Summary

On this note we will introduce mixture models, Gaussian Mixture Models, and the Expectation-Maximization (EM) algorithm. You can find an example implementation of the EM algorithm to estimate Gaussian Mixture Models parameters in R here: https://ccvr.github.io/index.html

Mixture models - Introduction

Probability distributions are useful to model data. For example, you can model the rate of an event occurring using a Poisson distribution or you can model the average width of a given species of tree using a Gaussian distribution.

But what happens when data is more complex? What happens if your observations clearly come from more than one distribution? For example, female angler fish are much larger than males, therefore, to accurately model angler fish weight it would be appropriate to model angler fish weight as coming from 2 different Gaussian distributions (one for each sex). This is where mixture models come in.

Mixture models - Definition

We have $X_1, \ldots, X_n$ data point, where each X_i is sampled from one of K distributions (also called mixture components). Usually, we do not know which of the K distributions X_i came from, and it is useful to describe this unobserved label using another variable, Z_i

The probability of each data point X_i can be expressed as:

$$P(X_i = x) = \sum_{k=1}^{K} P(X_i = x \mid Z_i = k)P(Z_i = k) = \sum_{k=1}^{K} P(X_i = x \mid Z_i = k)\pi_k$$

Where $P(Z_i = k) = \pi_k$, and $\sum_{k=1}^{K} \pi_k = 1$

The probability mass function for the mixture model $p(x)$ can be described as:

$$p(x) = \sum_{k=1}^{K} p(x \mid Z_k)\pi_k$$

Where $p(x \mid Z_k)$ can be any probability mass function. For $X_1, \ldots, X_n$ independent observations, the likelihood function is:

$$L(\pi) = \prod_{i=1}^{n} \sum_{k=1}^{K} P(X_i = x \mid Z_i = k)\pi_k$$

Although I've described mixture models in the discrete context, the same definitions apply on the continuous setting.

Gaussian mixture models

Gaussian mixture models are a type of mixture models, where the probability density function is Gaussian. In this context, for $X_1, \ldots, X_n$ independent observations, the likelihood function is:

$$L(\pi) = \prod_{i=1}^{n} \sum_{k=1}^{K} P(X_i = x \mid Z_i = k)\pi_k = \prod_{i=1}^{n} \sum_{k=1}^{K} N(X_i = x \mid \mu_k, \sigma_k^2)\pi_k$$

Now it's time to answer the important question: How do we estimate the parameters of the underlying distributions? Remember, all this began because we do not know what the underlying distributions are, and we don't know which data point comes from which distribution. Because we don't know these two things, we won't be able to solve analytically for our distribution parameters. Instead, we can use the Expectation-Maximization algorithm.

Expectation-Maximization Algorithm

1. Initialize the μ_k, σ_k^2, π_k parameters and evaluate the log-likelihood.

2. E-step: Evaluate $P(Z_i = k \mid X_i)$ with the current set of parameters (see equations 1 below).
3. M-step: Estimate all the parameters, μ_k, σ_k^2, π_k, using the $P(Z_i = k \mid X_i)$ as weights (see equations 2-4 below).
4. Re-evaluate the likelihood. Stop if it has converged (does not change more than a small quantity ϵ). Repeat 2-4 if it has not converged.

The equations below can be derived using the standard maximum likelihood estimation framework. You can look at the section on additional resources for Gaussian Mixture models for the full derivation.

$$P(Z_i = k \mid X_i) = \frac{P(X_i \mid Z_i = k)P(Z_i = k)}{P(X_i)} = \frac{\pi_k N(\mu_k, \sigma_k^2)}{\sum_{k=1}^{K} \pi_k N(\mu_k, \sigma_k^2)} = \theta_{Z_i,k} \quad (1)$$

$$\mu_k^* = \frac{1}{N_k} \sum_{i=1}^{n} x_i \, \theta_{Z_i,k} \quad (2)$$

$$\sigma_k^{*2} = \frac{1}{N_k} \sum_{i=1}^{n} (x_i - \mu_k)^2 \, \theta_{Z_i,k} \quad (3)$$

$$\pi_k = \frac{N_k}{n} \quad (4)$$

Finally, the EM-Algorithm can be easily implemented in R, and has been implemented already in several R-Packages. My personal recommendation is the “mixtools” R-package. An example of a simpler implementation I wrote can be found here https://ccvr.github.io/ .

Other good resources for Gaussian Mixture models:
https://stephens999.github.io/fiveMinuteStats/index.html
https://en.wikipedia.org/wiki/Mixture_model