Chapter 1

Clustering and the EM Algorithm

Notation

- N - number of time samples
- M - number of states
- k - index of states
- L - dimension of observation vector

Imagine the following problem. You have measured the height of each plant in a garden. There are N plants, and you know that some have been regularly fertilized, and the remainder have not been fertilized at all. Unfortunately, fertilizing records were lost, and you no longer know which were fertilized, and which were not.

Your measurements, Y_n, of the plant height for the n^{th} plant could have been modeled as Gaussian with mean μ and variance σ^2 if they had all been treated equally; but since they have not, the fertilized plants will, on average, be taller.

Since we have lost the fertilization records, we can model this unknown by a random variable X_n which is 0 if the plant has not been fertilized and 1 if it has. With this assumption, then it is reasonable to model the hight, Y_n, as conditionally Gaussian with distribution $N(\mu_1, \sigma_1)$ if the plant was fertilized, and $N(\mu_0, \sigma_0)$ if it was not. The complete stochastic model for

Figure 1.1: Example of the distribution we might expect in plant height for the two populations. Notice that the two populations create two modes in the distribution.

this problem is then that $\{X_n\}_{n=0}^{N-1}$ are i.i.d. Bernoulli random variables with $P\{X_n = k\} = \pi_k$ for $k \in \{0, 1\}$, and $\{Y_n\}_{n=0}^{N-1}$ are conditionally i.i.d. Gaussian with distribution $N(\mu_{X_k}, \sigma_{X_k})$. The value X_n is sometimes referred to as a *label* because it specifies the population to which Y_n belongs. Notice, that parameters of the conditional distribution, μ_{X_k} and σ_{X_k}, are both dependent on the random variable X_k. This type of stochastic process is sometimes referred to as a *doubly stochastic process* due to this structure. The parameters of this doubly stochastic process (Y_n, X_n) are then given by $\theta = [\mu_0, \sigma_0^2, \mu_1, \sigma_1^2, \pi_0, \pi_1]$ where $\pi_0 + \pi_1 = 1$.

The question then arises of how to estimate the parameter θ of this distribution? This is an important practical problem because we may want to measure the effect of fertilization on plant growth, so we would like the know how much μ_0 and μ_1 differ. Figure 1.1 illustrates the situation. Notice that the two populations create two modes in the distribution of plant height. In order to estimate the mean of each mode, it seems that we would need to know X_n, the label of each plant. However, casual inspection of the distribution of Fig. 1.1 suggests that one might be able to estimate the unknown means, μ_0 and μ_1, by looking at the combined distribution of the two populations.

One possibility for estimating μ_0 and μ_1 is to first estimate the labels X_n. This can be done by applying a threshold at the valley between the two modes, and classifying the value.

$$\hat{X}_n = \begin{cases} 0 & Y_n < \text{threshold} \\ 1 & Y_n \geq \text{threshold} \end{cases}$$

The results of this classification can be more compact represented by the two sets $S_0 = \{n : X_n = 0\}$ and $S_1 = \{n : X_n = 1\}$. With the result of this classification, we can the estimate the two means

$$\begin{aligned} \hat{\mu}_0 &= \frac{1}{|S_0|} \sum_{\{n:X_n=0\}} Y_n \\ \hat{\mu}_1 &= \frac{1}{|S_1|} \sum_{\{n:X_n=1\}} Y_n \ , \end{aligned}$$

where $|S_0|$ and $|S_1|$ denote the number of plants that have been classified as unfertilized and fertilized respectively, and then $\hat{\mu}_0$ and $\hat{\mu}_1$ are the means of the two groups.

While this is an intuitively appealing approach, it has a very serious flaw. Since we have separated the two groups by their hight, it is inevitable that we will measure a larger value for μ_1 than for μ_0. In fact, even when the two means are quite different the resulting estimates of $\hat{\mu}_0$ and $\hat{\mu}_1$ will be systematically wrong no matter how large the value of N. This is much worse than simply being biased, these estimates are inconsistent because the estimated values do not converge to the true parameters as $N \rightarrow \infty$.

Another approach to solving this estimation problem is to attempt to directly estimate the value of the parameter vector $\theta = [\mu_0, \sigma_0^2, \mu_1, \sigma_1^2, \pi_0, \pi_1]$ from the data Y using the ML estimate. Formally, this can be stated as

$$\hat{\theta} = \arg \max_{\theta \in \Theta} \log p(y|\theta) \ .$$

This seems to be a much more promising approach since it is known that the ML estimate is not only consistent, but it is asymptotically efficient, which means that asymptotically achieves the accuracy of the Cramer-Rao bound. [**?**] However, there is still a problem. The distribution $p(y|\theta)$ is not explicitly available for this problem. In fact, it requires the evaluation of a sum that makes direct ML estimate difficult.

$$p(y|\theta) = \sum_x p(y, x|\theta) p(x|\theta)$$

So simple closed form expressions for the ML estimate are no longer possible.

The purpose of the expectation-maximization (EM) algorithm is to provide systematic methodology for estimating parameters such as μ_k when there is missing data, X_n. In particular, the EM algorithm provides a method for determining the ML estimates of the complete parameter vector, θ, which in turn contains the estimates of μ_k. The EM algorithm is quite clever and perhaps surprising, but it results in some very intuitive algorithms when it is properly applied. In fact, the EM algorithm is more than a simple algorithm.

The following sections will lay out the different aspects of the EM algorithm. One way of thinking about the EM algorithm is as a set of inequalities that insure that each step of the EM recursion will improve the likelihood of a parameter estimate. In addition, there are important graphical interpretations of the EM algorithm as a special case of a more general method for optimization using substitute functionals. Finally, we will present a general formulation of the EM algorithm for exponential distributions.

1.1 EM Algorithm Inequalities and Recursions

The EM algorithm is based on the concept that you can separate the log likelihood into the sum of two functions.

$$\log p(y|\theta') = Q(\theta', \theta) + H(\theta', \theta)$$

where

$$\begin{aligned} Q(\theta', \theta) &\stackrel{\triangle}{=} E\left[\log p(y, X|\theta')|Y = y, \theta\right] \\ H(\theta', \theta) &\stackrel{\triangle}{=} -E\left[\log p(X|y, \theta')|Y = y, \theta\right] \end{aligned}$$

To see this result, we have the following sequence of equalities.

$$\begin{aligned} \log p(y|\theta') &= E\left[\log p(y|\theta')|Y = y, \theta\right] \\ &= E\left[\log\left\{\frac{p(y, x|\theta')}{p(x|y, \theta')}\right\}\middle| Y = y, \theta\right] \\ &= E\left[\log p(y, x|\theta')|Y = y, \theta\right] - E\left[\log p(x|y, \theta')|Y = y, \theta\right] \\ &= Q(\theta', \theta) + H(\theta', \theta) \end{aligned}$$

where we use the fact that $p(y, x|\theta') = p(x|y, \theta')p(y|\theta)$ for the second equality. Of course, this result is only valid when $p(y, x|\theta)$ is guaranteed to be strictly positive for all $\theta \in \Theta$.

The key insight is then that the function $H(\theta', \theta)$ takes on its minimum value when $\theta' = \theta$. More precisely, for all $\theta \in \Theta$ and for all $\theta \in \Theta$, we have that

$$H(\theta, \theta) \leq \arg\min_{\theta' \in \Theta} H(\theta', \theta) \tag{1.1}$$

To see that this is true, we have the following set of inequalities

$$\begin{aligned} 0 &= \log\left\{\int p(x|y, \theta')dx\right\} \\ &= \log\left\{\int \frac{p(x|y, \theta')}{p(x|y, \theta)} p(x|y, \theta)dx\right\} \\ &\geq \int \log\left\{\frac{p(x|y, \theta')}{p(x|y, \theta)}\right\} p(x|y, \theta)dx \\ &= \int \log p(x|y, \theta')p(x|y, \theta)dx - \int \log p(x|y, \theta)p(x|y, \theta)dx \\ &= H(\theta', \theta) - H(\theta, \theta) \end{aligned}$$

Using this insight, yields the fundamental result of the EM algorithm. Increasing the value of the function $Q(\theta', \theta)$, we are guaranteed to increase the value of the likelihood. More precisely, for all $\theta \in \Theta$ and for all $\theta \in \Theta$, we have that if $Q(\theta', \theta) > Q(\theta, \theta)$, then we know that $p(y|\theta') > p(y|\theta)$. The proof of this key result is then quite simple.

$$\begin{aligned} \log p(y|\theta') &= Q(\theta', \theta) + H(\theta', \theta) \\ &> Q(\theta, \theta) + H(\theta, \theta) \\ &= \log p(y|\theta) \end{aligned}$$

With is in hand, we have the basic recursion that defines the EM algorithm.

$$\theta^{(k+1)} = \arg \max_{\theta \in \Theta} Q(\theta, \theta^{(k)}) \tag{1.2}$$

where

$$Q(\theta', \theta) = E\left[\log p(y, X|\theta')|Y = y, \theta\right]$$

1.2 Clustering with EM Algorithm

1.3 Convergence and Optimization using Substitute Functions

1.4 General Methods for EM Updates with Exponential Distributions

While the calculation of the Q function is typically complex, it is clear that there is a general pattern to the result. For example, in the clustering example of Section 1.2 the final EM update equations appear much like the conventions ML estimates, except that the means are weighted by the probability that an sample is from the particular class. This pattern turns out to have a underlying explanation which can be used to make derivation of the EM updates much simpler. In fact, it turns out that for all exponential distributions, the form of the EM update is quite simple.

In the following chapter, we introduce the concepts necessary to understand this simplification, and we show how to derive EM updates for any distribution with exponential form.

1.4.1 Exponential Distributions and their Natural Sufficient Statistics

In order to derive the simplified form of the EM update, we first must introduce the concept of an exponential distribution and its natural sufficient statistics. We base our definitions on a N dimensional random vector Y with density function $p(y|\theta)$ with parameter vector $\theta \in \Theta$. Then we have the following two important definitions.

Definition: A *statistic* is any function $T(Y)$ of the data Y.

Definition: We say that a statistic $T(Y)$ is a *sufficient statistic* for θ if there exist functions $g(\cdot, \cdot)$ and $h(\cdot)$ such that

$$p(y|\theta) = h(y)\, g(T(y), \theta) \tag{1.3}$$

for all $y \in I\!R^N$ and $\theta \in \Theta$.

Inituitively, a sufficient statistic distills all the information from the data, Y, necessary to estimate the parameter θ. For example, the ML estimator of θ must be a function of the sufficient statistic $T(Y)$. To see this, notice that

$$\begin{aligned}
\hat{\theta}_{ML} &= \arg\max_{\theta\in\Theta} \log p(y|\theta) \\
&= \arg\max_{\theta\in\Theta} \{\log h(y) + \log g(T(y), \theta)\} \\
&= \arg\max_{\theta\in\Theta} \log g(T(y), \theta) \\
&= f(T(y))
\end{aligned}$$

for some function $f(\cdot)$.

Many commonly used distributions such as Gaussian, exponential, Poisson, Bernoulli, and binomial have a structure which makes them particularly useful. These distributions are known as exponential families and have the following special property.

Definition: A family of density functions $p(y|\theta)$ for $y \in R^N$ and $\theta \in \Theta$ is said to be a *k-parameter exponential family* if there exist functions $g(\theta) \in I\!R^k$, $s(y)$, $d(\theta)$ and statistic $T(y) \in I\!R^k$ such that

$$p(y|\theta) = \exp\{< g(\theta), T(y) > + d(\theta) + s(y)\} \tag{1.4}$$

for all $y \in \mathbb{R}^N$ and $\theta \in \Theta$ where $< \cdot, \cdot >$ denotes the inner product. We refer to $T(y)$ as the *natural sufficient statistic* or *natural statistic* for the exponential distribution.

Exponential distributions are extremely valuable because the log of its density forms an inner product that is easily manipulated when computing ML parameter estimates.

Here are some examples of exponential distributions and their natural sufficient statistics.

Example 1.4.1: Let $\{Y_n\}_{n=0}^{N-1}$ be i.i.d. random variables with distribution $N(\mu, 1)$ and parameter $\theta = [\mu]$. Then $p(y|\theta)$ is an exponential distribution with natural sufficient statistic

$$t_1 = \sum_{n=0}^{N-1} Y_n \ ,$$

and the ML estimate of θ is given by

$$\hat{\mu} = \frac{t_1}{N} \ .$$

The proof of this fact is given in Appendix A.1.

So here we see that the sample average, which is the ML estimate of the mean for i.i.d. Gaussian random variables, can be expressed as t_1/N. We can extend this same structure to the case when both the mean and variance are unknown.

Example 1.4.2: Let $\{Y_n\}_{n=0}^{N-1}$ be i.i.d. random variables with distribution $N(\mu, \sigma^2)$ and parameter $\theta = [\mu, \sigma^2]$. Then $p(y|\theta)$ is an exponential distribution with natural sufficient statistics

$$\begin{aligned} t_1 &= \sum_{n=0}^{N-1} Y_n \\ t_2 &= \sum_{n=0}^{N-1} Y_n^2 \ , \end{aligned}$$

and the ML estimate of θ is given by

$$\begin{aligned} \hat{\mu} &= \frac{t_1}{N} \\ \hat{\sigma}^2 &= \frac{t_2}{N} - \left(\frac{t_1}{N}\right)^2 \ . \end{aligned}$$

The proof of this fact is given in Appendix A.2.

The example can be further generalized by allowing the observations, Y_n, to be multivariate Gaussian vectors. In this case, the parameters are the mean and covariance of the Gaussian vector, and the ML estimates are given by the sample average and the sample covariance.

Example 1.4.3: Let $\{Y_n\}_{n=0}^{N-1}$ be i.i.d. random vectors of dimension L with distribution $N(\mu, R)$ and parameter $\theta = [\mu, R]$. Then $p(y|\theta)$ is an exponential distribution with natural sufficient statistics

$$\begin{aligned} t_1 &= \sum_{n=0}^{N-1} Y_n \\ t_2 &= \sum_{n=0}^{N-1} Y_n Y_n^t \ , \end{aligned}$$

and the ML estimate of θ is given by

$$\begin{aligned} \hat{\mu} &= \frac{t_1}{N} \\ \hat{R} &= \frac{t_2}{N} - \frac{t_1 t_1^t}{N^2} \ . \end{aligned}$$

The proof of this fact is given in Appendix A.3.

We will also be very interested in the *Bernoulli distribution* which is also exponential. Bernoulli random variables can be thought of as the outcome of a coin flip, for an unfair coin where the probability of "heads" is π_1, and the probability of "tails" is $\pi_0 = 1 - \pi_1$.

Example 1.4.4: Let $\{X_n\}_{n=0}^{N-1}$ be i.i.d. random variables with the Bernoulli distribution given by $P\{X_n = 0\} = \pi_0$ and $P\{X_n = 1\} = \pi_1$ and parameter $\theta = [\pi_0, \pi_1]$. Then $p(x|\theta)$ is an exponential distribution with natural sufficient statistics

$$\begin{aligned} N_0 &= \sum_{n=0}^{N-1} \delta(X_n) \\ N_1 &= \sum_{n=0}^{N-1} \delta(X_n - 1) \end{aligned}$$

where $\delta(\cdot)$ is a Kroniker delta function, and the ML estimate of θ is given by

$$\begin{aligned}\hat{\pi}_0 &= \frac{N_0}{N} \\ \hat{\pi}_1 &= \frac{N_1}{N}\end{aligned}$$

The proof of this fact is given in Appendix A.4.

For clustering problem of Section 1.2, we had an observed vector of N conditionally Gaussian random variables Y_n, and a corresponding set of binary random variables which determined their distribution. More specifically, Y_n was conditionally Gaussian with mean and variance given by μ_{x_n} and $\sigma^2_{x_n}$, and X_n are i.i.d. Bernoulli random variables with parameters π_0 and π_1. In this case, the joint distribution of both Y and X is exponential with parameter $\theta = [\mu_0, \sigma_0^2, \mu_1, \sigma_1^2, \pi_0, \pi_1]$.

Example 1.4.5: Let $\{X_n\}_{n=0}^{N-1}$ be i.i.d. Bernoulli random variables with $P\{X_n = 0\} = \pi_0$ and $P\{X_n = 1\} = \pi_1 = 1 - \pi_0$. Let $\{Y_n\}_{n=0}^{N-1}$ be conditionally i.i.d. Gaussian random variables with conditional mean and variance given by μ_{x_n} and $\sigma^2_{x_n}$ respectively. Then $p(y, x|\theta)$ is an exponential distribution with parameter $\theta = [\mu_0, \sigma_0^2, \mu_1, \sigma_1^2, \pi_0, \pi_1]$ and natural sufficient statistics

$$N_k = \sum_{n=0}^{N-1} \delta(x_n - k) \tag{1.5}$$

$$t_{1,k} = \sum_{n=0}^{N-1} y_n \delta(x_n - k) \tag{1.6}$$

$$t_{2,k} = \sum_{n=0}^{N-1} y_n^2 \delta(x_n - k) \tag{1.7}$$

where $k \in \{0, 1\}$ and the ML estimate of θ is given by

$$\hat{\mu}_k = \frac{t_{1,k}}{N_k} \tag{1.8}$$

$$\hat{\sigma}_k^2 = \frac{t_{2,k}}{N_k} - \left(\frac{t_{1,k}}{N_k}\right)^2 \tag{1.9}$$

$$\hat{\pi}_k = \frac{N_k}{N} \ . \tag{1.10}$$

The proof of this fact is given in Appendix A.5.

Of course the problem with the ML estimates of (1.8), (1.9), and (1.10) is that we we may not know the labels X_n. This is the so called incomplete data problem that the EM algorithm addresses. In the next section, we will see how the EM algorithm can be simply derived for any such exponential distribution.

1.4.2 General Formulation of EM Update

One reason that the EM algorithm is so useful is that for many practical situations the distributions are exponential, and in this case the EM updates have a particularly simple form. Let Y is the observed or incomplete data and let X be the unobserved data, and assume that the joint density of (Y, X) is from and exponential family with parameter vector θ. Then we know that

$$p(y, x|\theta) = \exp\{< g(\theta), T(y, x) > +d(\theta) + s(y, x)\}$$

for some natural sufficient statistic $T(y, x)$. Assuming the ML estimate of θ exists, then it is given by

$$\begin{aligned} \theta_{ML} &= \arg\max_{\theta\in\Theta} \{< g(\theta), T(y, x) > +d(\theta)\} && (1.11) \\ &= f(T(y, x)) && (1.12) \end{aligned}$$

where $f(\cdot)$ is some function of $T(y, x)$.

Recalling the form of the Q function, we have

$$Q(\theta', \theta) = E\left[\log p(y, X|\theta')|Y = y, \theta\right]$$

where Y is the observed data and X is the unknown data. Using the assumed structure of the exponential distribution, we have that

$$\begin{aligned} Q(\theta', \theta) &= E\left[\log p(y, X|\theta')|Y = y, \theta\right] \\ &= E\left[< g(\theta'), T(y, X) > +d(\theta') + s(y, X)|Y = y, \theta\right] \\ &= < g(\theta'), \bar{T}(y) > +d(\theta') + constant \end{aligned}$$

where

$$\bar{T}(y) = E\left[T(y, X)|Y = y, \theta\right]$$

is the conditional expectation of the sufficient statistic $T(y, x)$, and *constant* is a constant which does not depend on θ'. Since our objective is to maximize

Q with respect to θ', this constant can be dropped. A single update of the EM algorithm is then given by the recursion

$$\begin{aligned} \theta'' &= \arg\max_{\theta' \in \Theta} Q(\theta', \theta) \\ &= \arg\max_{\theta' \in \Theta} \left\{ < g(\theta'), \bar{T}(y) > + d(\theta') \right\} \\ &= f(\bar{T}(y)) \end{aligned} \tag{1.13}$$

Intuitively, we see that the EM update has the same form as the computation of the ML estimate, but with the expected value of the statistic, $\bar{T}$, replacing the actual statistic, T.

To see how useful this result can be, we can use it to easily derive the EM update parameters so for the clustering Example 1.4.5.

Example 1.4.6: Let $\{X_n\}_{n=0}^{N-1}$ be i.i.d. Bernoulli random variables with parameters $[\pi_0, \pi_1]$, and let $\{Y_n\}_{n=0}^{N-1}$ be conditionally i.i.d. Gaussian random variables with conditional mean and variance given by μ_{x_n} and $\sigma_{x_n}^2$ respectively. We would like derive the EM updates for the parameters $\theta = [\mu_0, \sigma_0^2, \mu_1, \sigma_1^2, \pi_0, \pi_1]$ assuming that Y is observed, but X is not. We know from Example 1.4.5 that $p(y, x|\theta)$ is an exponential distribution with natural sufficient statistics given by equations (1.5), (1.6), and (1.7), and ML parameter estimates given by (1.8), (1.9), and (1.10).

In order to derive the EM update, we only need to replace the sufficient statistics in the ML estimate by their expected values. So to compute the EM update of the parameter θ, we first compute the conditional expectation of the statistics in the E-step

$$\begin{aligned} \bar{N}_k &= \sum_{n=0}^{N-1} P\{X_n = k | Y = y, \theta\} \\ \bar{t}_{1,k} &= \sum_{n=0}^{N-1} y_n P\{X_n = k | Y = y, \theta\} \\ \bar{t}_{2,k} &= \sum_{n=0}^{N-1} y_n^2 P\{X_n = k | Y = y, \theta\} \ . \end{aligned}$$

and then then we use these new statistics to computed updated values of the parameters in the M-step.

$$\hat{\mu}_k \leftarrow \frac{\bar{t}_{1,k}}{\bar{N}_k} \tag{1.14}$$

$$\hat{\sigma}_k^2 \leftarrow \frac{\bar{t}_{2,k}}{\bar{N}_k} - \left(\frac{\bar{t}_{1,k}}{\bar{N}_k}\right)^2 \tag{1.15}$$

$$\hat{\pi}_k \leftarrow \frac{\bar{N}_k}{N} \tag{1.16}$$

By repeating this process, we increase the likelihood of the observations. Assuming that the likelihood has a minimum [1] and that one is not trapped in a local minimum, then repeated application of the EM iterations will converge to the ML estimate of θ.

[1] For the case of a Gaussian mixture with no lower bound on σ_k^2, the likelihood is not bounded. However, in practice a local minimum of the likelihood usually provides a good estimate of the parameters.

Appendix A

EM Algorithm Derivations

This appendix contains derivations of results from Chapter 1.

A.1 Example 1.4.1 Derivation

Let $\{Y_n\}_{n=0}^{N-1}$ be i.i.d. random variables with distribution $N(\mu, 1)$. Define the following statistic corresponding to the sample mean of the random variables.

$$t_1 = \sum_{n=0}^{N-1} y_n$$

By writing the density function for the sequence Y as

$$\begin{aligned}
p(y|\mu) &= \prod_{n=0}^{N-1} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(y_n-\mu)^2\right\} \\
&= \frac{1}{\left(\sqrt{2\pi}\right)^N} \exp\left\{-\frac{1}{2}\sum_{n=0}^{N-1}(y_n-\mu)^2\right\} \\
&= \frac{1}{\left(\sqrt{2\pi}\right)^N} \exp\left\{-\frac{1}{2}\sum_{n=0}^{N-1}(y_n^2-2y_n\mu+\mu^2)\right\} \\
&= \frac{1}{\left(\sqrt{2\pi}\right)^N} \exp\left\{-\frac{1}{2}\sum_{n=0}^{N-1} y_n^2\right\} \exp\left\{\frac{1}{2}(2t_1\mu - N\mu^2)\right\} \\
&= \frac{1}{\left(\sqrt{2\pi}\right)^N} \exp\left\{-\frac{1}{2}\left(\sum_{n=0}^{N-1} y_n^2 + t_1^2/N\right)\right\} \exp\left\{-\frac{N}{2}(t_1/N-\mu)^2\right\} \ ,
\end{aligned}$$

where we can see that it has the form of equation (1.3). Therefore, t_1 is a sufficient statistic for the parameter μ. Computing the ML estimate yeilds

the following.

$$\begin{aligned}
\hat{\mu}_{ML} &= \arg\max_{\mu} \log p(y|\mu) \\
&= \arg\max_{\mu} \left\{ -\frac{N}{2}(t_1/N - \mu)^2 \right\} \\
&= \arg\min_{\mu} (t_1/N - \mu)^2 \\
&= \frac{t_1}{N}
\end{aligned}$$

A.2 Example 1.4.2 Derivation

Let $\{Y_n\}_{n=0}^{N-1}$ be i.i.d. random variables with distribution $N(\mu, \sigma^2)$. Define the following statistics corresponding to the sample mean and variance of the random variables.

$$\begin{aligned}
t_1 &= \sum_{n=0}^{N-1} y_n \\
t_2 &= \sum_{n=0}^{N-1} y_n^2
\end{aligned}$$

Then we may write the density function for Y in the following form.

$$\begin{aligned}
p(y|\mu,\sigma^2) &= \prod_{n=0}^{N-1} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2}(y_n - \mu)^2 \right\} \\
&= \frac{1}{\left(\sqrt{2\pi\sigma^2}\right)^N} \exp\left\{ -\sum_{n=0}^{N-1} \frac{1}{2\sigma^2}(y_n - \mu)^2 \right\} \\
&= \frac{1}{\left(\sqrt{2\pi\sigma^2}\right)^N} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{n=0}^{N-1} (y_n^2 - 2y_n\mu + \mu^2) \right\} \\
&= \frac{1}{\left(\sqrt{2\pi\sigma^2}\right)^N} \exp\left\{ -\frac{1}{2\sigma^2} t_2 + 2\frac{\mu}{2\sigma^2} t_1 - \frac{N}{2\sigma^2}\mu^2 \right\} \\
&= \exp\left\{ -\frac{1}{2\sigma^2} t_2 + 2\frac{\mu}{2\sigma^2} t_1 - \frac{N}{2\sigma^2}\mu^2 - \frac{N}{2}\log(2\pi\sigma^2) \right\} \\
&= \exp\left\{ \left[\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2} \right] \begin{bmatrix} t_1 \\ t_2 \end{bmatrix} - \frac{N}{2\sigma^2}\mu^2 - \frac{N}{2}\log(2\pi\sigma^2) \right\}
\end{aligned}$$

Using the following definitions

$$\begin{aligned} g(\theta) &= \left[\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2}\right] \\ T(y) &= \begin{bmatrix} t_1 \\ t_2 \end{bmatrix} \\ d(\theta) &= -\frac{N}{2\sigma^2}\mu^2 - \frac{N}{2}\log(2\pi\sigma^2) \\ s(y) &= 0 \end{aligned}$$

we can see that $p(y|\mu, \sigma^2)$ has the form of equation (1.4) with suffient statistic $T(y)$. With some calculations it may be easily shown that the ML estimates of μ and σ^2 are given by

$$\begin{aligned} \hat{\mu}_{ML} &= \frac{t_1}{N} \\ \hat{\sigma}^2_{ML} &= \frac{t_2}{N} - \left(\frac{t_1}{N}\right)^2 \end{aligned}$$

A.3 Example 1.4.3 Derivation

A.4 Example 1.4.4 Derivation

A.5 Example 1.4.5 Derivation

Example 3: Let $\{X_n\}_{n=0}^{N-1}$ be i.i.d. random variables with $P\{X_n = 0\} = \pi_0$ and $P\{X_n = 1\} = \pi_1 = 1 - \pi_0$. Let $\{Y_n\}_{n=0}^{N-1}$ be conditionally i.i.d random variables given X, and let the conditional distribution of Y_n given X_n be Gaussian $N(\mu_{X_n}, \sigma^2_{X_k})$ where μ_0, μ_1, σ_0, and σ_1 are parameters of the distribution. Then the complete set of parameters for the density of (Y, X) are given by

$$\theta = [\mu_0, \mu_1, \sigma_0, \sigma_1, \pi_0] \ .$$

Define the statistics

$$N_k = \sum_{n=0}^{N-1} \delta(x_n - k)$$

$$
\begin{aligned}
t_{1,k} &= \sum_{n=0}^{N-1} y_n \delta(x_n - k) \\
t_{2,k} &= \sum_{n=0}^{N-1} y_n^2 \delta(x_n - k)
\end{aligned}
$$

where $k \in \{0, 1\}$ and $\delta(\cdot)$ is a Kroniker delta function. We know that if both Y and X are known then the ML estimates are given by

$$\hat{\mu}_k = \frac{t_{1,k}}{N_k} \tag{A.1}$$

$$\hat{\sigma}_k^2 = \frac{t_{2,k}}{N_k} - \left(\frac{t_{1,k}}{N_k}\right)^2 \tag{A.2}$$

$$\hat{\pi}_k = \frac{N_k}{N} \, . \tag{A.3}$$

We can express the density function for $p(y|x, \theta)$ by starting with the expressions derived in example 2 for each of the two classes corresponding to $X_n = 0$ and $X_n = 1$.

$$
\begin{aligned}
p(y|x,\theta) &= \prod_{k=0}^{1} \exp\left\{ \left[\frac{\mu_k}{\sigma_k^2}, -\frac{1}{2\sigma_k^2}\right] \begin{bmatrix} t_{1,k} \\ t_{2,k} \end{bmatrix} - \frac{N_k}{2\sigma_k^2}\mu_k^2 - \frac{N_k}{2}\log(2\pi\sigma_k^2) \right\} \\
&= \prod_{k=0}^{1} \exp\left\{ \left[\frac{\mu_k}{\sigma_k^2}, -\frac{1}{2\sigma_k^2}, -\frac{\mu_k^2}{2\sigma_k^2}\right] \begin{bmatrix} t_{1,k} \\ t_{2,k} \\ N_k \end{bmatrix} - \frac{N_k}{2}\log(2\pi\sigma_k^2) \right\} \\
&= \prod_{k=0}^{1} \exp\left\{ \left[\frac{\mu_k}{\sigma_k^2}, -\frac{1}{2\sigma_k^2}, -\frac{\mu_k^2}{2\sigma_k^2} - \frac{1}{2}\log(2\pi\sigma_k^2)\right] \begin{bmatrix} t_{1,k} \\ t_{2,k} \\ N_k \end{bmatrix} \right\} \\
&= \exp\left\{ \sum_{k=0}^{1} \left[\frac{\mu_k}{\sigma_k^2}, -\frac{1}{2\sigma_k^2}, -\frac{\mu_k^2}{2\sigma_k^2} - \frac{1}{2}\log(2\pi\sigma_k^2)\right] \begin{bmatrix} t_{1,k} \\ t_{2,k} \\ N_k \end{bmatrix} \right\}
\end{aligned}
$$

The distribution for X also has exponential form with

$$
\begin{aligned}
p(x|\theta) &= \pi_0^{N_0} \pi_1^{N_1} \\
&= \exp\left\{ \sum_{k=0}^{1} N_k \log \pi_k \right\}
\end{aligned}
$$

This yields the joint density for (Y, X) with the following form.

$$\begin{aligned} p(y, x|\theta) &= p(y|x, \theta)p(x|\theta) \\ &= \exp\left\{ \sum_{k=0}^{1} \left[\frac{\mu_k}{\sigma_k^2}, -\frac{1}{2\sigma_k^2}, -\frac{\mu_k^2}{2\sigma_k^2} - \frac{1}{2}\log(2\pi\sigma_k^2) + \log \pi_k \right] \begin{bmatrix} t_{1,k} \\ t_{2,k} \\ N_k \end{bmatrix} \right\} \end{aligned}$$