} \\ &= \frac{\psi_{k,d} \cdot \lambda_k}{\sum_{l=1}^K \psi_{l,d} \cdot \lambda_l}.\end{aligned}$$

POSTERIOR INFERENCE

- Next, sample each cluster vector $\boldsymbol{\psi}_k = (\psi_{k,1}, \dots, \psi_{k,D})$ from

$$\pi[\boldsymbol{\psi}_k | \dots] \propto \pi(Z, \boldsymbol{\psi}, \boldsymbol{\lambda} | Y)$$

$$\propto \left(\prod_{i=1}^n \prod_{d=1}^D \psi_{z_i,d}^{1[y_i=d|z_i]} \right) \cdot \left(\prod_{i=1}^n \prod_{k=1}^K \lambda_k^{1[z_i=k]} \right) \cdot \left(\prod_{k=1}^K \prod_{d=1}^D \psi_{k,d}^{a_d-1} \right) \cdot \left(\prod_{k=1}^K \lambda_k^{\alpha_k-1} \right)$$

$$\propto \left(\prod_{d=1}^D \psi_{k,d}^{n_{k,d}} \right) \cdot \left(\prod_{d=1}^D \psi_{k,d}^{a_d-1} \right)$$

$$= \left(\prod_{d=1}^D \psi_{k,d}^{a_d+n_{k,d}-1} \right)$$

$$\equiv \text{Dirichlet}(a_1 + n_{k,1}, \dots, a_D + n_{k,D}).$$

where $n_{k,d} = \sum_{i:z_i=k} 1[y_i = d]$, the number of individuals in cluster k that are assigned to category d of the levels of y .

POSTERIOR INFERENCE

- Finally, sample $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_K)$, the vector of cluster probabilities from

$$\begin{aligned}\pi[\boldsymbol{\lambda} | \dots] &\propto \pi(Z, \boldsymbol{\psi}, \boldsymbol{\lambda} | Y) \\ &\propto \left(\prod_{i=1}^n \prod_{d=1}^D \psi_{z_i, d}^{1[y_i=d|z_i]} \right) \cdot \left(\prod_{i=1}^n \prod_{k=1}^K \lambda_k^{1[z_i=k]} \right) \cdot \left(\prod_{k=1}^K \prod_{d=1}^D \psi_{k, d}^{\alpha_d-1} \right) \cdot \left(\prod_{k=1}^K \lambda_k^{\alpha_k-1} \right) \\ &\propto \left(\prod_{i=1}^n \prod_{k=1}^K \lambda_k^{1[z_i=k]} \right) \cdot \left(\prod_{k=1}^K \lambda_k^{\alpha_k-1} \right) \\ &\propto \left(\prod_{k=1}^K \lambda_k^{n_k} \right) \cdot \left(\prod_{k=1}^K \lambda_k^{\alpha_k-1} \right) \\ &\propto \left(\prod_{k=1}^K \lambda_k^{\alpha_k+n_k-1} \right) \\ &\equiv \text{Dirichlet}(\alpha_1 + n_1, \dots, \alpha_K + n_K),\end{aligned}$$

where $n_k = \sum_{i=1}^n 1[z_i = k]$, the number of individuals assigned to cluster k .

WHAT'S NEXT?

MOVE ON TO THE READINGS FOR THE NEXT MODULE!