STA 360/602L: Module 8.2

Finite mixture models: univariate categorical data

Dr. Olanrewaju Michael Akande

Multinomial model recap

- Suppose $y_i, \ldots, y_n | \boldsymbol{\theta} \overset{iid}{\sim} \text{Categorical}(\boldsymbol{\theta})$, 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

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

- So that the posterior is

$$\pi(\boldsymbol{\theta} | Y) = \text{Dirichlet}(\alpha_1 + n_1, \ldots, \alpha_d + n_d)$$

- 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 $z_i \in \{1, \ldots, K\}$.

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

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

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

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

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

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

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

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

Posterior inference

- Note that the unobserved variables and parameters are $Z = (z_1, \ldots, z_n)$, $\boldsymbol{\psi}$, and $\boldsymbol{\lambda}$.
- So, the joint posterior is

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

Posterior inference

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

$$
\begin{aligned}
\Pr[z_i = k | \ldots] &= \Pr[z_i = k | y_i, \boldsymbol{\psi}_k, \lambda_k] \\
&= \frac{\Pr[y_i, z_i = k | \boldsymbol{\psi}_k, \lambda_k]}{\sum_{l=1}^{K} \Pr[y_i, z_i = l | \boldsymbol{\psi}_l, \lambda_l]} \\
&= \frac{\Pr[y_i | z_i = k, \boldsymbol{\psi}_k] \cdot \Pr[z_i = k, \lambda_k]}{\sum_{l=1}^{K} \Pr[y_i | z_i = l, \boldsymbol{\psi}_l] \cdot \Pr[z_i = l, \lambda_l]} \\
&= \frac{\psi_{k,d} \cdot \lambda_k}{\sum_{l=1}^{K} \psi_{l,d} \cdot \lambda_l}.
\end{aligned}
$$

Posterior inference

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

$$
\begin{aligned}
\pi[\boldsymbol{\psi}_k | \ldots] &\propto \pi\left(Z, \boldsymbol{\psi}, \boldsymbol{\lambda} | Y\right) \\
&\propto \left(\prod_{i=1}^{n} \prod_{d=1}^{D} \psi_{z_i,d}^{1[y_i = d | z_i]} \right) \cdot \left(\prod_{i=1}^{n} \prod_{k=1}^{K} \lambda_k^{1[z_i = k]} \right) \cdot \left(\prod_{k=1}^{K} \prod_{d=1}^{D} \psi_{k,d}^{a_d - 1} \right) \cdot \left(\prod_{k=1}^{K} \lambda_k^{\alpha_k - 1} \right) \\
&\propto \left(\prod_{d=1}^{D} \psi_{k,d}^{n_{k,d}} \right) \cdot \left(\prod_{d=1}^{D} \psi_{k,d}^{a_d - 1} \right) \\
&= \left(\prod_{d=1}^{D} \psi_{k,d}^{a_d + n_{k,d} - 1} \right) \\
&\equiv \text{Dirichlet}\left(a_1 + n_{k,1}, \ldots, a_d + n_{k,D}\right).
\end{aligned}
$$

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 $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_K)$, the vector of cluster probabilities from

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

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!