Notes on Gaussian Mixtures

Bradley Gram-Hansen

September 27, 2017

1 Mixture models

The learning outcomes are as follows:

1. Learn what sort of data mixture models should be used to model
2. Perform posterior inference in a mixture model

Mixture models are models in which we want to, I suppose, learn the *label* of our particular datum. Or, in another way, we aim to associate that datum with a number of other datum which our model learns to have the same characteristics and hence find the distribution over the whole data, that characterizes this.

The latent variables x in a mixture model correspond to a mixture component. Where the mixture component takes values in a discrete set $\{1, \ldots, K\}$. K need not be fixed. In general, a mixture model assumes data are generated by the following process: first we sample x and then we sample the observables $\mathbf{y}$ from a distribution that depends on the latent variables i.e $p(x, \mathbf{y}) = p(x)p(\mathbf{y}|x)$. In mixture models $p(x)$ is always a multinomial distribution. $p(\mathbf{y}|x)$ can take a variety of forms. In particular, it takes a Gaussian form in a 'Gaussian mixture model'.

Mathematically we can write this as:

$$p(\mathbf{y}) = \sum_{\mathbf{x}} p(\mathbf{x})p(\mathbf{y}|\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{y}|\mu_k, \Sigma_k) \tag{1}$$

our latent parameters $\mathbf{x}$ in general will be a member of $\mathbf{x} \in \mathbb{Z}/2\mathbb{Z}$ and so we say $\mathbf{x}$ has a 1-of-K representation. In which one element of the latent variables is equal to 1 and all other elements are equal to 0. This means that $x_n \in \{0, 1\}$ and $\sum_k^K x_k = 1$. This means that marginal distribution over $p(\mathbf{x})$ is specifed in terms of the mixing coefficients π_k such that $p(x_k = 1) = \pi_k$. Where π_k is the *Multinomial* distribution, which is also called the *Categorical* distribution. Elements of the Categorical distribution must satisfy the following constraints:

$$0 \leq \{\pi_k\} \leq 1 \tag{2}$$

$$\sum_{K}^{k=1} \pi_k = 1 \tag{3}$$

Because we use the 1-of-K representation we may write the marginal distribution of the latent parameters as:

$$p(\mathbf{x}) = \Pi_{k=1}^{K} \pi_k^{x_k} \tag{4}$$

Likewise, the conditional distribution of $\mathbf{y}$ given a particular value of $\mathbf{x}$ is a Gaussian given as: $p(\mathbf{y}|\mathbf{x}) = \Pi_{k=1}^{K} \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)^{x_k}$ AS $\mathbf{x}$ as $K-1$ zero elements, which means that the product of the terms would be $1 * 1 * 1$..[term where $x_m = 1$].. $* 1$... If we gave several observations points $\mathbf{y}_1, \ldots, \mathbf{y}_N$ then each observations has a corresponding latent variable $\mathbf{x}_1, \ldots \mathbf{x}_N$

1.1 Simple model

In the first model we have the following:

$$x \sim \mathbf{Cat}(0.7, 0.3) \tag{5}$$

$$y|x = 1 \sim \mathcal{N}(0, 1) \tag{6}$$

$$y|x = 2 \sim \mathcal{N}(6, 2) \tag{7}$$

therefore the marginal $p(y) = 0.7\mathcal{N}(0, 1) + 0.3 \cdot \mathcal{N}(6, 2)$

1.2 Posterior inference

Assuming we have already chosen the parameter models, we can infer which class a particular datum y is a member of via Bayes rule. That is $p(x|\mathbf{y}) \propto p(x)p(\mathbf{y}|x)$ and from example 1, that means that we have the following:

$$p(x = 1|\mathbf{y}) = \frac{p(x = 1)p(\mathbf{y}|x = 1)}{0.7\mathcal{N}(0, 1) + 0.3 \cdot \mathcal{N}(6, 2)} \tag{8}$$

1.3 Another Simple Model

Consider the following 2-D mixture of Gaussians model, where y_1 and y_2 are conditionally independent given x.

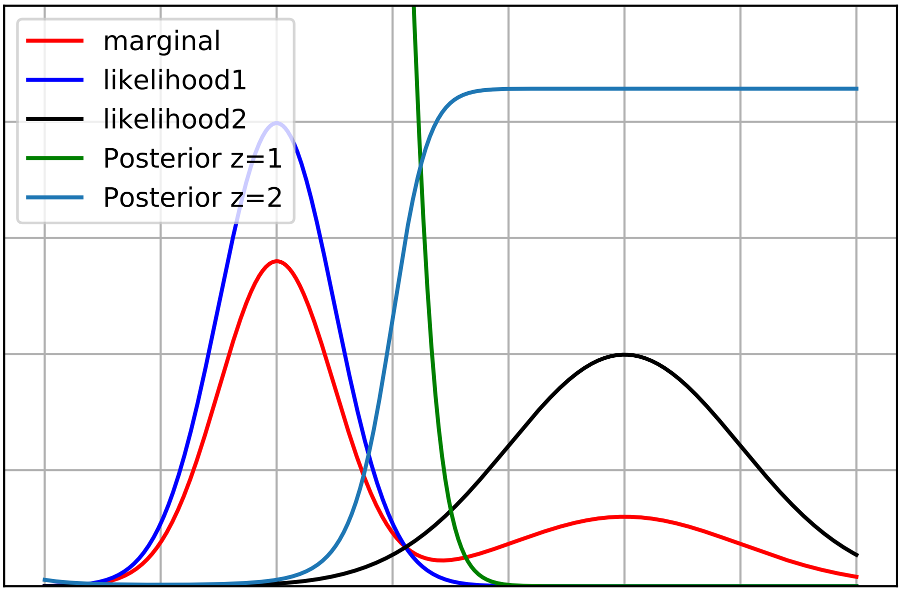
$$x \sim \mathrm{Cat}(0.4, 0.6) \tag{9}$$

$$y_1|x = 1 \sim \mathcal{N}(0, 1) \tag{10}$$

$$y_2|x = 1 \sim \mathcal{N}(6, 1) \tag{11}$$

$$y_1|x = 2 \sim \mathcal{N}(6, 2) \tag{12}$$

$$y_2|x = 2 \sim \mathcal{N}(3, 2) \tag{13}$$


marginal
likelihood1
likelihood2
Posterior z=1
Posterior z=2