Gaussian Mixtures

Data = [6.47, 5.82, 8.7, 4.76, 7.62, 6.95, 7.44, 6.73, 3.38, 5.89, 7.81, 6.93, 7.23, 6.25, 5.31, 7.71, 7.42, 5.81, 4.03, 7.09, 7.1, 7.62, 7.74, 6.19, 7.3, 7.37, 6.99, 2.97, 3.3, 7.08, 6.23, 3.67, 3.05, 6.67, 6.5, 6.08, 3.7, 6.76, 6.56, 3.61, 7.25, 7.34, 6.27, 6.54, 5.83, 6.44, 5.34, 7.7, 4.19, 7.34]

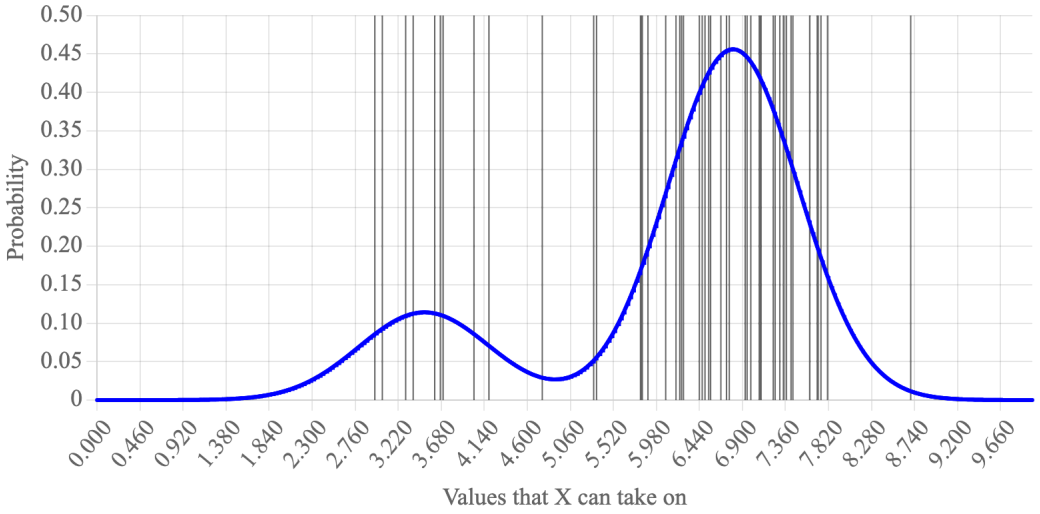
Parameter t: 0.2

Parameter μ_a: 3.5

Parameter σ_a: 0.7

Parameter μ_b: 6.8

Parameter σ_b: 0.7



Likelihood: 1.847658621579746e-34

Log Likelihood: -77.7

Best Seen: -77.7

What is a Gaussian Mixture?

A Gaussian Mixture describes a random variable whose PDF could come from one of two Gaussians (or more, but we will just use two in this demo). There is a certain probability the sample will come from the first gaussian, otherwise it comes from the second. It has five parameters: 4 to describe the two gaussians and one to describe the relative weighting of the two gaussians.

Generative Code

```
from scipy import stats
def sample():
   # choose group membership
   membership = stats.bernoulli.rvs(0.2)
   if membership == 1:
      # sample from gaussian 1
      return stats.norm.rvs(3.5,0.7)
   else:
      # sample from gaussian 2
      return stats.norm.rvs(6.8,0.7)
```

Probability Density Function

$$f(X = x) = t \cdot f(A = x) + (1 - t) \cdot f(B = x)$$

st

$$A \sim N(\mu_a, \sigma_a^2)$$

$$B \sim N(\mu_b, \sigma_b^2)$$

Putting it all together, the PDF of a Gaussian Mixture is:

$$f(x) = t \cdot \left(\frac{1}{\sqrt{2\pi}\sigma_a} e^{-\frac{1}{2}(\frac{x-\mu_a}{\sigma_a})^2}\right) + (1-t) \cdot \left(\frac{1}{\sqrt{2\pi}\sigma_b} e^{-\frac{1}{2}(\frac{x-\mu_b}{\sigma_b})^2}\right)$$

MLE for Gaussian Mixture

Special note: even though the generative story has a bernoulli (group membership) it is never observed. MLE maximizes the likelihood of the observed data.

Let $\vec{\theta} = [t, \mu_a, \mu_b, \sigma_a, \sigma_b]$ be the parameters. Because the math will get long I will use θ as notation in place of $\vec{\theta}$. Just keep in mind that it is a vector.

The MLE idea is to chose values of θ which maximize log likelihood. All optimization methods require us to calculate the partial derivatives of the thing we want to optimize (log likelihood) with respect to the values we can change (our parameters).

Likelihood function

$$\begin{aligned} L(\theta) &= \prod_i^n f(x_i|\theta) \\ &= \prod_i^n \left[t \cdot \left(\frac{1}{\sqrt{2\pi}\sigma_a} e^{-\frac{1}{2}(\frac{x_i-\mu_a}{\sigma_a})^2}\right) + (1-t) \cdot \left(\frac{1}{\sqrt{2\pi}\sigma_b} e^{-\frac{1}{2}(\frac{x_i-\mu_b}{\sigma_b})^2}\right)\right] \end{aligned}$$

Log Likelihood function

$$\begin{aligned} LL(\theta) &= \log L(\theta) \\ &= \log \prod_i^n f(x_i|\theta) \\ &= \sum_i^n \log f(x_i|\theta) \end{aligned}$$

That is sufficient for now, but if you wanted to expand out the term you would get:

$$LL(\theta) = \sum_i^n \log \left[t \cdot \left(\frac{1}{\sqrt{2\pi}\sigma_a} e^{-\frac{1}{2}(\frac{x_i-\mu_a}{\sigma_a})^2}\right) + (1-t) \cdot \left(\frac{1}{\sqrt{2\pi}\sigma_b} e^{-\frac{1}{2}(\frac{x_i-\mu_b}{\sigma_b})^2}\right)\right]$$

Derivative of LL with respect to θ

Here is an example of calculating a partial derivative with respect to one of the parameters, μ_a. You would need a derivative like this for all parameters.

Caution: When I first wrote this demo I thought it would be a simple derivative . It is not so simple because the log has a sum in it. As such the log term doesn't reduce. The log still serves to make the outer $\prod$ into a $\sum$. As such the LL partial derivatives are solvable, but the proof uses quite a lot of chain rule.

Takeaway: The main takeaway from this section (in case you want to skip the derivative proof) is that the resulting derivative is complex enough that we will want a way to compute argmax without having to set that derivative equal to zero and solving for μ_a. Enter gradient descent!

A good first step when doing a huge derivative of a log likelihood function is to think of the derivative for the log of likelihood of a single datapoint. This is the inner sum in the log likelihood expression:

$$\frac{d}{d\mu_a} \log f(x_i|\theta)$$

Before we start: notice that μ_a does not show up in this term from $f(x_i|\theta)$:

$$(1-t) \cdot \left(\frac{1}{\sqrt{2\pi}\sigma_b} e^{-\frac{1}{2}(\frac{x_i-\mu_b}{\sigma_b})^2}\right) = K$$

In the proof, when we encounter this term, we are going to think of it as a constant which we call K. Ok, lets go for it!

$$
\begin{aligned}
&\frac{d}{d\mu_a}\log f(x_i|\theta) \\
&\quad= \frac{1}{f(x_i|\theta)}\frac{d}{d\mu_a}f(x_i|\theta) && \text{chain rule on } \log \\
&\quad= \frac{1}{f(x_i|\theta)}\frac{d}{d\mu_a}\Big[t\cdot\Big(\frac{1}{\sqrt{2\pi}\sigma_a}e^{-\frac{1}{2}(\frac{x_i-\mu_a}{\sigma_a})^2}\Big)+K\Big] && \text{substitute in } f(x_i|\theta) \\
&\quad= \frac{1}{f(x_i|\theta)}\frac{d}{d\mu_a}\Big[t\cdot\Big(\frac{1}{\sqrt{2\pi}\sigma_a}e^{-\frac{1}{2}(\frac{x_i-\mu_a}{\sigma_a})^2}\Big)\Big] && \frac{d}{d\mu_a}K=0 \\
&\quad= \frac{t}{f(x_i|\theta)\sqrt{2\pi}\sigma_a}\cdot\frac{d}{d\mu_a}e^{-\frac{1}{2}(\frac{x_i-\mu_a}{\sigma_a})^2} && \text{pull out const} \\
&\quad= \frac{t}{f(x_i|\theta)\sqrt{2\pi}\sigma_a}\cdot e^{-\frac{1}{2}(\frac{x_i-\mu_a}{\sigma_a})^2}\cdot\frac{d}{d\mu_a}-\frac{1}{2}(\frac{x_i-\mu_a}{\sigma_a})^2 && \text{chain on } e^x \\
&\quad= \frac{t}{f(x_i|\theta)\sqrt{2\pi}\sigma_a}\cdot e^{-\frac{1}{2}(\frac{x_i-\mu_a}{\sigma_a})^2}\cdot\Big[-(\frac{x_i-\mu_a}{\sigma_a})\frac{d}{d\mu_a}(\frac{x_i-\mu_a}{\sigma_a})\Big] && \text{chain on } x^2 \\
&\quad= \frac{t}{f(x_i|\theta)\sqrt{2\pi}\sigma_a}\cdot e^{-\frac{1}{2}(\frac{x_i-\mu_a}{\sigma_a})^2}\cdot\Big[-(\frac{x_i-\mu_a}{\sigma_a})\cdot\frac{-1}{\sigma_a}\Big] && \text{final derivative} \\
&\quad= \frac{t}{f(x_i|\theta)\sqrt{2\pi}\sigma_a^3}\cdot e^{-\frac{1}{2}(\frac{x_i-\mu_a}{\sigma_a})^2}\cdot(x_i-\mu_a) && \text{simplify}
\end{aligned}
$$

That was for a single data-point. For the full dataset:

$$
\begin{aligned}
\frac{dLL(\theta)}{d\mu_a} &= \sum_i^n \frac{d}{d\mu_a}\log f(x_i|\theta) \\
&= \sum_i^n \frac{t}{f(x_i|\theta)\sqrt{2\pi}\sigma_a^3}\cdot e^{-\frac{1}{2}(\frac{x_i-\mu_a}{\sigma_a})^2}\cdot(x_i-\mu_a)
\end{aligned}
$$

This process should be repeated for all five parameters! Now, how should we find a value of μ_a, which, in the presence of the other settings to parameters, and the data, makes this derivative zero? Setting the derivative = 0 and solving for μ_a is not going to work.

Use an Optimizer to Estimate Params

Once we have a LL function and the derivative of LL with respect to each parameter we are ready to compute argmax using an optimizer. In this case the best choice would probably be gradient ascent (or gradient descent with negative log likelihood).

$$
\nabla_\theta LL(\theta) = \begin{bmatrix}
\frac{dLL(\theta)}{dt} \\
\frac{dLL(\theta)}{d\mu_a} \\
\frac{dLL(\theta)}{d\mu_b} \\
\frac{dLL(\theta)}{d\sigma_a} \\
\frac{dLL(\theta)}{d\sigma_b}
\end{bmatrix}
$$