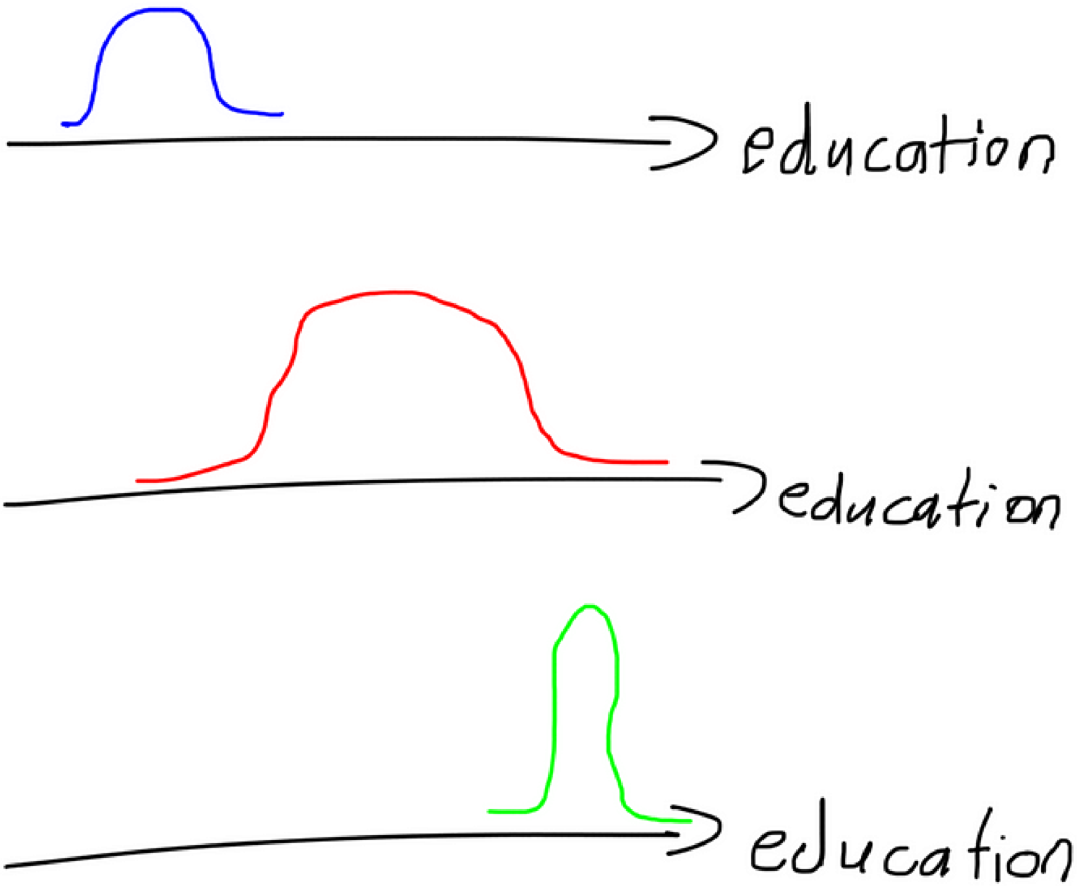
Gaussian Mixture Models Clustering Algorithm Explained

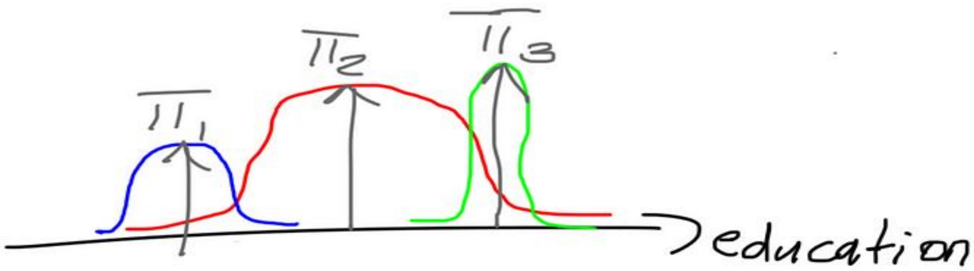
As the name implies, a Gaussian mixture model involves the mixture (i.e. superposition) of multiple Gaussian distributions. For the sake of explanation, suppose we had three distributions made up of samples from three distinct classes.

The blue Gaussian represents the level of education of people that make up the lower class. The red Gaussian represents the level of education of people that make up the middle class, and the green Gaussian represents the level of education of people that make up the upper class.



Not knowing what samples came from which class, our goal will be to use Gaussian Mixture Models to assign the data points to the appropriate cluster.

After training the model, we'd ideally end up with three distributions on the same axis. Then, depending on the level of education of a given sample (where it is located on the axis), we'd place it in one of the three categories.



Every distribution is multiplied by a weight π to account for the fact that we do not have an equal number of samples from each category. In other words, we might only have included 1000 people from the upper class and 100,000 people from the middle class.

Since, we're dealing with probabilities, the weights should add to 1, when summed.

$$\pi = [\overset{\pi_1}{0.31}, \overset{\pi_2}{0.48}, \overset{\pi_3}{0.21}]$$

$$\sum_{K=1}^{K} \pi_K = 1$$

Let’s suppose we wanted to know what is the **likelihood that the ith sample came from Gaussian *k***. We can express this as:

$$p(z_i = k \mid \theta)$$

Where theta represents the mean, covariance and weight for each Gaussian.

$$\theta = \{\mu_1, \mu_2, \mu_3, \Sigma_1, \Sigma_2, \Sigma_3, \pi_1, \pi_2, \pi_3\}$$

You may also come across the equation written as π. This is not to be confused with the weight associated with each Gaussian.

$$p(z_i = k \mid \theta) = \pi_k$$

Next, we express the **likelihood of observing a data point given that it came from Gaussian K** as:

$$p(x_i \mid z_i = k, \mu_k, \Sigma_k)$$

The latter is sometimes written as follows (I believe the N comes from Normal Distribution):

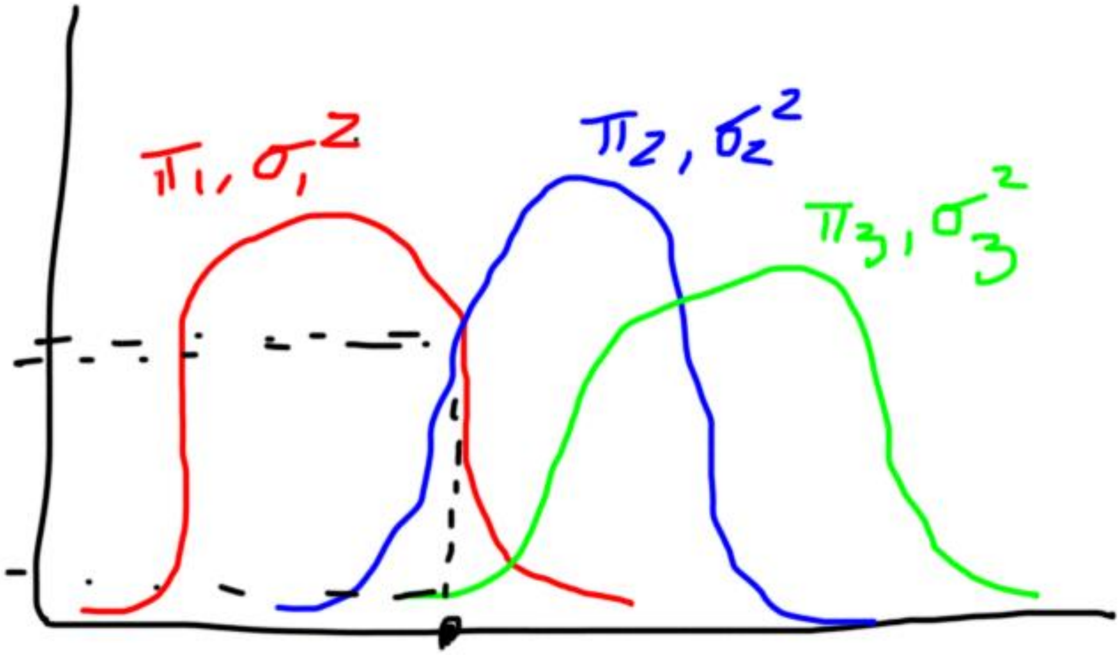
$$p(x_i | z_i = k, \mu_k, \Sigma_k) = N(x_i | \mu_k, \Sigma_k)$$

Suppose we had a Gaussian distribution where the horizontal axis is the different IQ scores an individual could possibly get, lowest through highest. We can find out how likely it is for an individual to have an IQ of 120 by drawing a vertical line from the position along the x-axis to the curve and then looking at the corresponding value on the y-axis. The value of y at any point is equal to the equation above.

If we'd like to know the likelihood of observing the sample ***i*** while taking into account all the different distributions, we simply sum the likelihoods of observing the sample given that it came from each of the possible Gaussian.

$$p(x_i | \theta) = \sum_{k=1}^{3} p(x_i | z_i = k, \theta) p(z_i = k | \theta)$$

Said differently, we take one sample (row) from our dataset, look at a single feature (i.e. level of education), plot its position on the x-axis and sum the corresponding y values (likelihood) for each distribution.



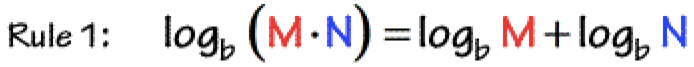
In order to extend this to all samples in our dataset. We assume the likelihood of observing one sample is independent from all the others and then we can simply multiply them.

$$L(\theta) = p(x|\theta) = \prod_{i=1}^{N} \sum_{k=1}^{3} p(x_i | z_i = k, \theta)\, p(z_i - k | \theta)$$

We can rewrite the equation using the nomenclature we saw previously as follows

$$L(\theta) = p(x|\theta) = \prod_{i=1}^{N} \sum_{k=1}^{3} \mathcal{N}(x_i | \mu_k, \Sigma_k)\, \pi_k$$

More often than not, we take the log of the likelihood because the multiplication of two numbers inside of a log is equal to the sum of the logs of its constituents, and it's easier to add numbers than to multiply them.



Rule 1: $\log_b(M \cdot N) = \log_b M + \log_b N$

$$\log(p(x|\theta)) = \sum_{i=1}^{N} \log(\sum_{K=1}^{3} N(x_i|\mu_k, \Sigma_k)\pi_k$$