Mixture Models

Robert Jacobs
Department of Brain & Cognitive Sciences
University of Rochester
Rochester, NY 14627, USA

August 8, 2008

Mixture models are often used for clustering; i.e., to summarize data that has multiple modes. This is often the case when the data is sampled from multiple sub-populations (i.e., categories), but individual data items are not labeled as to which of the sub-populations they come from.

Consider the task of summarizing the data in Figure 1. A common technique for performing this task is to use a statistical model known as a mixture model. Relative to many other models for estimating densities, mixture models have a number of advantages. First, mixture models can summarize data that contain multiple modes. In this sense, they are more powerful than distributions from the exponential family (e.g., Gaussian, binomial, Poisson, etc.). Second, mixture models are parametric models. Methods based on probability theory, such as maximum likelihood and Bayesian inference methods, are often easily applied to mixture models. Third, mixture models are parsimonious in the sense that they typically combine distributions that are simple and relatively well-understood. In the conventional statistics literature, the components of mixture models are nearly always members of the exponential family of distributions (but this has recently begun to change; we will talk about this more later in the semester).

A mixture model summarizing the data above might contain two mixture components, each a Gaussian distribution. The two Gaussians would have different mean vectors and covariance matrices. The mean of one Gaussian would roughly be the point (3, 3); the mean of the second Gaussian would be the point (7, 7). Mixture models provide a principled way of combining the two (uni-modal) Gaussian distributions into a single (multi-modal) distribution that summarizes the entire data set. As this example illustrates, mixture models are "piecewise estimators" in the sense that different components are used to summarize different subsets of the data. The subsets do not, however, have hard boundaries; as discussed below, a data item might simultaneously be a member of multiple subsets.

As a second example, suppose that we measure the height of a large number of adults. It is quite likely that the distribution of heights is bimodal. This is because the distribution of heights for males and females is different. In short, we can model the distribution of heights as the combination of a distribution for the height of males and a distribution for the height of females.

For convenience, let's restrict our attention to mixture models that are a mixture of Gaussian (Normal) distributions. We assume that the environment generates the data in the following way. For each data item $x^{(t)}$:

1. One of the Gaussian distributions is selected at random from some probability distribution. Let $\pi(i)$ denote the probability of selecting the i^{th} Gaussian;

2. The i^{th} Gaussian is sampled. This sample is the data item $x^{(t)}$. The probability that Gaussian i generates the value $x^{(t)}$ is given by (assuming a one-dimensional Gaussian distribution):

$$p(x^{(t)}|i) = \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{1}{2\sigma_i^2}(x^{(t)}-\mu_i)^2} \tag{1}$$

where μ_i is the mean of the i^{th} Gaussian and σ_i^2 is its variance.

Note that the overall probability of the value $x^{(t)}$ is given by:

$$\begin{aligned} p(x^{(t)}) &= \sum_i p(i, x^{(t)}) && (2)\\ &= \sum_i p(i)\, p(x^{(t)}|i) && (3)\\ &= \sum_i \pi(i)\, p(x^{(t)}|i) && (4)\\ &= \sum_i \pi(i) \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{1}{2\sigma_i^2}(x^{(t)}-\mu_i)^2}. && (5) \end{aligned}$$

If we regard a mixture model as an instance of a graphical model, then we get the graph shown in Figure 2. Note that to generate a data item (i.e. a set of values for the visible variables), a hidden variable is selected at random (e.g., one of the Gaussian distributions is selected), and the selected hidden variable generates the visible data (e.g., a sample is drawn from the selected Gaussian distribution; in the figure we've assumed that this is a four-dimensional Gaussian distribution). It will be very important for us to somehow try to estimate which hidden variable (e.g., which Gaussian distribution) was responsible for generating each data item. Data items generated by the same hidden variable are said to belong to the same cluster, whereas data items generated by different hidden variables are said to belong to different clusters.

We have discussed the generation of a particular data item $x^{(t)}$. In general, there will be many data items: $\mathcal{X} = \{x^{(t)}\}_{t=1}^{T}$. The data items are independent and identically distributed so the probability of getting the entire sample $\mathcal{X}$ is:

$$\begin{aligned} p(\mathcal{X}) &= \prod_t \sum_i \pi(i)\, p(x^{(t)}|i) && (6)\\ &= \prod_t \sum_i \pi(i) \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{1}{2\sigma_i^2}(x^{(t)}-\mu_i)^2}. && (7) \end{aligned}$$

The product arises from the fact that the individual data items are generated independently.

Suppose that we consider data item $x^{(t)}$ and we want to know what Gaussian distribution it came from. That is, we want to know the probabilities $p(i|x^{(t)})$. Using Bayes' rule, we get:

$$p(i|x^{(t)}) = \frac{\pi(i)\, p(x^{(t)}|i)}{p(x^{(t)})} \tag{8}$$

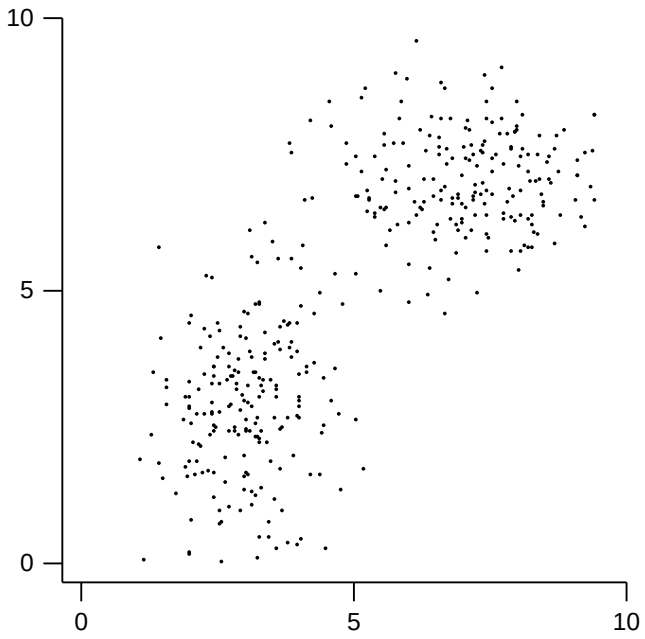

Figure 1: Data items to be summarized via a mixture model.

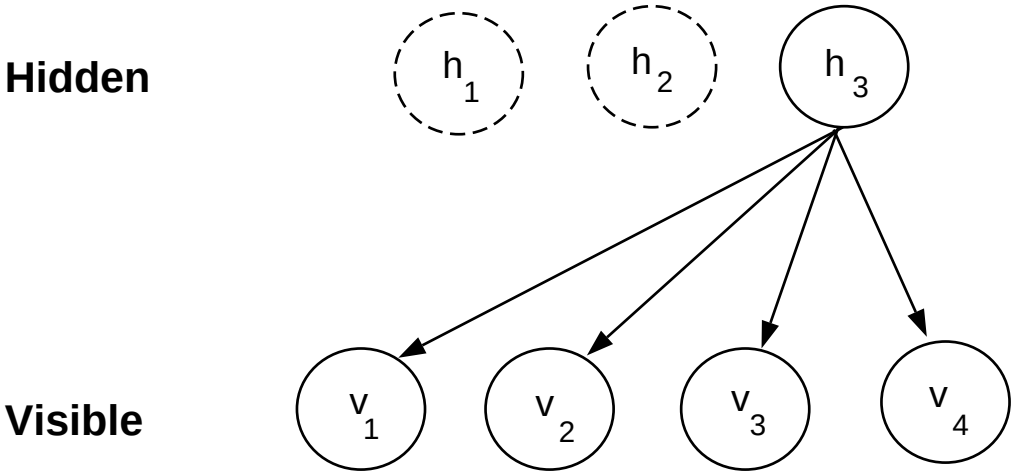

Figure 2: Graphical representation of a mixture model.

$$= \frac{\pi(i)\, p(x^{(t)}|i)}{\sum_j \pi(j)\, p(x^{(t)}|j)} \tag{9}$$

$$= \frac{\pi(i)\, \frac{1}{\sqrt{2\pi}\sigma_i}\, e^{-\frac{1}{2\sigma_i^2}(x^{(t)}-\mu_i)^2}}{\sum_j \pi(j) \frac{1}{\sqrt{2\pi}\sigma_j}\, e^{-\frac{1}{2\sigma_j^2}(x^{(t)}-\mu_j)^2}}. \tag{10}$$

Note that this is a powerful tool. We are not told what Gaussian (e.g., category) a particular data item was sampled from, but we are able to determine the probabilities that it came from the different distributions. Much of the pattern classification literature is based on this set of equations. Using Bayesian terminology, we will refer to $p(i) = \pi(i)$ as the prior probability of Gaussian i, and to $p(i|x^{(t)})$ as its posterior probability.

Recall that for maximum likelihood estimation, we are interested in finding the values of the Gaussian means and variances that maximize the log likelihood of the probability of the data. That is, we want to adjust our parameters (via gradient ascent for the purposes of this note) so as to maximize the log likelihood function:

$$\log L = \log p(\mathcal{X}) = \sum_t \log \sum_i \pi(i)\, p(x^{(t)}|i). \tag{11}$$

After some simplifications, the derivatives for the means are:

$$\frac{\partial \log L}{\partial \mu_i} = \sum_t \frac{p(i|x^{(t)})}{\sigma_i^2}(x^{(t)} - \mu_i). \tag{12}$$

That is, the mean of Gaussian i is moved towards the data item $x^{(t)}$, but only in proportion to the probability that it generated $x^{(t)}$ [the posterior probability $p(i|x^{(t)})$]. This makes sense, right? A Gaussian that was very likely to have generated $x^{(t)}$ should change its parameters a lot; a Gaussian that was very unlikely to have generated $x^{(t)}$ shouldn't change its parameters much at all. The derivatives for the variances are:

$$\frac{\partial \log L}{\partial \sigma_i^2} = \sum_t \frac{p(i|x^{(t)})}{2\sigma_i^4}[(x^{(t)} - \mu_i)^2 - \sigma_i^2]. \tag{13}$$

That is, the variance σ_i^2 is moved towards the sample variance $(x^{(t)} - \mu_i)^2$, but only in proportion to the probability that Gaussian i generated $x^{(t)}$ [the posterior probability $p(i|x^{(t)})$].

Note that this framework is not limited to one-dimensional data. In multiple dimensions, each cluster is a multi-dimensional Gaussian distribution with a mean vector and a covariance matrix. It is frequently the case that the covariance matrix for the j^{th} cluster is restricted to the form $\sigma_j^2\mathbf{I}$ where $\mathbf{I}$ is the identity matrix. Based on what we've covered in class so far, you should be able to understand how to update the mean vector and the covariance matrix $\sigma_j^2\mathbf{I}$. You should also think about how to update covariance matrices of other forms (such as diagonal matrices in which each diagonal entry has a different value, non-diagonal matrices, etc.).

Also note that this whole game can be played with distributions other than Gaussian. For example, $x^{(t)}$ could be a Bernoulli variable (e.g., binary variable taking the values 0 and 1, or head and tail, or true and false, etc.). In this case, we have a mixture of Bernoulli densities. The probability of the data is:

$$p(\mathcal{X}) = \prod_t \sum_i \pi(i)\, p(x^{(t)}|i) \tag{14}$$

$$= \prod_t \sum_i \pi(i)\, [\mu_i]^{x^{(t)}}\, [1-\mu_i]^{1-x^{(t)}} \tag{15}$$

where $\pi(i)$ is the (prior) probability of selecting the i^{th} Bernoulli distribution and μ_i is the mean of the i^{th} Bernoulli distribution.

In general, one can have a mixture model where the mixture components are any probability model. This can be extremely powerful. For example, we could have a mixture of factor analyzers, a mixture of hidden Markov models, a mixture of conditional distributions (known as a mixture of experts), or many other possibilities.

In this note, we've considered estimating the parameters of a mixture model via gradient ascent. It is more common in the literature to estimate these parameters using other algorithms, such as the Expectation-Maximization (EM) algorithm. However, that is a topic for another day.

Lastly, recall that we've only considered mixture models in which the number of mixture components is fixed. There are ways of growing the number of mixture components during the course of training based on the characteristics of the training data. These models, known as Dirichlet Process Mixture Models, are beyond the scope of this note.