SDS 383D: Modeling II **Spring 2017**

Section 5: Mixture models

Mixture models

So far, we've assumed that our data are conditionally exchangeable given their covariates. In other words, for every unique set of covariates there exists a set of parameters, conditioned on which, the data with those covariates are i.i.d. We used various distributions over functions to learn a distribution over these parameters, for all covariate settings.

A common setting was when our data was normally distributed, with mean $\beta^T x_i$ and variance σ^2. If we did not have the covariate values x_i, our data would no longer be normally distributed.

Exercise 5.1 *Download the dataset restaurants.csv. This contains profit information for restaurants, based on seating capacity and whether they are open for dinner. Run a Bayesian regression of Profit vs SeatingCapacity and a dummy for DinnerService (you can reuse code from 2.12) (I'd suggest whitening Profit, it will make later prior specification easier). Do the residuals look normal? (e.g. plot histograms, qq plots). Now, let's just look at the raw Profit data: Does it look normal?*

Let's assume we're in the situation where we don't know any of these covariate values. For now, let's ignore the continuous-valued covariate (SeatingCapacity), and try to infer the categorical covariate. Let's say we know that half our restaurants are open for dinner. We could assume that each restaurant is associated with a *latent* indicator variable Z_i, that assigns them to one of two groups, so that

$$Z_i \sim \text{Bernoulli}(\pi)$$

As in the regression setting, conditioned on the latent variable, we will assume that the observed profits are i.i.d. normal. Again, as in the basic regression setting, we will assume the variances of the two normals are the same, but the means are different, i.e.

$$X_i|Z_i = z \sim \text{Normal}(\mu_z, \sigma^2).$$

If we marginalize over these binary indicators, our observations are assumed to be distributed according to a mixture of two Gaussians:

$$X_i \sim 0.5N(\mu_1, \sigma_1^2) + 0.5(\mu_2, \sigma_2^2)$$

We can then look at the posterior distribution over each indicator variable, conditioned on the class probabilities and parameters:

$$\begin{aligned}
&\mathbf{P}(Z_i = z|X_i, \pi, \mu_1, \sigma^2) \propto P(Z_i = z|\pi)p(X_i|\mu_z, \sigma^2)\\
\text{so,} \qquad &\mathbf{P}(Z_i = 1|X_i, \pi, \mu_1, \sigma^2) \propto \pi p(X_i|\mu_1, \sigma^2)\\
&\mathbf{P}(Z_i = 0|X_i, \pi, \mu_1, \sigma^2) \propto P(Z_i = 0|\pi)p(X_0|\mu_z, \sigma^2)
\end{aligned}$$

Conditioned on the Z_i, we can update the means of the Gaussians using conjugacy.

Note that we are not guaranteed to find latent clusters that correspond to the covariate we were expecting! If there is a more parsimonious partitioning of the data, then the posterior will tend to favor that partitioning.

Exercise 5.2 *Let's assume (as is the case if our latent variables correspond to the actual DinnerService covariate) that the class proportions are roughly equal, and fix $\pi = 0.5$. Using the conditional distributions $P(Z_i|X_i, \pi, \mu_1, \mu_2, \sigma^2)$ and $p(\mu_k|\{X_i : Z_i = k\}, \theta)$, where θ are appropriate (shared) prior parameters for μ_k, implement a Gibbs sampler that samples the means and the latent indicator variables. I'd suggest using the parameters of the initial regression to pick your hyperparameters.*

Compare the clustering obtained with the "true" clustering due to the DinnerService variable.

OK, let's now assume we don't know π, and that the two classes have different values of σ^2. Let's put a Beta(α, β) prior on π, since it is conjugate to the Bernoulli distribution.

Exercise 5.3 *Let's assume we want to integrate out π. What is the conditional distribution $P(Z_i|Z_{\neg i}, X_i, \mu_1, \mu_2, \sigma_1, \sigma_2, \alpha, \beta)$, where $Z_{\neg i}$ means all the values of Z except Z_i?*

Exercise 5.4 *How about if we want to integrate out all of the continuous variables? What is the conditional distribution $P(Z_i|Z_{\neg i}, X, \theta)$, where θ is the set of all hyperparameters?*

Exercise 5.5 *Implement a Gibbs sampler for this new model where we learn the cluster proportions. You can either implement one of the variants in the previous two exercises, or the fully uncollapsed model where we sample Z, π, μ_1, μ_2, σ_1^2 and σ_2^2.*

Let's now consider the case where we have more than two classes. Here, we need to replace our Bernoulli distribution with a multinomial parametrized by some probability vector π, so that:

$$P(Z_i = k) = \pi_k$$

Exercise 5.6 *Much as the multinomial is the multivariate generalization of the binomial distribution, the Dirichlet$(\alpha_1, \ldots, \alpha_K)$ distribution, which has pdf*

$$\frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_{k=1}^{K} \pi_k^{\alpha_k},$$

is the multivariate generalization of the beta distribution. Show that the Dirichlet is conjugate to the multinomial, and derive the posterior predictive distribution

$$P(Z_{n+1}|Z_{1:n}) = \int_{\mathcal{M}} P(Z_{n+1}|\pi)p(\pi)d\pi$$

You may find it helpful to note that, if $\pi \sim$ Dirichlet$(\alpha_1, \ldots, \alpha_K)$, then $E[\pi] = \frac{(\alpha_1, \ldots, \alpha_K)}{\sum_k \alpha_k}$.

Exercise 5.7 *Modify your previous Gibbs sampler to allow multiple classes, and two-dimensional data. Generate some data according to a Dirichlet mixture of 5 Gaussians in $\mathbb{R}^2$, and test your code on it.*