Topic 2
The Maximum Likelihood and Maximum a Posteriori Paradigms
The Expectation Maximization Algorithm
(Draft)

Instructor: Farid Alizadeh

February 1, 2020

1 Density and likelihood functions

The maximum likelihood method is a way of estimating unknown parameters using the density functions. Let us describe the set up.

Suppose a random variable X or even a random vector has a density function $f_X(t|\theta)$ where $\theta = (\theta_1, \ldots, \theta_k)$ is a list of unknown parameters. For example, if X could be a normal random variable and in that case $\theta = (\mu, \sigma)$ where μ is the mean, and σ is the standard deviation; both or one of them are unknown.

Now suppose we have several *observations* of the random variable X, say $X_1, X_2, \ldots, X_n$. They are assumed to be independent and all have the same distribution. In short we say the X_i are *i.i.d.*, (independently and identically distributed). The question is how do we estimate the parameters $\theta_1, \ldots, \theta_k$ from these n observations of the X. We some times call the X_i's *realizations*a of the random variable X. In statistics we simply call them *the sample* or data. For instance, suppose we believe that the income level of people in the state of New Jersey follows the normal distribution, but we don't know μ the mean income and σ, the standard deviation of the income. So, if we take a sample of 100 people and record their income, then these 100 incomes are each a realization of the normal random variable.

The *likelihood function* is the *joint* density function of observations $X_1, \ldots, X_n$, **but treated as a function of the (unknown) parameters**. Here are a couple of examples.

Example 1: Likelihood function for chance of success in Bernoulli Trials Suppose you wish to see what proportion of the population is willing to buy an

electric car in the next fiscal year. Suppose we ask 10 people and three of them say they plan to buy an electric car next year. If the probability that a given person plans to buy an electric car next year is p, then the number of people out of ten who plan to buy an electric car next year follows the binomial distribution with density function

$$b(k|p,n) = \binom{n}{k} p^k (1-p)^{n-k}$$

Here we know $n = 10$, and we know that $k = 3$, so the *likelihood function* is the function of the unknown parameter p, and is given by

$$\text{Lik}(p) = \binom{10}{3} p^3 (1-p)^7$$

Another way of looking at this problem is the following. Suppose you were told that of the ten people asked only the first, the third and the seventh said that they plan to buy an electric car next year. Each person's answer is a Bernoulli random variable which takes the value of 1 with probability p and 0 with probability $(1-p)$. So the density function is $p^t(1-p)^{1-t}$ where t can be either 0 or 1. So the survey of the ten people above can be encoded as a vector of realizations $\mathbf{X} = (1,0,1,0,0,0,1,0,0,0)$. The likelihood of this function is given by

$$p\times(1-p)\times p\times(1-p)\times(1-p)\times(1-p)\times p\times(1-p)\times(1-p)\times(1-p) = p^3(1-p)^7$$

Notice that this likelihood is up to a constant the same as the one in the binomial approach.

Example 2: Likelihood function for multiple choices Generalizing Bernoulli trials, suppose we have m different choices called $\{1,2,\ldots,m\}$. We run a number of experiments and observe that k_1 times '1' occurred, k_2 times '2', and so on until k_m times 'm'. We wish to compute the maximum likelihood estimates of probability p_i for each of outcomes $i = 1,2,\ldots,m$. For example, suppose $m = 3$ and the data we have is the following: $2,1,1,3,1,3,3,2,1,3$. We are seeking p_1, p_2, and $p_3 = 1 - p_1 - p_2$. If in an experiment, a '1' comes out, the likelihood is given by p_1, and if a '2' or '3' pops up it is given by p_2, or p_3. Assuming each of the occurrences are independents of the others, the likelihood of this sequence is given by $\text{lik}(p_1,p_2,p_3) = p_2p_1p_1p_3p_3p_2p_1p_3 = p1^4p_2^2p_3^4$. So in the general case the likelihood is given by

$$\text{lik}(p_1,p_2,\ldots,p_m) = p_1^{k_1}p_2^{k_2}\ldots p_m^{k_m}$$

Since $p_1+p_2+\cdots+p_m = 1$, we can eliminate one of these parameters, say p_m, set $n = k_1+k_2+\cdots+k_m$, and rewrite the likelihood function as:

$$\text{lik}(p_1,p_2,\ldots,p_{m-1}) = p_1^{k_1}p_2^{k_2}\ldots p_{m-1}^{k_{m-1}}\left(1-p_1-p_2-\cdots-p_{m_1}\right)^{n-k_1-k_2-\cdots-k_{m-1}}$$

Example 3: Likelihood function of mean and standard deviation for normal data Suppose you believe the income of New Jersey residents follows the normal

distribution, but you don't know the mean μ and the standard deviation σ. Suppose You ask three randomly chosen people about their income and record them as $\mathbf{X} = (\$35, \$45, \$40)$ in thousands of dollars. Now, we know that the formula for the pdf of the normal distribution is given by:

$$f_X(t) = \phi(t|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2\right)$$

The *likelihood* function is calculated by plugging in for t the values of 35, 45 and 40. We then multiply these values, since the three observations or realizations of the random variable X are i.i.d., so they are independent. Thus, the likelihood function is given by:

$$\begin{aligned}\mathrm{Lik}(\mu,\sigma) &= \phi(35|\mu,\sigma)\times\phi(45|\mu,\sigma)\times\phi(40|\mu,\sigma)\\ &= \left(\frac{1}{\sqrt{2\pi}}\right)^3\frac{1}{\sigma^3}\exp\left(-\frac{1}{2}\left(\frac{35-\mu}{\sigma}\right)^2\right)\exp\left(-\frac{1}{2}\left(\frac{45-\mu}{\sigma}\right)^2\right)\exp\left(-\frac{1}{2}\left(\frac{40-\mu}{\sigma}\right)^2\right)\\ &= \left(\frac{1}{\sqrt{2\pi}}\right)^3\frac{1}{\sigma^3}\exp\left(-\frac{1}{2}\left[\left(\frac{35-\mu}{\sigma}\right)^2+\left(\frac{45-\mu}{\sigma}\right)^2+\left(\frac{40-\mu}{\sigma}\right)^2\right]\right)\end{aligned}$$

2 The Maximum Likelihood Approach

Suppose we wish to estimate a vector of parameters $\theta = (\theta_1,\ldots,\theta_p)$. The method of *maximum likelihood (ML)* says that in order to estimate θ, we should find the one that maximizes the likelihood function of θ:

$$\text{most likely } \theta = \theta_{\mathrm{ML}} = \mathrm{argmax}_\theta\,\mathrm{Lik}(\theta)$$

In many cases it is more convenient to maximize the *log likelihood* function $\log\mathrm{Lik}(\theta)$ or *minimize* the *negative log likelihood*. Mathematically it does not make a difference whether we maximize $\mathrm{Lik}(\theta)$ or $\log\mathrm{Lik}(\theta)$ or minimize $-\log\mathrm{Lik}(\theta)$. However, in many cases it is more convenient to maximize $\log\mathrm{Lik}(\theta)$; we will see some cases below.

Example 1 Continued: We saw that the likelihood function for the probability of buying an electric car in the next year, p, is given by $Cp^3(1-p)^7$, where C is a constant not depending on p. So in this case the vector of unknown parameters contains only one item: p. Now the log likelihood is

$$\log\mathrm{Lik}(p) = c + 3\log(p) + 7\log(1-p)$$

In this case it is straightforward to find the p that maximizes the log likelihood function. Let's solve it for arbitrary numbers. Suppose we had asked n people whether they will buy an electric car within a year, and k people said they will. Then the log likelihood function would have been

$$\log\mathrm{Lik}(p) = \log\left[Cp^k(1-p)^{n-k}\right] = c + k\log(p) + (n-k)\log(1-p)$$

where, $c = \log(C)$ is a constant. To maximize this we simply take the derivative with respect to p and set it equal to zero:

$$\begin{aligned}\frac{d}{dp}\log \text{Lik}(p) &= \frac{d}{dp}\big(c + k\log(p) + (n-k)\log(1-p)\big)\\ &= \frac{k}{p} - \frac{n-k}{1-p}\\ &= \frac{k(1-p) - (n-k)p}{p(1-p)}\end{aligned}$$

Setting this derivative equal to zero we get

$$\begin{aligned}\frac{k(1-p) - (n-k)p}{p(1-p)} &= 0\\ k - kp - (n-k)p &= 0\\ np &= k\\ p &= \frac{k}{n}\end{aligned}$$

So we say, that $p = k/n$ is the *maximum likelihood estimator* of the parameter p. This is of course the most natural estimation, but it is also verified by the ML method.

Example 2 Continued: The log likelihood function is given by:

$$\begin{aligned}\log \text{lik}(p_1, p_2, \ldots, p_{m-1}) = k_1\log(p_1) + k_2\log(p_2) + \cdots + k_{m-1}\log(p_{m-1}) +\\ (n - k_1 - k_2 \cdots - k_{m-1})\log\big(1 - p_1 - p_2 - \cdots - p_{m-1}\big)\end{aligned}$$

We can now take the derivative, the gradient, with respect to each of variables $p_1, p_2, \ldots, p_{m-1}$ and set it equal to zero and try to solve the problem:

$$\begin{aligned}\frac{\partial}{\partial p_1}\log \text{lik}(p_1, p_2, \ldots, p_{m-1}) &= \frac{k_1}{p_1} - \frac{n - k_1 - k_2 - \cdots - k_{m-1}}{1 - p_1 - p_2 - \cdots - p_{m-1}} = 0\\ \frac{\partial}{\partial p_2}\log \text{lik}(p_1, p_2, \ldots, p_{m-1}) &= \frac{k_2}{p_2} - \frac{n - k_1 - k_2 - \cdots - k_{m-1}}{1 - p_1 - p_2 - \cdots - p_{m-1}} = 0\\ &\cdots\cdots\cdots\\ \frac{\partial}{\partial p_{m-1}}\log \text{lik}(p_1, p_2, \ldots, p_{m-1}) &= \frac{k_{m-1}}{p_{m-1}} - \frac{n - k_1 - k_2 - \cdots - k_{m-1}}{1 - p_1 - p_2 - \cdots - p_{m-1}} = 0\end{aligned}$$

Solving this system of $m-1$ equations for p_i we get:

$$\frac{k_1}{p_1} = \frac{k_2}{p_2} = \cdots = \frac{k_{m-1}}{p_{m-1}} = \frac{n - k_1 - k_2 - \cdots - k_{m-1}}{1 - p_1 - p_2 - \cdots - p_{m-1}} = \frac{n}{1}$$

The last equality is the result if addiing the numerators and denominators of the equal ratios. This set of equations now yields:

$$p_1 = \frac{k_1}{n}, p_2 = \frac{k_2}{n}, \ldots p_{m-1} = \frac{k_{m-1}}{n}, p_m = \frac{k_m}{n}.$$

So, again, the sample proportion is the most likely estimate of probabilities.

Example 3 Continued: We saw earlier three observations of the annual income of New Jersey residents in thousands of dollars): $\mathbf{X} = (\$35, \$45, \$40)$. We also assumed that these income levels follow the normal distribution, with unknown mean μ and unknown standard deviation σ. The three observations were three independent and identically distributed (i.i.d) *realizations* of the random variable. Therefore, in this setting the vector of unknown parameters $\theta = (\mu, \sigma)$. We derived the likelihood function earlier. We now derive the log likelihood function and attempt to minimize it. The negative log likelihood function is:

$$\begin{aligned} -\log \text{Lik}(\mu, \sigma) &= 3\log\sqrt{2\pi} + 3\log(\sigma) + \frac{1}{2}\left[\left(\frac{35-\mu}{\sigma}\right)^2 + \left(\frac{45-\mu}{\sigma}\right)^2 + \left(\frac{40-\mu}{\sigma}\right)^2\right] \\ &= 3\log\sqrt{2\pi} + 3\log(\sigma) + \frac{1}{2\sigma^2}\left[(35-\mu)^2 + (45-\mu)^2 + (40-\mu)^2\right] \end{aligned}$$

To find the minimum again we need to take derivatives with respect to both μ and σ and set them equal to zero. This will result in a system of two equations in two unknowns:

$$\frac{\partial}{\partial\mu}\left[3\log\sqrt{2\pi} + 3\log(\sigma) + \frac{1}{2\sigma^2}\left((35-\mu)^2 + (45-\mu)^2 + (40-\mu)^2\right)\right] = 0$$

$$\frac{\partial}{\partial\sigma}\left[3\log\sqrt{2\pi} + 3\log(\sigma) + \frac{1}{2\sigma^2}\left((35-\mu)^2 + (45-\mu)^2 + (40-\mu)^2\right)\right] = 0$$

In this case taking derivatives with respect to μ and setting it equal to zero results in a linear function involving only μ:

$$\frac{1}{2\sigma^2} \times 2 \times \left[(35-\mu) + (45-\mu) + (40-\mu)\right] = 0$$

which results in the solution

$$\mu = \bar{X} = \frac{35+45+40}{3}$$

In other words the most likely estimate of μ is the average of the observed data, that is the *sample mean* $\bar{X}$. Once we plug in for μ in the second equation $\bar{X}$ and take the derivative with respect to σ^2 (note: derivative is taken with respect to σ^2 not σ), set to zero and solve, we get

$$\sigma^2 = \frac{(45-\bar{X})^2 + (35-\bar{X})^2 + (40-\bar{X})^2}{3}$$

which is the sample variance (but divided by $n = 3$ not $n-1 = 2$ which is the more common definition of sample variance).

This example can be extended to the case where there are N observed data (realization of the normal i.i.d random variables). Then the maximum likelihood estimated of the mean is the sample mean, and the maximum likelihood estimate of $\sigma^2 =$ is the sample variance (divided by N not $N-1$):

$$\mu_{ML} = \bar{X} = \frac{X_1 + \cdots + X_N}{N} \quad \sigma^2_{ML} = s^2 = \frac{(X_1-\bar{X})^2 + \cdots + (X_N-\bar{X})^2}{N}$$

These three examples are not very typical in that in most problems, there is no *neat formula* for the optimal value of the unknown parameters. Instead the optimal values have to be estimated by iterative procedures for optimization of the likelihood function or the log likelihood function. In general, minimization or maximization of functions of several variables require specialized optimization algorithms. We will deal with this in future lectures when we discuss the topic of logistic regression.

3 Properties of the Maximum Likelihood Estimation

3.1 ML estimates are approximately normal for large N

Suppose we are trying to estimate a parameter θ using the maximum likelihood method and from i.i.d observations $X_1, \ldots, X_N$. Here θ is *any* parameter you can imagine, for example the mean, the variance, the coefficients of a linear function, the proportion, etc. Suppose θ_{extML*} is the maximum likelihood estimation after these N observations, that is

$$\theta_{ML} = \text{argmax}_\theta \, \text{Lik}\big(\theta | X_1, \ldots, X_N\big).$$

Then, in the frequentist framework, the θ_{ML} itself is a random variable (because it is estimated from random variables X_i). It can be shown that under certain conditions, the maximum likelihood estimate θ_{ML} follows approximately the normal distribution $N\big(\theta, \sigma_N(\theta)\big)$ and this approximation becomes more accurate as the number of observations N gets larger. It can also be shown that under these conditions $\sigma_N(\theta) \to 0$ as N approaches infinity, meaning that the maximum likelihood estimate θ_{ML} will be very close to the *true* θ.

The standard deviation $\sigma_n(\theta)$ is related to the second derivative of the log likelihood function. This quantity itself tends to zero as N gets larger, indicating that the maximum likelihood estimate θ_{ML} will be very close to "real" θ for very large N.

This fact gives justification for using the maximum likelihood method in parameter estimation in general, and machine learning in particular. We will look at the applications of the maximum likelihood method in *logistic regression* in future.

3.2 Likelihood ratio statistic

Suppose we have a vector of parameters $\theta = (\theta_1, \ldots, \theta_k, \theta_{k+1}, \ldots, \theta_{k+m})$. Suppose that the first k parameters are the ones we are really interested in, and the other m parameters are needed as part of the density function of the data. For example, suppose we are interested in the variance of income in a country or state. Suppose we believe that the income follows the normal distribution $N(\mu, \sigma)$. Here we are not really interested in the average income μ but on the

variance σ^2. Suppose we have a sample of $X_1, X_2, \ldots X_N$. We could estimate σ^2 in two different ways.

1. **Method 1:** We could assume μ is known and equal to μ_0, a fixed and known number (for example \$30 thousand). Then we can take the -log likelihood function and minimize it with respect to σ^2:

   $$-\log \text{Lik}(\sigma^2 | X_1, \ldots, X_n, \mu = \mu_0) = N \log(2\pi) + N \log(\sigma^2) + \frac{(X_1 - \mu_0)^2 + \cdots + (X_N - \mu_0)^2}{\sigma^2}$$

   In this form, the negative log likelihood is a function of σ^2 only. Taking derivatives (with respect to σ^2) and setting it equal to zero we get:

   $$\sigma^2_{ML} = \frac{(X_1 - \mu_0)^2 + \cdots + (X_N - \mu_0)^2}{N}$$

2. **Method 2:** We would not assume anything about μ. We simply decide to estimate it from the data along with σ^2. We have seen that in this case

   $$\mu_{ML} = \bar{X} \quad \sigma^2_{ML} = \frac{(X_1 - \bar{X})^2 + \cdots + (X_N - \bar{X})^2}{N}$$

The question is whether the smaller model (the one that assumes $\mu = \mu_0$ is enough, or whether to estimate σ^2 we must also simultaneously estimate μ as well. There is a hypothesis test for this question. We state it for the more general case.

Suppose we wish to estimate parameters $\theta_1, \ldots, \theta_k$. But the pdf, and thus the likelihood function not only depends on these parameters, it also depends on m additional parameters $\theta_{k+1}, \ldots, \theta_{k+m}$. We can have two models:

1. **The reduced model:** The parameters $\theta_{k+1}, \ldots, \theta_{k+m}$ are fixed. In this case the log likelihood function is

   $$\log \text{Lik}\big(\theta_1, \ldots, \theta_k | X_1, \ldots, X_n, \theta_{k+1}, \ldots, \theta_{k+m}\big)$$

   that is, it is only a function $\theta_1, \ldots, \theta_k$ and the other parameters are known.

2. **The Complete Model:** All parameters are unknown and have to be estimated from the data. In this case the log likelihood function is

   $$\log \text{Lik}(\theta_1, \ldots, \theta_{k+m} | X_1, \ldots, X_n)$$

   that is, it is a function of the $k + m$ parameters $\theta_1, \ldots, \theta_{k+m}$.

To decide which model is correct we can formulate the following hypothesis test:
The Null hypothesis: The claimed values of parameters $\theta_{k+1}, \ldots, \theta_{k+m}$ are correct (that is the reduced model is correct).

The Alternative hypothesis: The claimed values of the parameters $\theta_{k+1}, \ldots, \theta_{k+M}$ are not correct (that is the complete model should be used.)

The approach taken to find the p-value of this test is based on the following fact:

> **Wilk's Theorem:** Define
>
> $$\Lambda = \log \text{Lik}(\theta_1, \ldots, \theta_k) - \log \text{Lik}(\theta_1, \ldots, \theta_k, \theta_{k+1}, \ldots, \theta_{k+m})$$
>
> If the Null hypothesis is correct $2\Lambda \sim \chi^2(m)$ that is 2Λ follows the Chi-square distribution with m degrees of freedom.

by computing the p-value (using the $\chi^2(m)$ distribution) we can decide if the reduced or the complete model should be used.

This test will become handy in determining whether adding new features to a classification or regression problem improves the model or not.

4 Maximum a Posteriori (MAP) or the Bayesian Approach

In the Bayesian approach, we incorporate our *prior* knowledge on the distribution of parameters into our model. First, there is a fundamental difference of point of view between Bayesian and non-Bayesian (often called *frequentist*) approaches to parameters:

The frequentist approach to parameter estimation:

- Parameters are assumed to be fixed (non-random) numbers out there which are unknown to us. We collect data and use them to estimate these parameters. For example, when the parameter of interest is the proportion of customers who plan to buy an electric car next year, this proportion is assumed to be an unknown *constant* that needs to be estimated. Or the average income of New Jersey residents is assumed to be an unknown fixed number that needs to be estimated.

- Data (or the sample) are assumed to be random, and one possible realization of many other possibilities. For instance, if we ask two people to choose ten random persons and ask them if they plan to buy an electric car next year, the first person's sample is randomly chosen, and similarly the second person's sample is also random. So their *sample proportion*, which is a random variable, may be different.

- More generally, and philosophically, probability of an event is modeled as if an experiment can be repeated many times, and the number *the frequency* of times that the event occurs approximates its probability. (This explains the term "frequentist").

- In frequentist estimation, parameters are estimated by a numbers; these numbers are *point estimation* of the corresponding parameters. They express their uncertainty or degree of confidence about their estimations in the form of *confidence intervals* (for a single parameter), or *confidence region* for a vector of parameters.

The Bayesian approach to parameter estimation:

In the Bayesian approach:

- Parameters themselves are considered random variables with a distribution. The idea is that it is not important whether the unknown parameters are fixed numbers or not. What is crucial is that we don't know their value, and that *ignorance* is encoded by a probability distribution.

- More generally, and philosophically, the probability of an event is modeled as the amount of information we have to expect its occurrence (see below for further explanation).

- The data are often considered fixed, or if not, the distributions are conditioned on the given data, so they are effectively treated as fixed.

- To start a statistical or learning task we first encode our ignorance about it through a probability distribution. For instance, if we are interested in New Jersey average household income we could say that this quantity follows the normal distribution with mean \$50,000 and standard deviation \$10,000. Note that this distribution encodes our current ignorance *before* any data is collected. We may have some idea about the New Jersey household income. May be similar surveys were made a few years before, or we know something about the income of neighboring states. This distribution is the *prior distribution*. In this example, the size of standard deviation indicates our degree of ignorance. Large σ means we are not sure; small σ means we are more confident.

- Once data is collected we use Bayes formula to *adjust* our probability distribution, that is our level of knowledge, about parameters (see below for details). Thus, after asking, say 100 New Jersey residents about their income, we can combine this information with our prior, and obtain a *posterior distribution* of the average income. Hopefully, this distribution encodes more information by having a smaller σ.

- Bayesians are not content with a simple point estimation, and search for a probability distribution to capture all their knowledge about a a parameter or a set of parameters.

The first point above is crucial for the Bayesian point of view. What this view requires us to provide is not just a point estimate of the the unknown parameter.

Rather we must provide a *probability distribution* for it. This distribution will encode our degree of belief, or confidence about this parameter. For instance, if we encode our degree of belief for parameter θ by the normal distribution with mean 100 and $\sigma = 3$, we are essentially saying that we believe $\theta \approx 100 \pm 3$. On the other hand if we encode our degree of belief for this same parameter θ with the normal distribution with mean 100 and $\sigma = .1$ we are saying that we believe $\theta \approx 100 \pm 0.1$. Obviously the second distribution is much more concentrated near 100, indicating our strong confidence that $\theta = 100$. This same point of view is applicable to multiple parameters. Here we need to produce a *joint distribution* of the parameters that encode our degree of belief.

Parameters as random variables

The idea is that even if the unknown parameters are some constants out there, it is our *lack of knowledge, and ambiguity about them* that makes them random. For instance, consider the statement *"there is a 50% chance that it will rain tomorrow."* Exactly what does the "chance" or randomness refer to here? A simple minded interpretation of the frequentist point of view would be that there are multiple possible tomorrows, half of which are rainy and the other half not rainy. The actual tomorrow then is assumed to be drawn randomly from this basket of multiple tomorrows!

In the Bayesian approach, when we say the chance of rain tomorrow is 50%, we are expressing our current lack of knowledge, our uncertainty, about tomorrow's weather. We use the expertise of a meteorologist and based on that knowledge asses the chance of rain to be 50%. The "experience" or "expertise" itself could be seen as a more sophisticated variation of the frequentist approach: meteorologists may use their experience to come up with this probability. They may say that in the past when similar atmospheric conditions existed, and in similar time an geographical location, half the time it resulted in rain the next day. This is the expert using a frequentist approach (dubbed as "experience" and "expertise") to estimate the chance of rain tomorrow.

From the Bayesian point of view, we have some knowledge about the parameters $\theta = (\theta_1, \ldots, \theta_p)$, *before any data is collected.* This knowledge is encoded as a *prior distribution* (or joint distribution for multiple parameters) and encapsulated in a (joint) pdf $f(\theta|\gamma)$. Here γ is a set of parameters for the prior distribution of parameters. To distinguish them from parameters of the data (that is θ) the numbers in γ are sometimes called *hyperparameters*. Let's look at a couple of examples to see what is going on.

Example 1 Continued: The maximum likelihood approach can give poor results if the data size is small. For instance, suppose that you ask ten people if they are going to buy an electric car next year, and none of them says yes. By the maximum likelihood approach, $k = 0$ and your maximum likelihood estimate for proportion of people buying electric cars next year would be zero. This situation is not uncommon at all, especially when the "true proportion" is small with respect to your sample size. For instance, if in reality only 1% of customers are going to buy an electric car next year, and you have asked

90 people, then there is a very good chance that none of them would say they would buy an electric car next year.

Now suppose before collecting any data we interview expert car salespersons and ask them what they think the proportion of people who are going to buy an electric next year is. And suppose that they, based on their knowledge, experience, other forms of market analysis, trends and similar information, come up with the estimate of about 10%. However, just having such an estimate is not enough. They also need to supply in some way their confidence on their estimate.

If you recall, in the Laplace Smoothing process in the Naive Bayes method, we replaced the estimate $\frac{k}{n}$ by $\frac{k+\alpha}{n+\alpha+\beta}$. The[1] interpretations roughly was our implicit assumption that we had already made $\alpha+\beta$ *virtual* observations and among them α were planning to buy an electric car next year. Note that we are providing more information than simply the ratio $\frac{\alpha}{\alpha+\beta}$. For example, the quantity $\frac{k+1}{n+10}$ is not the same $\frac{k+10}{n+100}$. Both have a prior estimate of $\frac{\alpha}{\alpha+\beta}=0.1$, but one assumes ten prior (virtual) observations out of which one was success, while the other assumes one hundred (virtual) observations of which ten were success. The latter is more confident about the accuracy of the prior ratio. Let us see how this can arise from a Bayesian point of view.

Suppose we assume that the prior ratio p has a distribution with the following density function:

$$f(p|\alpha,\beta) = Bp^{\alpha-1}(1-p)^{\beta-1}$$

Random variables with this type of density function are said to follow the *beta distribution*. Note that this is a *continuous distribution* (since p the variable can be any real number between zero and one.) The constant $B = B(\alpha,\beta)$ is chosen so that the area under the graph of $f(p|\alpha,\beta)$ equals one[2]. This distribution is defined only for $0 \le p \le 1$ and for parameters $\alpha,\beta > 0$. In this case α,β are the hyperparameters.

With this prior density let us see what happens to the maximum likelihood estimate:

$$\begin{aligned} \text{Lik}(p|k,n) &\propto f(k|p,n)f(p) \\ &\propto p^k(1-p)^{n-k}p^{\alpha-1}(1-p)^{\beta-1} \\ &= p^{k+\alpha-1}(1-p)^{n-k+\beta-1} \end{aligned}$$

As we saw earlier using log likelihood and basic calculus we find that the maximizing value for p is $p^* = \frac{k+\alpha-1}{n+\alpha+\beta-2}$. This is equivalent to us having observed $\alpha+\beta-2$ prior (virtual) observations, and $\alpha-1$ of them were success. After running the experiment (receiving new data) we saw k success among n trials. In essence α and β hyperparameters encode both our prior estimate of

[1] We made a small change in that in the denominator we add $\alpha+\beta$ not just β. This ensures that the item added to the denominator is larger than the one in the numerator.

[2] The constant $B(\alpha,\beta)$ as a function of α,β is called the *beta function*, thus the name beta distribution.

$p \approx \frac{\alpha-1}{\alpha+\beta-2}$, and our confidence in our belief: if α and β are both small it indicates small confidence; if they are large (with their ratio fixed) it indicates high confidence. To see this suppose, we choose $\alpha = 2$ and $\beta = 11$. If we made $n = 30$ (actual) observations and $k = 2$ were success, our Bayesian estimate would be $\frac{2+2-1}{30+11+2-2} = 7.3\%$. But if we make additional (actual) observations and where this time $n = 300$ and $k = 20$ were success, then the Bayesian estimate would be $\frac{20+2-1}{300+11+2-2} = 6.5\%$ which is closer to the pure ML value of $k/n = 2/30 = 6.67\%$ number without the Bayesian prior. On the other hand if $\alpha = 10, \beta = 110$ were given, then in the first case our estimate would've been $\frac{2+10-1}{30+10+110-2} = 7.4\%$ and in the second case we'd get $\frac{20+10-1}{300+10+110-2} = 6.9\%$. With $\alpha = 10, \beta = 110$ the influence of the prior is more than when $\alpha = 1, \beta = 11$. In both cases though, as data gets larger, the influence of the prior diminishes. However for larger α and β the pace is slower and the influence of the prior lingers longer.

The Bayesian estimate $p = \frac{k+\alpha-1}{n+\alpha+\beta-2}$ is called the *Maximum À Posteriori (MAP)* estimate of p, and is denoted by p_{MAP}.

Note that if we had chosen $\alpha = \beta = 1$, then the optimal estimate would become $p^* = k/n$ and this encodes the case that zero prior virtual observations and zero virtual success ($\alpha - 1 = 0, \beta - 1 = 0$). So we have no prior knowledge of p at all. So for $\alpha = \beta = 1$ the beta distribution turns into the uniform distribution, which assigns equal likelihood to all values of between 0 and 1. This situation is sometimes called *non-informative prior*. In this case the MAP and maximum likelihood estimates coincide.

In Bayesian point of view, computing the most likely value for p is only one part of the process of estimation. We now have to give a *posterior distribution* for p. Here are what we know:

- Before any observation, our prior distribution for p was the beta distribution with parameters α, β:

  $$f(p) = Bp^{\alpha-1}(1-p)^{\beta-1}$$

  α and β are the hyper parameters, giving a shape to the probability distribution encoding our prior belief.

- We collected data, that is asked n people and observed k successes. We would like to update the posterior distribution of p, which by using Bayes theorem is given by

  $$\text{posterior pdf:} \quad f(p|n,k) = \frac{f(k|n,p)f(p|\alpha,\beta)}{f(k)}$$

  Again the denominator does not involve p and is treated as a constant. The quantity $f(k|n,p)$ is the probability that k successes occur given n trials and probability of one success equal to p. This is just the *binomial* distribution:

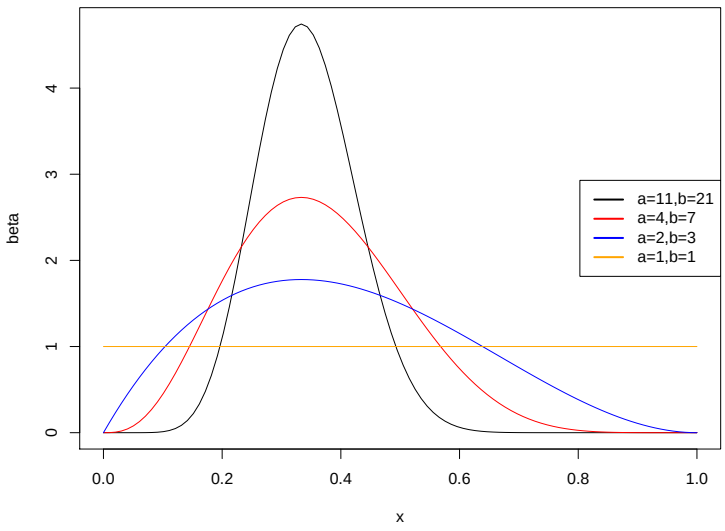
  $$f(k|n,p) = \binom{n}{k} p^k (1-p)^{n-k}$$

- Thus the posterior probability is given by

$$\begin{aligned} f(p|n,k) &= \frac{\binom{n}{k}p^k(1-p)^{n-k} \times Bp^{\alpha-1}(1-p)^{\beta-1}}{f(k)} \\ &= \text{Constant} \times p^{k+\alpha-1}(1-p)^{n-k+\beta-1} \end{aligned}$$

  So the posterior also follows the beta distribution, but with updated parameters. In fact, since $\alpha_{\text{new}} = k+\alpha$ and $\beta_{\text{new}} = n-k+\beta$, the parameters of the new beta distribution are larger. This has the effect of making the beta distribution more concentrated on the most likely ratio $\frac{k+\alpha-1}{n-k+\beta-1}$.

Here are plots of the beta distribution for various values of α, β



The particular choice of the beta distribution for the prior made our calculations very simple. Had we chosen a different prior distribution on p, we would probably have got a different result, and in addition the computation of the optimal estimate would have been more complicated.

The fact that the combination using the beta distribution for prior, and the binomial distribution for the number of successes results in again a beta distributed posterior is very convenient. In this case we say that the binomial and beta distributions are *conjugate distributions* of each other. For example, we can use this fact to set up for when the data is presented to us as a sequence. For instance, suppose we choose our prior with $\alpha = 2$ and $\beta = 21$, indicating that we think 1 out of 20 or 5% of customers plan to buy an electric car next year. We ask $n_1 = 100$ people, and say, $k_1 = 7$ say they plan to buy electric cars. Now our posterior distribution is beta with parameters $\alpha + k_1 = 9$ and

$\beta + (n_1 - k_1) = 112$. Now if new data comes, that another $n_2 = 50$ people were asked and $k_2 = 3$ said they plan to buy electric cars, then we can use the previous posterior as the new prior and run the experiment again. This means that now our parameters are $\alpha + k_1 + k_2$ and $\beta + (n_1 + n_2 - k_1 - k_2)$.

Example 2 Continued: We saw that the maximum likelihood estimate of the mean of a normally distributed i.i.d observations $X_1, \ldots, X_N$ is the sample mean $\bar{X} = \frac{X_1 + \cdots + X_N}{N}$. What if we have some prior knowledge about where the mean actually is. For instance, suppose that before any data is taken we think that average income of New Jersey residents is roughly around \$50 thousand. This belief could come from, for instance, surveys from previous years, and by factoring in typical salary increases. But, just as in the probability estimation for proportions, we also need to encode our confidence in our prior belief. One way to do this could be to assume that the prior probability distribution for the mean is normal with mean equal to, say $\mu_P = \$50$ thousand and standard deviation $\sigma_P = \$10$ thousand. Once the three observations (data) of $X_1 = \$35$, $X_2 = \$45$,nd $X_3 = \$35$ is included, the Bayesian likelihood function is given by

$$\begin{aligned}\text{Lik}(\mu, \sigma) = \\ &= \left(\frac{1}{\sqrt{2\pi}}\right)^3 \frac{1}{\sigma^3} \frac{1}{\sqrt{2\pi}\sigma_P} \times \\ &\exp\left(-\frac{1}{2}\left(\frac{35-\mu}{\sigma}\right)^2\right) \times \exp\left(-\frac{1}{2}\left(\frac{45-\mu}{\sigma}\right)^2\right) \times \exp\left(-\frac{1}{2}\left(\frac{40-\mu}{\sigma}\right)^2\right) \times \\ &\exp\left(-\frac{1}{2}\left(\frac{\mu-\mu_P}{\sigma_P}\right)^2\right)\end{aligned}$$

In this case the *hyperparameters* $\mu_P = 50$ and $\sigma_P = 10$ not only encode our *prior* estimate (μ_P), but also our confidence on our prior through σ_P. If we were very sure that \$50 thousand is the average income in New Jersey then we would have chosen our σ_P smaller, say only \$1 thousand. On the other hand, if we were less sure, we would have chosen a larger σ_P, say \$15 thousand. So here the assumption of a normal prior distribution, along with σ_P encodes our prior knowledge (and confidence in that knowledge) before we see the data.

Let us assume, for simplicity, that σ is known. Then taking the log likelihood function, removing terms that do not depend on μ and taking derivative with respect to μ and setting equal to zero, we get:

$$\mu_{\text{MAP}} = \frac{\sigma^2}{N\sigma_P^2 + \sigma^2}\mu_P + \frac{N\sigma_P^2}{N\sigma_P^2 + \sigma^2}\bar{X}$$

where $\bar{X}$ is the sample mean, which is the maximum likelihood estimate of μ (without considering the Bayesian prior), and noting that $\mu_P = 50, \sigma_P = 10$, and $N = 3$.

Notice that μ_{MAP} is a weighted average of our prior estimate μ_P, and the estimate derived from the data ($\bar{X}$). Our confidence of the prior is encoded in the prior standard deviation σ_P, or more accurately in the inverse $\frac{1}{\sigma_P^2}$. If σ_P is

small (and so $\frac{1}{\sigma_P^2}$ is large), our confidence is high and this is reflected in larger weight for the prior. If our σ_P is large, our confidence is less, and this is reflected in a higher weight for the ML estimate.

Now keep σ_P fixed and let N, the number of observations to grow larger. In this case, the weight of μ_P, that is, $\frac{\sigma^2}{N\sigma_P^2+\sigma^2} \to 0$, while the weight of $\bar{X}$, that is $\frac{N\sigma_P^2}{N\sigma_P^2+\sigma^2} \to 1$. So as the number of observation gets larger, the effect of prior diminishes.

However, in the Bayesian approach, we are not simply interested in a point estimate μ_{MAP} of the mean. We wish to give a *distribution* for it. Again, writing the Bayes formula in this case we get the following for the posterior distribution of μ:

$$\begin{aligned} f(\mu|X) &\propto f(X|\mu) f_P(\mu) \\ &= \prod_{i=1}^{n} \phi\big(x_i|\mu_{MAP}, \sigma_N\big)\phi(\mu|\mu_P, \sigma_p) \\ &= \phi\big(\mu_N, \sigma_N\big) \end{aligned}$$

The last equality can be obtained by considerable algebraic manipulation. So, it turns out this product gives rise to a normal distribution again, with mean $\mu_N = \mu_{MAP}$ as its mean, and σ_N^2 as its variance, where

$$\frac{1}{\sigma_N^2} = \frac{1}{\sigma_P^2} + \frac{N}{\sigma_{ML^2}}$$

So, the conjugate of the normal distribution when σ is assumed to be known is again a normal distribution. Since σ_N gets smaller as N gets larger, this means that the posterior distribution is more *informative*: because of smaller σ_N we are more confident about the exact value of μ.

The discussion above can be extended to when σ is also an unknown parameter. In that case we need to have a *joint* prior distribution for μ and σ and derive the joint posterior distribution after observing the data. Suitable conjugate distributions exists that makes computations more manageable. Also, all of this discussion can be further extended to multivariate normal distribution where we need to estimate the vector of means $\boldsymbol{\mu}$ and the covariance matrix Σ.

5 The Expectation Maximization (EM) Algorithm

In many situations the data available to us, the $N \times d$ table, do not explicitly contain all relevant data that we need to estimate parameters. So the data can be divided into two parts X, Z, where X is the *observed* data, and Z is the *unobserved* data or the *latent variables*. Based on this we wish to estimate a set of parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$. Before we continue let's give an example.

Example 3: Mixed Gaussian Model: Recall the airline example we used in Topic 1 about Bayes decision rule. We wished to guess which of the three categories of tickets, First Class, Economy Plus, or Economy should be targeted to people given their income. In that example, we had complete information about the distribution of incomes of each group of ticket buyers, including the relative proportion of each group. So we used the Bayes decision rule to assign a probability to each new person to buy each type of ticket, solely based on their income.

Now assume that we don't have any knowledge about the parameters, that is mean and variance of each group of ticket buyers (we only assume they all follow the normal distribution, but we don't know the parameters). All we have is a collection of incomes from potential customers of the airline: $X_1, X_2, \ldots, X_N$. And we don't know which one buys which type of airline ticket. From here we need to determine:

- An estimate of the mean income of ticket buyers in each group of First Class, Economy Plus, and Economy tickets, μ_1, μ_2, μ_3,
- An estimate of the variance of income in each group $\sigma_1, \sigma_2, \sigma_3$,
- An estimate of the relative proportion of each group τ_1, τ_2 and τ_3 with $\tau_1 + \tau_2 + \tau_3 = 1$ (so there are only two parameters here),
- for each data point X_i estimate the probability that the person is a potential First Class buyer, Economy Plus buyer, or Economy ticket buyer, p_{1i}, p_{2i} and p_{3i} with $p_{i1} + p_{i2} + p_{i3} = 1$.

The vector of parameters we need to estimate is $\theta = (\mu_1, \mu_2, \mu_3, \sigma_1, \sigma_2, \sigma_3, \tau_1, \tau_2)$. On the other hand there is another set of variables at play here which are not given to us: For each customer i the categorical (nonnumerical) variable Z_i, where $Z_i \in \{1, 2, 3\}$. Z_i indicates whether the i^{th} customer is a First Class buyer ($Z_i = 1$), or Economy Plus buyer ($Z_i = 2$) or an Economy buyer ($Z_i = 3$). Note that if the value of Z_i were known, then we would have a straightforward problem: We would estimate the mean and variance of each group from the data by the maximum likelihood methods we discussed earlier, and we would also use sample proportion to estimate the proportion of each population of ticket buyers. But the *latent variables* Z_i are not given and yet play a role in determining these parameters.

So the question is how do we calculate the most likely values for the parameters when parts of the data, or information about some features are hidden or missing.

Formally we need to compute the *marginal likelihood* of the parameters. This means that we need to compute the average likelihood function, where the

average is taken over the latent variables:

$$\begin{aligned}\log \mathrm{lik}(\theta|X) &= \log\Big(\mathbb{E}_Z\big(\mathrm{lik}(\theta|\mathbf{X},\mathbf{Z})\big)\Big)\\ &= \log\Big(\sum_z \mathrm{lik}(\theta \mid \mathbf{X},\mathbf{Z}=z)\Big) \quad \text{If } \mathbf{Z} \text{ is discrete}\\ &= \log\Big(\int_{\mathcal{D}} \mathrm{lik}(\theta|\mathbf{X},\mathbf{Z}=z)\mathrm{d}z\Big) \quad \text{If } \mathbf{Z} \text{ is continuous and ranges over } \mathcal{D}\end{aligned}$$

where $\mathrm{lik}(\theta \mid x, z)$ is obtained from the (joint) pdf of the latent variables $\mathbf{Z}$ and the observed variables $\mathbf{X}$, $f(x, z \mid \theta)$.

Note that in the expressions above we have "the logarithm of expected value of the likelihood function. In general, we are interested in situations where the integral (or sum) above is computationally hard to compute. For instance, for the airline ticket example, one possible algorithm is to consider all combinations of assigning of ticket class to each data point X_i, and compute the average likelihood. This will require 3^N possible cases, which is computationally infeasible, even for moderate sizes of N. On the other hand, we assume that the *conditional* likelihood, that is $\mathrm{lik}(\theta \mid \mathbf{X}, \mathbf{Z} = z)$ is easier to handle. For instance, in the airline ticket, if the values of $\mathbf{Z}_i$ are given, that is we know which ticket type each customer i is going to buy, then the likelihood, and subsequently the maximum likelihood estimation of parameters, is straightforward.

The *Expectation Maximization (EM)* algorithm attempts to take advantage of the easiness of estimating parameters if the latent variables, or their distribution were known, and solves the problem in two stages. Consider the following facts:

- Suppose we know the values of unknown parameters, θ. Then we could use the Bayes decision rule for discrete values of $\mathbf{Z}$ and assign a probability to each possible value of $\mathbf{Z}_i$. Even if $\mathbf{Z}$ is a continuous latent variable, having an estimate of θ we can estimate the distribution of $\mathbf{Z}$.

- Now suppose we know the distribution latent variables $\mathbf{Z}$. Then the problem will reduce to an ordinary maximum likelihood problem, and the values of parameters can be estimated by maximum likelihood methods.

Thus, knowledge of probability distribution of values of $\mathbf{Z}_i$ makes the problem tractable, and knowledge of parameters make computing the probability distribution of $\mathbf{Z}_i$ tractable.

The expectation maximization algorithm exploits this fact to estimate alternately the distribution of the latent variables, and the unknown parameters, and the probability distribution of the latent variables. At the k^{th} iteration of the algorithm, we already have some estimate of the parameters, say, $\theta^{(k)}$. Assuming the latent variables $\mathbf{Z}_i$ are discrete, then we can use Bayes decision rule to estimate the $p_{ia}^{(k)} = \Pr[\text{the } i^{\text{th}} \text{ elements is in class } a]$. This *updates* the estimates of these probabilities over the previous iterations, they are based on better estimates of parameters θ^k. This step is called the *Expectation step*.

Once the $p_{ia}^{(k+1)}$, the updated values of probability distribution of Z_i are computed, we could use those values to get a better estimate of the parameters, say θ^{k+1} . This step is called the *Maximization Step*.

The algorithm goes through a series of expectation and maximization steps until convergence, usually when the distance of $\|\theta^{k-1} - \theta^{k}\|$ is small, and $\|P^{k+1} - P^{k}\|$ is small. Here P^k is an $N \times (r-1)$ matrix where P_{ij}^k is the k^{th} stage estimate of the probability that the i^{th} data point in the table belongs to class a.

The only thing left is the initialization step: what is the initial estimate of parameters $\theta^{(0)}$. In general, any random choice should work.

Here is the general formulation of the algorithm.

Expectation Maximization Algorithm:

Given: • A data set $\mathbf{X}$ with N iid data and d features and a set of latent (unobserved) variables $\mathbf{Z}$
• A joint distribution $f(X, Z \mid \theta)$ with unknown parameters θ

Output: An estimate of θ that maximizes $Q(\theta \mid \mathbf{X}) = \log\Big(\mathbb{E}_Z\big(\text{lik}(\theta \mid \mathbf{X}, \mathbf{Z})\big)\Big)$ and the distribution of $\mathbf{Z}$

Method: 1) Start with an arbitrary grouping of data into k classes
For each group estimate
- θ^0 using the maximum likelihood method
- $\boldsymbol{\tau}^0 = (\tau_, \ldots, \tau_k)$ the proportion of data in each group using ML

3) **While** not converged **do**
Expectation step:
Use the current estimate $\theta^{(k)}$ to estimate the density $p_Z^{(k+1)}$ (e.g. using Bayes rule)
Maximization step:
Use the $p^{(k+1)}$ and update the estimate of parameters $\theta^{(k+1)}$)
If $\|\theta^{(k+1)} - \theta^{(k)}\| < \epsilon$ and $\|p_Z^{(k+1)} - p_Z^{(k)}\| < \epsilon$ converged=**true**
end

Example 3 continued: Mixed Gaussian Model one-dimensional case Suppose we have a data set of incomes, $\mathbf{X} = (30, 290, 90, 110, 55, 57, 370, 115, 135)$. We need to estimate two things. Assume C1 is First Class, C2, Economy Plus, and C3, Economy. Here are a few steps of EM algorithm:

- **Initialize:** Randomly assign each item to one of the three ticket class and then compute the ML mean and variance of each group (using sample mean and sample variance). For example, let's assume that our random selection gave the following assignments to each class: C1=(30,290,90), C2=(110,55,57), and C3=(370,115,135). We can estimate the three pairs of parameters $(\mu_1, \sigma_1), (\mu_2, \sigma_2), (\mu_3, \sigma_3)$, as follows:

- the means are $\mu_1 = 136.67$, $\sigma_1^2 = 12355.56$, $\mu_2 = 74$, $\sigma_2^2 = 648.67$, $\mu_3 = 206.67$, and $\sigma_3^2 = 13405.56$. Also for each group probability of belonging to that group is simply $1/3$. $\tau_1 = \tau_2 = \tau_3 = 1/3$

- The first expectation step is to find for each X_i the probability that it belongs to each ticket class, using current estimates of mean and variance. This can be accomplished by applying Bayes rule, just as we did in Topic 1. So for instance, $X_1 = 30$, and we can see that

$$p_1 = \frac{\phi(30 \mid \mu_1 = 136.67, \sigma_1^2 = 12355.56) \times (1/3)}{A}$$
$$p_2 = \frac{\phi(30 \mid \mu_2 = 74, \sigma_1^2 = 648.67) \times (1/3)}{A}$$
$$p_3 = \frac{\phi(30 \mid \mu_3 = 206.67, \sigma_1^2 = 13405.56) \times (1/3)}{A}$$

  where A is sum of the numerators.

- The first maximization step uses these update probabilities and computes new estimates of mean, variance and class probabilities.

$$\mu_1 = \frac{\sum_{i=1}^{N} X_i p_{1i}}{\sum_i p_{1i}} = 144.90$$
$$\mu_2 = \frac{\sum_{i=1}^{N} X_i p_{2i}}{\sum_i p_{2i}} = 75.12$$
$$\mu_3 = \frac{\sum_{i=1}^{N} X_i p_{3i}}{\sum_i p_{3i}} = 221.78$$

- And similarly we can update variances:

$$\sigma_1^2 = \frac{\sum_{i=1}^{N} (X_i - \mu_1)^2 p_{1i}}{\sum_i p_{1i}} = 10765$$
$$\sigma_2^2 = \frac{\sum_{i=1}^{N} (X_i - \mu_2)^2 p_{2i}}{\sum_i p_{2i}} = 921.7$$
$$\sigma_3^2 = \frac{\sum_{i=1}^{N} (X_i - \mu_3)^2 p_{3i}}{\sum_i p_{3i}} = 15052.9$$

- And update the class probabilities $\boldsymbol{\tau}$:

$$\tau_1 = \frac{\sum_i p_{1i}}{N} = 0.28$$
$$\tau_2 = \frac{\sum_i p_{2i}}{N} = 0.41$$
$$\tau_3 = \frac{\sum_i p_{3i}}{N} = 0.30$$

- the process is repeated until there is not much of improvement on estimated parameters.

In more useful applications of mixed models, the data X might have several features, say d. The joint distribution of each class may be considered to be multi-variate normal with mean vector $\boldsymbol{\mu}_i$ and covariance matrix Σ_i, all unknown. Also class probabilities τ_j are also unknown. We could use the EM algorithm, much the same way as we did for the single feature case, and estimate all these parameters, as well as the probability that each data point belongs to each class.

Also, the EM algorithm can be used for distributions other than the normal. For instance the data can follow the exponential distribution, or the gamma, or beta, or even discrete distributions such as binomial and multinomial distributions.

There is also a Bayesian analog to the EM algorithm. Suppose we have prior knowledge of the unknown parameters expressed in the form of some joint probability distribution. Then the EM algorithm can be adjusted to find the optimal values of the unknown parameters, in the presence of latent variables.

Also check the Wikipedia page expectation maximization