Mixture Models

Matthew Stephens (edited and modified by Heejung Shim)

Learning goals

- Know what generative process is assumed in a mixture model, and what sort of data it is intended to model.
- Be able to obtain MLE of the parameters in a mixture model using the Expectation-Maximization (EM) algorithm (after learning the EM algorithm)
- Be able to perform clustering inference in a mixture model. Specifically, for a given sample, be able to compute the probability of the sample belonging to each cluster.

Motivating examples

- You are a biologist interested in studying behaviour, and you have collected lots of videos of bees. You want to "cluster" their behaviours into meaningful groups, so that you can look at each cluster and figure out what it means.
- You have DNA data from 500 individuals. Could you cluster them into multiple (hypothesized) ancestral groups?

Generating data from a mixture model

mixture model
mixture component
mixture proportion

Example 1 Suppose we are interested in simulating the price of a randomly chosen book. Since paperback books are typically cheaper than hardbacks, it might make sense to model the price of paperback books separately from hardback books. In this example, we will model the price of a book as a **mixture model**. We will have two **mixture components** in our model - one for paperback books, and one for hardbacks.

Let's say that if we choose a book at random, there is a 50% chance of choosing a paperback and 50% of choosing hardback. These proportions are called **mixture proportions**. Assume the price of a paperback book is normally distributed with mean 9 and standard deviation 1 and the price of a hardback is normally distributed with a mean 20 and a standard deviation of 1.

How would you simulate the price of a randomly chosen book?

We could simulate book prices P_i in the following way:

- Sample $Z_i \sim \text{Bernoulli}(0.5)$.
- If $Z_i = 0$ draw P_i from the paperback distribution $N(9, 1)$. If $Z_i = 1$, draw P_i from the hardback distribution $N(20, 1)$.

Can you implement R code to simulate 5000 book prices? See 'Simulating the price of a randomly chosen book' in Mixture_Model.pdf.

We see that the resulting probability density for all books is bimodal, and is therefore not normally distributed. In this example, we modeled the price of a book as a **mixture** of two **components** where each component was modeled as a Gaussian distribution. This is called a **Gaussian mixture model** (GMM).

Example 2 Now assume our data are the heights of students at the University of Melbourne. Assume the height of a randomly chosen male is normally distributed with a mean equal to 175 cm and a standard deviation of 6 cm and the height of a randomly chosen female is normally distributed with a mean equal to 160 cm and a standard deviation of 6 cm. However, instead of 50/50 mixture proportions, assume that 75% of the population is female, and 25% is male.

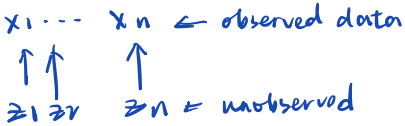
We can simulate heights in a similar fashion as above, with the corresponding changes to the parameters. See 'Simulating the heights of students at the University of Melbourne' in Mixture_Model.pdf.

We see that the Gaussian mixture model is unimodal in this example. But we can see that the population density is not symmetric, and therefore not normally distributed.

Generating data from a mixture model

These two illustrative examples above give rise to the general notion of a **mixture model** which assumes each observation is generated from one of K mixture components. We will formalize this notion.

Mixture model



Assume we observe independent samples $X_1, \ldots, X_n$ and that each X_i is sampled from one of K **mixture components**. In the second example above, the mixture components were {male,female}.

Associated with each random variable X_i is a label $Z_i \in \{1, \ldots, K\}$ which indicates which component X_i came from. In the second example, Z_i would be either 1 or 2 depending on whether X_i was a male or female height. Often times we don't observe Z_i (e.g. we might just obtain a list of heights with no gender information), so the Z_i's are sometimes called **latent variables**.

$$p(x, z) = p(x|z) \cdot p(z)$$

Mixture model - distribution of Z_i

We assume that $Z_1, \ldots, Z_n$ are independent and identically distributed as follows. For $i = 1 \ldots, n$,

$$Z_i \sim \text{categorical } (\pi_1, \ldots, \pi_K),$$

where $\pi_k > 0$ for $k = 1, \ldots, K$, and $\sum_{k=1}^{K} \pi_k = 1$.

Categorial distribution:

- See definition here :
  `https://en.wikipedia.org/wiki/Categorical_distribution`
- Informally, it is equivalent to

$$Z_i = k \text{ with probability } \pi_k \text{ for } k = 1, \ldots, K.$$

Mixture model - distribution of Z_i

We assume that $Z_1, \ldots, Z_n$ are independent and identically distributed as follows. For $i = 1 \ldots, n$,

$$Z_i \sim \text{categorical } (\pi_1, \ldots, \pi_K),$$

where $\pi_k > 0$ for $k = 1, \ldots, K$, and $\sum_{k=1}^{K} \pi_k = 1$.

Here, the π_k are called **mixture proportions** or **mixture weights** and they represent the probability that X_i belongs to the k-th mixture component.

The mixture proportions are nonnegative and they sum to one, $\sum_{k=1}^{K} \pi_k = 1$.

Mixture model - distribution of X_i conditional on Z_i

We assume that $X_1, \ldots, X_n$ conditional on $Z_1, \ldots, Z_n$ are independent and identically distributed as follows. For $i = 1 \ldots, n$,

$$X_i | Z_i = k \sim P(X_i | \Theta_k), \ k = 1, \ldots, K.$$

We call $P(X_i|\Theta_k)$ the **mixture component**, and it represents the distribution of X_i assuming it came from component k (i.e., $Z_i = k$).

The mixture components in our examples above were normal distributions, leading to

$$X_i | Z_i = k \sim N(\mu_k, \sigma_k^2), \ k = 1, \ldots, K.$$

from the 2nd example

$Z_i = \begin{cases} 1 \to \text{male} \\ 2 \to \text{female} \end{cases}$ $\quad X_i | Z_i = 1 \sim N(175, 6^2)$ $\quad X_i | Z_i = 2 \sim N(160, 6^2)$.

Mixture model

If we observe independent samples $X_1, \ldots, X_n$ from this mixture model, the probability of $X_1, \ldots, X_n$ can be written as

$$\prod_{i=1}^{n} P(X_i) = \prod_{i=1}^{n} \left[\sum_{k=1}^{K} P(X_i, Z_i = k) \right] = \prod_{i=1}^{n} \left[\sum_{k=1}^{K} P(X_i | Z_i = k) P(Z_i = k) \right]$$

$$= \prod_{i=1}^{n} \left[\sum_{k=1}^{K} P(X_i | \Theta_k) \pi_k \right].$$

For Gaussian mixture model (GMM),

$$\prod_{i=1}^{n} P(X_i) = \prod_{i=1}^{n} \left[\sum_{k=1}^{K} f(x_i; \mu_k, \sigma_k^2) \pi_k \right],$$

where $f(x_i; \mu_k, \sigma_k^2)$ is the probability density function of a normal (gaussian) distribution with mean $= \mu_k$ and variance $= \sigma_k^2$.

MLE of the parameters in a mixture model

For GMM, the likelihood function of the parameters, $\theta = (\pi_1, \ldots, \pi_{K-1}, \mu_1, \sigma_1, \ldots, \mu_K, \sigma_K)$, is

$$L(\theta) = \prod_{i=1}^{n} \left[\sum_{k=1}^{K} f(x_i; \mu_k, \sigma_k^2) \pi_k \right].$$

A natural next question to ask is how to estimate the parameters, θ, from our observations $X_1, \ldots, X_n$. We illustrate an EM algorithm to find MLE of θ in the next lecture.

Clustering inference in a mixture model

Here, we assume we've already obtained MLE of the parameters in the model. We want to infer, given a data point X_i, which component it is likely to belong to. More precisely, we want to compute $P(Z_i = k|X_i)$ for $k = 1, \ldots, K$.

$$P(Z_i = k|X_i) = \frac{P(Z_i = k, X_i)}{P(X_i)} = \frac{P(X_i|Z_i = k)P(Z_i = k)}{\sum_{k'=1}^{K} P(X_i|Z_i = k')P(Z_i = k')}$$

Example : See 'Computing the probability of each book being paperback or hardback' and 'Computing the probability of each student being female or male' in Mixture_Model.pdf.

Note on mixture model

We make one small note that sometimes confuses students new to mixture models.

You might recall that if X^a and X^b are normal random variables, then $X = \pi_a X^a + \pi_b Y^b$ is also a normally distributed random variable. From this, you might wonder why the mixture models above aren't normal. Why?

The reason is that $\pi_a X^a + \pi_b Y^b$ is **not** a mixture of normals. It is a linear combination of normals.

A random variable sampled from a simple Gaussian mixture model can be thought of as a two stage process. First, we randomly sample a component (e.g. male or female), then we sample our observation from the normal distribution corresponding to that component. This is clearly different than sampling X^a and X^b from different normal distributions, then computing a linear combination of them.

Learning goals

- Know what generative process is assumed in a mixture model, and what sort of data it is intended to model.
- Be able to obtain MLE of the parameters in a mixture model using the Expectation-Maximization (EM) algorithm (after learning the EM algorithm)
- Be able to perform clustering inference in a mixture model. Specifically, for a given sample, be able to compute the probability of the sample belonging to each cluster.