STA 360/602L: Module 8.6

Finite mixture models: multivariate continuous data

Dr. Olanrewaju Michael Akande

Finite mixture of univariate normal (recap)

- For a location-scale mixture of univariate normals, we can specify
  - $y_i | z_i \sim \mathcal{N}\left(\mu_{z_i}, \sigma^2_{z_i}\right)$, and
  - $\Pr(z_i = k) = \lambda_k \equiv \prod_{k=1}^{K} \lambda_k^{1[z_i = k]}$.
- Priors:
  - $\pi[\boldsymbol{\lambda}] = \text{Dirichlet}(a_1, \ldots, a_K)$,
  - $\mu_k \sim \mathcal{N}(\mu_0, \gamma_0^2)$, for each $k = 1, \ldots, K$, and
  - $\sigma_k^2 \sim \mathcal{IG}\left(\frac{\nu_0}{2}, \frac{\nu_0 \sigma_0^2}{2}\right)$, for each $k = 1, \ldots, K$.

Finite mixture of multivariate normals

- It is relatively easy to extend this to the multivariate case.
- As with the univariate case, given a sufficiently large number of mixture components, a scale-location multivariate normal mixture model can be used to approximate any multivariate density.
- We have

$$\mathbf{y}_i \overset{iid}{\sim} \sum_{k=1}^{K} \lambda_k \cdot \mathcal{N}_p(\boldsymbol{\mu}_k, \Sigma_k)$$

- Or equivalently,

$$\mathbf{y}_i | z_i, \boldsymbol{\mu}_{z_i}, \Sigma_{z_i} \sim \mathcal{N}_p(\boldsymbol{\mu}_{z_i}, \Sigma_{z_i})$$

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

Posterior inference

- We can then specify priors as

$$\pi(\boldsymbol{\mu}_k) = \mathcal{N}_p\left(\boldsymbol{\mu}_0, \Lambda_0\right) \quad \text{for } k = 1, \ldots, K;$$

$$\pi(\Sigma_k) = \mathcal{IW}_p\left(\nu_0, S_0\right) \quad \text{for } k = 1, \ldots, K;$$

$$\pi[\boldsymbol{\lambda}] = \text{Dirichlet}(a_1, \ldots, a_K).$$

- We can also just use the conjugate option for $\pi(\boldsymbol{\mu}_k, \Sigma_k)$ to avoid specifying Λ_0, so that we have

$$\begin{aligned}\pi(\boldsymbol{\mu}_k, \Sigma_k) &= \pi(\boldsymbol{\mu}_k | \Sigma_k) \cdot \pi(\Sigma_k) \\ &= \mathcal{N}_p\left(\boldsymbol{\mu}_0, \frac{1}{\kappa_0}\Sigma_k\right) \cdot \mathcal{IW}_p\left(\nu_0, S_0\right) \quad \text{for } k = 1, \ldots, K;\end{aligned}$$

$$\pi[\boldsymbol{\lambda}] = \text{Dirichlet}(a_1, \ldots, a_K).$$

- Gibbs sampler for both options follow directly from what we have covered so far.

Label switching again

- To avoid label switching when fitting the model, we can constrain the order of the $\boldsymbol{\mu}_k$'s.
- Here are three of many approaches:

1. Constrain the prior on the $\boldsymbol{\mu}_k$'s to be

$$\boldsymbol{\mu}_k|\boldsymbol{\Sigma}_k \sim \mathcal{N}_p(\boldsymbol{\mu}_0, \frac{1}{\kappa_0}\Sigma_k) \quad \boldsymbol{\mu}_{k-1} < \boldsymbol{\mu}_k < \boldsymbol{\mu}_{k+1},$$

   which does not always seem reasonable.

2. Relax option 1 above to only the first component of the mean vectors

$$\boldsymbol{\mu}_k|\boldsymbol{\Sigma}_k \sim \mathcal{N}_p(\boldsymbol{\mu}_0, \frac{1}{\kappa_0}\Sigma_k) \quad \mu_{1,k-1} < \mu_{1,k} < \mu_{1,k+1}.$$

3. Try an ad-hoc fix. After sampling the $\boldsymbol{\mu}_k$'s, rearrange the labels to satisfy $\mu_{1,k-1} < \mu_{1,k} < \mu_{1,k+1}$ and reassign the labels on $\boldsymbol{\Sigma}_k$ accordingly.

DP MIXTURE OF NORMALS (TEASER)

- To avoid setting K apriori, we can extend this finite mixture of normals to a Dirichlet process (DP) mixture of normals.
- The first level of the model remains the same. That is,

$$\mathbf{y}_i | z_i, \boldsymbol{\mu}_{z_i}, \Sigma_{z_i} \sim \mathcal{N}_p(\boldsymbol{\mu}_{z_i}, \Sigma_{z_i}) \quad \text{for each } i;$$

$$\pi(\boldsymbol{\mu}_k, \Sigma_k) = \pi(\boldsymbol{\mu}_k | \Sigma_k) \cdot \pi(\Sigma_k)$$

$$= \mathcal{N}_p\left(\boldsymbol{\mu}, \frac{1}{\kappa_0}\Sigma_k\right) \cdot \mathcal{IW}_p\left(\nu_0, S_0\right) \quad \text{for each } k.$$

DP MIXTURE OF NORMALS (TEASER)

- For the prior on $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_K)$, use the following stick breaking representation of the Dirichlet process.

$$
\begin{aligned}
P(z_i = k) &= \lambda_k; \\
\lambda_k &= V_k \prod_{l<k} (1 - V_l) \text{ for } k = 1, \ldots, \infty; \\
V_k &\overset{iid}{\sim} \text{Beta}(1, \alpha); \\
\alpha &\sim \text{Gamma}(a, b).
\end{aligned}
$$

- As an approximation, use $\lambda_k = V_k \prod_{l<k} (1 - V_l)$ for $k = 1, \ldots, K^\star$ with $K^\star$ set to be as large as possible!
- This specification forces the model to only use as many components as needed, and usually, no more. Also, the Gibbs sampler is relatively straightforward.
- Other details are beyond the scope of this course, but I am happy to provide resources for those interested!

What's next?

Well.....nothing!

You made it to the end of this course.

Hope you enjoyed the course and that you have learned a lot about Bayesian inference.