Assignment 1

Elements of Machine Learning

Department of Computer Science
University of Copenhagen

Casper Lisager Frandsen <fsn483@alumni.ku.dk>,
Marc Pedersen <wfj327>

Version 1
Due: February 16th, 23:55

Contents

1. Book Exercises

(a) 8.3

First we will show that $p(a)$ and $p(b)$ are marginally dependent by showing that $p(a)p(b) \neq p(a,b)$. We calculate each of these as the sum of the probability of all rows where the corresponding random variable $= 1$. With this we get the following:

$$\begin{aligned} p(a) &= 0.400 \\ p(b) &= 0.408 \\ p(c) &= 0.520 \end{aligned}$$

We can also easily calculate combined probabilities, by summing the probability of all rows where all the random variables are $= 1$:

$$\begin{aligned} p(a,b) &= 0.144 \\ p(a,c) &= 0.160 \\ p(b,c) &= 0.312 \end{aligned}$$

Now we can evaluate $p(a,b) \neq p(a)p(b)$:

$$\begin{aligned} p(a,b) &= 0.144 \\ p(a)p(b) &= 0.400 \cdot 0.408 = 0.1632 \\ 0.144 \neq 0.1632 &\implies p(a,b) \neq p(a)p(b) \end{aligned}$$

Thus, we have shown that they are marginally dependent. Now, to show that they become independent when conditioned on c, we have to show that $p(a,b|c) = p(a|c)p(b|c)$. We know that $p(a|b) = \dfrac{p(a,b)}{p(b)}$, so we calculate the following:

$$\begin{aligned} p(a,b|c) &= \frac{p(a,b,c)}{p(c)} = \frac{0.096}{0.52} = 0.1\overline{846153} \\ p(a|c) &= \frac{p(a,c)}{p(c)} = \frac{0.16}{0.52} = 0.\overline{307692} \\ p(b|c) &= \frac{p(b,c)}{p(c)} = \frac{0.312}{0.52} = 0.6 \end{aligned}$$

Now we simply calculate and compare:

$$\begin{aligned} p(a,b|c) &= 0.1\overline{846153} \\ p(a|c)p(b|c) &= 0.\overline{307692} \cdot 0.6 = 0.1\overline{846153} \end{aligned}$$

These are obviously equal, so we have shown that $p(a,b|c) = p(a|c)p(b|c)$ for $c = 1$. To show the same for $c = 0$, we follow pretty much the same procedure. If a probability is not named below, it is the same as for the calculations above:

$$p(c) = 0.48$$

$$
\begin{aligned}
p(a,c) &= 0.240 \\
p(b,c) &= 0.096 \\
p(a,b,c) &= 0.048 \\
p(a|c) &= \frac{0.240}{0.48} = 0.5 \\
p(b|c) &= \frac{0.096}{0.48} = 0.2
\end{aligned}
$$

Now we can again calculate and compare:

$$
\begin{aligned}
p(a|c)p(b|c) &= 0.5 \cdot 0.2 = 0.1 \\
p(a,b|c) &= \frac{0.048}{0.48} = 0.1
\end{aligned}
$$

These are also obviously equal, and we have thus also shown that $p(a|c)p(b|c) = p(a,b|c)$ is true for $c = 0$.

(b) 8.4

We have the following numbers already from the exercise above:

$$
\begin{aligned}
p(a,b,c) &= 0.096 \\
p(a) &= 0.4 \\
p(a,c) &= 0.16 \\
p(c) &= 0.52 \\
p(b,c) &= 0.312
\end{aligned}
$$

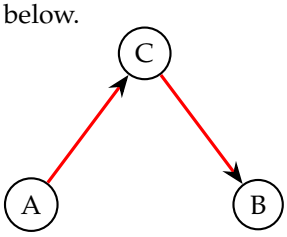
We can calculate the following:

$$
\begin{aligned}
p(c|a) &= \frac{p(a,c)}{p(a)} = \frac{0.16}{0.4} = 0.4 \\
p(b|c) &= \frac{p(b,c)}{p(c)} = \frac{0.312}{0.52} = 0.6
\end{aligned}
$$

Now we simply calculate $p(a)p(c|a)p(b|c)$ and $p(a,b,c)$:

$$
\begin{aligned}
p(a)p(c|a)p(b|c) &= 0.4 \cdot 0.6 \cdot 0.4 = 0.096 \\
p(a,b,c) &= 0.096
\end{aligned}
$$

Thus we have shown that $p(a)p(c|a)p(b|c) = p(a,b,c)$. The graph is shown below.



(c) 8.10

For part 1, we can intuitively see that it is true, using d-separation. To check whether two nodes are independent, all paths between them have to be blocked. There is a single path from a to b, through c. However, we can see that the path is blocked at the c node, because the arrows meet head to head, on a node that is obviously not $\in \emptyset$. We can also show by writing the probabilities of the graph out:

$$p(a,b,c,d) = p(d|c)p(c|a,b)p(a)p(b)$$

We can marginalise the components of c and d:

$$\begin{aligned}
p(a,b,c,d) &= p(d|c)p(c|a,b)p(a)p(b) \\
p(a,b) &= \sum_d \sum_c p(d|c)p(c|a,b)p(a)p(b) \\
&= \sum_d p(d|c) \sum_c p(c|a,b)p(a)p(b) \\
&= \underbrace{\sum_d p(d|c)}_{=1} \underbrace{\sum_c p(c|a,b)}_{=1} p(a)p(b) \\
p(a,b) &= p(a)p(b) \implies a \perp\!\!\!\perp b|\emptyset
\end{aligned}$$

Thus, we have proven that a and b are independent without any observed values. For the second part we have to prove the opposite, but with the node d observed. That is, we have to prove that $a \not\perp\!\!\!\perp b|d$. Intuitively we can see that this is true, using d-separation again. The only node that connects the two is stil c. However, because c has a descendant d, that is in the given set, c blocks the path, and thus $a \not\perp\!\!\!\perp b|d$. Using the probabilities, we can write it out the following way:

$$p(a,b,c|d) = \frac{p(d|c)p(c|a,b)p(a)p(b)}{p(d)}$$

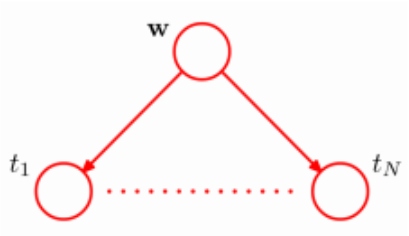
We can marginalise the components of c:

$$\begin{aligned}
p(a,b,c|d) &= \frac{p(d|c)p(c|a,b)p(a)p(b)}{p(d)} \\
p(a,b|d) &= \frac{\sum_c p(d|c)p(c|a,b)p(a)p(b)}{p(d)} \\
&= \frac{p(d|c)p(a)p(b)\overbrace{\sum_c p(c|a,b)}^{=1}}{p(d)} \\
&= \frac{p(d|c)p(a)p(b)}{p(d)} \\
&= \frac{p(d|c)p(a)p(b)}{p(d)} \neq p(a)p(b) \implies a \not\perp\!\!\!\perp b|d
\end{aligned}$$

(d) 9.5

As per the example from the lecture slides, pictured below, any arrows into a "plate" are repeated for all N elements on the plate. Because all the arrows meet tail-to-tail at the π, μ and Σ nodes, all paths are blocked between $z_1, ..., z_n$. According to the rules of D-separation, they must then all be independent. Because $p(Z|X, \mu, \Sigma, \pi)$ is the combined probability of all $\in Z$, the general independence rule that $p(a, b) = p(a)p(b)$ must also apply, and thus we have shown that:

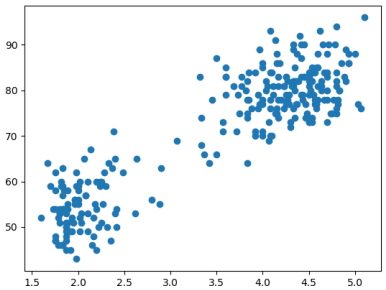
$$p(Z|X, \mu, \Sigma, \pi) = \prod_{n=1}^{N} p(z_n|x_n, \mu, \Sigma, \pi)$$



2. Old Faithful exercises

(a)

The data has been plotted from a .txt file, with line numbers as the first column.

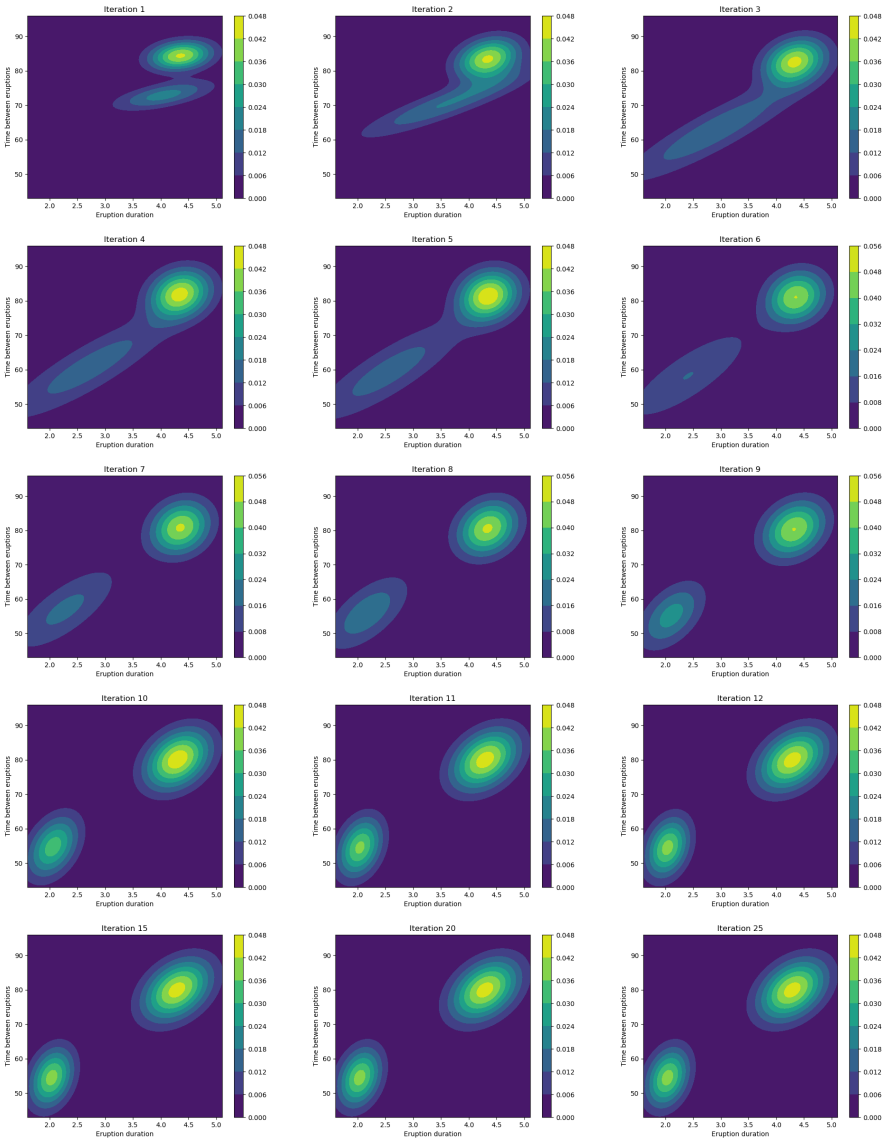


(b)

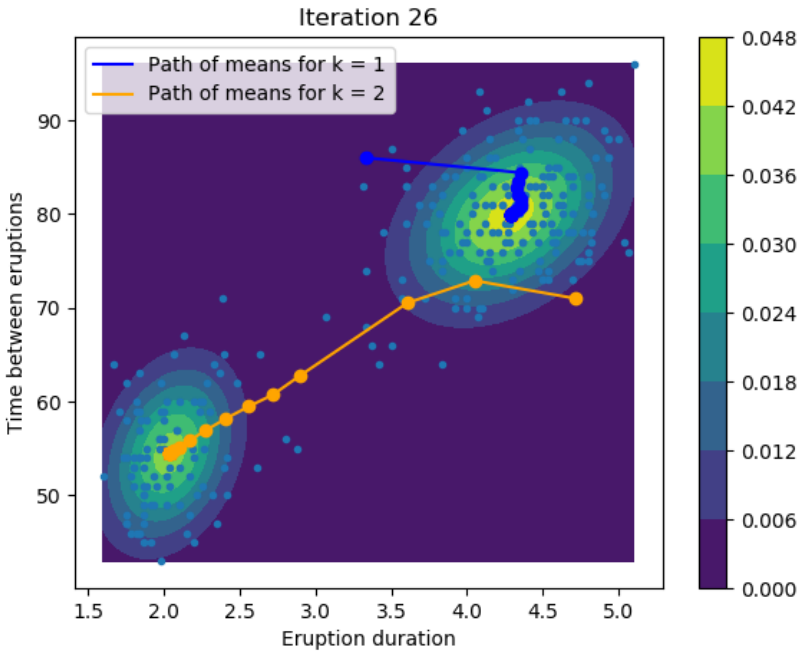
The EM algorithm has been implemented in the attached "old_faithful.py" file. The starting means have been chosen uniformly from the x and y coordinates in the given data. The starting covariance matrices have been generated as k stacked identity matrices.

(c)

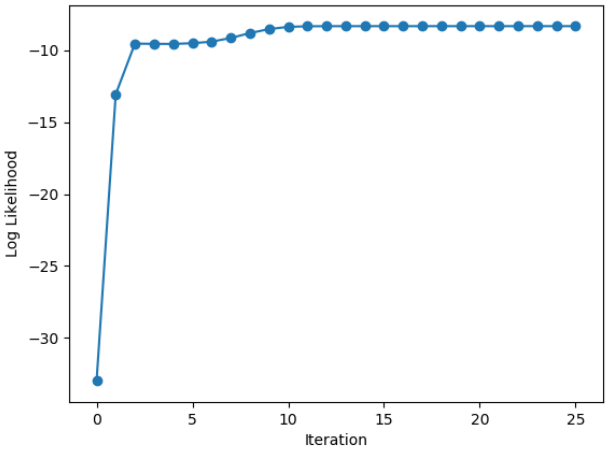
We generate three types of graphs from the attached code. The first series, shown below show the evolution of the probability distribution, as the means and covariance matrices are updated.



The next graph shows the exact positions of the means, as they are updated. This is overlaid on top of the data, and the contour plot. This helps visualise that the generated contour plots are relatively accurate with regards to the data.



The final graph shows the progress of the log likelihood function, over the different iterations. It makes sense intuitively, that it increases so significantly in the beginning if you also look at the previous graphs. The means and covariance matrices change significantly in the first few iterations, but rapidly converge to a point where the changes are minimal, and barely perceptible.



(d)

Using more than two mixtures in this case significantly increase the amount of iterations the algorithm has to go through before it converges. For $k = 2$,

the number of iterations it takes to converge hovers between $10 - 20$ usually, with unlucky starting means going as high as 40-odd iterations. The graphs shown below are from a run with $k = 3$, and went through 153 iterations before converging. It is also worth noting that the results don't seem to represent the data as well as with $k = 2$. This is likely due to the algorithm finding a local maximum, instead of a global maximum. An example of this can be seen in the contour plot below, which has significantly higher peaks that don't seem to align with the data. The graphs below are of the same type as those described in the previous exercise.