STA 360/602L: Module 8.4

Finite mixture models: univariate continuous data (illustration)

Dr. Olanrewaju Michael Akande

Location mixture of normals recap

- Sampling model with latent variable:
  - $y_i | z_i \sim \mathcal{N}\left(\mu_{z_i}, \sigma^2\right)$, and
  - $\Pr(z_i = k) = \lambda_k \equiv \prod_{k=1}^{K} \lambda_k^{1[z_i = k]}$.
- Priors:
  - $\pi[\boldsymbol{\lambda}] = \text{Dirichlet}(\alpha_1, \ldots, \alpha_K)$,
  - $\mu_k \sim \mathcal{N}(\mu_0, \gamma_0^2)$, for each $k = 1, \ldots, K$, and
  - $\sigma^2 \sim \mathcal{IG}\left(\frac{\nu_0}{2}, \frac{\nu_0 \sigma_0^2}{2}\right)$.

Full conditionals recap

- For $i = 1, \ldots, n$, sample $z_i \in \{1, \ldots, K\}$ from a categorical distribution (multinomial distribution with sample size one) with probabilities

$$\Pr[z_i = k | \ldots] = \frac{\lambda_k \cdot \mathcal{N}\left(y_i; \mu_k, \sigma^2\right)}{\sum_{l=1}^{K} \lambda_l \cdot \mathcal{N}\left(y_i; \mu_l, \sigma^2\right)}.$$

- Sample $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_K)$ from

$$\pi[\boldsymbol{\lambda} | \ldots] \equiv \text{Dirichlet}\left(\alpha_1 + n_1, \ldots, \alpha_K + n_K\right),$$

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

Full conditionals recap

- Sample the mean μ_k for each cluster from

$$\pi[\mu_k|\ldots] = \mathcal{N}(\mu_{k,n}, \gamma^2_{k,n});$$
$$\gamma^2_{k,n} = \frac{1}{\frac{n_k}{\sigma^2} + \frac{1}{\gamma_0^2}}; \qquad \mu_{k,n} = \gamma^2_{k,n}\left[\frac{n_k}{\sigma^2}\bar{y}_k + \frac{1}{\gamma_0^2}\mu_0\right],$$

- Finally, sample σ^2 from

$$\pi(\sigma^2|\ldots) = \mathcal{IG}\left(\frac{\nu_n}{2}, \frac{\nu_n\sigma_n^2}{2}\right).$$
$$\nu_n = \nu_0 + n; \qquad \sigma_n^2 = \frac{1}{\nu_n}\left[\nu_0\sigma_0^2 + \sum_{i=1}^{n}(y_i - \mu_{z_i})^2\right].$$

Practical considerations

- As we will see in the illustration very soon, the sampler for this model can suffer from label switching.
- For example, suppose our groups are men and women. Then, if we run the sampler multiple times (starting from the same initial values), sometimes it will settle on females as the first group, and sometimes on females are the second group.
- Specifically, MCMC on mixture models in general can suffer from label switching.
- Fortunately, results are still valid if we interpret them correctly.
- Specifically, we should focus on quantities and estimands that are invariant to permutations of the clusters. For example, look at marginal quantities, instead of conditional ones.

Other practical considerations

- So far we have assumed that the number of clusters K is known.
- What if we don't know K?
  - Compare marginal likelihood for different choices of K and select K with best performance.
  - Can also use other metrics, such as MSE, and so on.
  - Maybe a prior on K?
  - Go Bayesian non-parametric: Dirichlet processes!

Move to the R script here.

What's next?

Move on to the readings for the next module!