Solutions to exercises: week 4

Exercise 4.1
(a) It's not 99%. It actually depends on the prior probability of people knowing all there is to know about machine learning. So let's say that is one in 1,000,000. So, $P(K) = \frac{1}{1{,}000{,}00}$. We are interested in $P(K|T)$, which is $P(T|K)P(K)/P(T) \approx 0.001$.

Exercise 4.2
(a) This is $\prod p(x_i|\theta)$.
(b) This is the likelihood function as used in the maximum likelihood estimation.

Exercise 4.3
(a) We have $p(\theta|x_1, \ldots, x_N) = p(\theta) \prod p(x_i|\theta)/(\prod p(x_i))$.
(b) As the data does not change $\prod p(x_i)$ remains constant. That is, from the viewpoint of estimating θ, we have $p(\theta|x_1, \ldots, x_N) is proportional to p(\theta) \prod p(x_i|\theta)$.
(c) We get the ML estimate back.
(d) From the full distribution $p(x, \theta) = p(x|\theta)p(\theta)$, we get to $p(x)$ by integrating or summing out θ.

Exercise 4.4
(a) The posterior is a uniform distribution, so no unique maximum
(b) It does satisfy positivity on $[0, 1]$. The indefinite integral is $\log(q) - \log(1 - q) + C$. This does not integrate to 1 on $[0, 1]$.

Exercise 4.5
(a) Priors are denoted p_1, p_2, means are m_1, m_2, standard deviation is s. w is the set of all parameters $\{p_1, p_2, m_1, m_2, s\}$.
Full model is

$$p(x, y|w) = \begin{cases} p(y = 1|w)p(x|y = 1, w) = p_1 \frac{1}{\sqrt{2\pi}s} \exp(-\frac{(x-m_1)^2}{2s^2}) & \text{if } y = 1 \\ p(y = 2|w)p(x|y = 2, w) = p_2 \frac{1}{\sqrt{2\pi}s} \exp(-\frac{(x-m_2)^2}{2s^2}) & \text{if } y = 2 \end{cases}$$

(b) Let us assume we have N_1 samples from class 1 and $N_2 = N - N_1$ from class 2. The empirical logarithmic loss means we express the loss in terms of our observed data. We just take the empirical average of our previously defined log loss. As a result, the quantities $p(x, y = 1)$ and $p(x, y = 2)$, the true distributions which we typically don't know, will drop out of the equation. [Basically we use that $\int p(x)g(x)dx \approx \frac{1}{N} \sum_i g(x_i)$.]

$$\tfrac{1}{N}[-\sum_{i=1}^{N_1}(\log[p_1 \tfrac{1}{\sqrt{2\pi}s}] - \tfrac{(x_i-m_1)^2}{2s^2}) - \sum_{i=1}^{N_2}(\log[p_2 \tfrac{1}{\sqrt{2\pi}s}] - \tfrac{(x_i-m_2)^2}{2s^2})]$$

Sloppy notation! It is assumed here that the sums first sum runs only over samples from class 1 and the second sum only runs over samples from class 2.
(c) The estimates for all free and unknown parameters $w = \{p_1, p_2, m_1, m_2, s\}$ should be maxima of the likelihood, or minima of our log loss. One way or the other, finding this optimum is possible by finding that point in parameter space where the gradient is zero. That is, we take the five derivatives, for every parameter one, of the log loss, equate those to zero, and solve this set of equations.
Here's one example: We want to solve

$$\frac{d}{dm_1} \tfrac{1}{N}[-\sum_{i=1}^{N_1}(\log[p_1 \tfrac{1}{\sqrt{2\pi}s}] - \tfrac{(x_i-m_1)^2}{2s^2}) - \sum_{i=1}^{N_2}(\log[p_2 \tfrac{1}{\sqrt{2\pi}s}] - \tfrac{(x_i-m_2)^2}{2s^2})] = 0$$

Taking the derivative to m_1, and ignoring various constant multiplicative factors to simplify the whole expression, it basically means we have to solve

$$\sum_{i=1}^{N_1} \frac{(x_i - m_1)}{2s^2} = 0$$

Now solving for m_1, we get as the [maximum likelihood] estimate m_1^{ML}:

$$m_1^{\mathrm{ML}} = \frac{1}{N_1} \sum_{i=1}^{N_1} x_i$$

This expression should look familiar…
(d) Posterior of w means we consider $P(w|(x_1, y_1), \ldots, (x_N, y_N))$, which in turn is proportional to $P((x_1, y_1), \ldots, (x_N, y_N)|w)p(w)$. We only make a prior assumption on m_1 so $p(w) = p(m_1) = \frac{1}{\sqrt{2\pi\lambda}} \exp(-\frac{m_1^2}{2\lambda^2})$. Instead of maximizing $P(w|(x_1, y_1), \ldots, (x_N, y_N))$, we can also maximize its logarithm or $\log P((x_1, y_1), \ldots, (x_N, y_N)|w) + \log p(w)$. The first term in the last expression equals the log loss we have seen in 3. So we take that expression and just add $\log p(w)$ to it to get to the objective function for the MAP estimates.
(e) Determining the MAP solutions can be done as in above… only the expression becomes a bit more complicated. In principle, however, it doesn't get more complex than solving a second order equation by means of the quadratic formula. Other things that may come in handy [depending on how you intend to solve it] are techniques like 'completing the square' and knowing that the product of two Gaussians is again a Gaussian, although it is not normalized.
E.g. $\hat{m}_1$ now becomes equal to

$$m_1^{\mathrm{MAP}} = \frac{N_1\lambda^2}{N_1\lambda^2 + s^2} m_1^{\mathrm{ML}}$$

(f) $\lim_{\lambda\to\infty} m_1^{\mathrm{MAP}} = m_1^{\mathrm{ML}}$ $\lim_{\lambda\downarrow 0} m_1^{\mathrm{MAP}} = 0$
(g) $\lim_{N_1\to\infty} m_1^{\mathrm{MAP}} = m_1^{\mathrm{ML}}$ $\lim_{N_1\downarrow 0} m_1^{\mathrm{MAP}} = 0$ $\lim_{s\downarrow 0} m_1^{\mathrm{MAP}} = m_1^{\mathrm{ML}}$

Exercise 4.6
(a) $\log p(x_i, y_i|w)$ and $\log \sum_y p(x_i, y|w)$.
(b) ...
(c) The expectation maximization (EM) algorithm.

Exercise 4.7
(a) We have a uniform distribution.
(b) I am not going to defend the solution, but just give the way to construct it. Calculate ML estimate for the mean, which is just the average of our N observations. If this mean is within the interval, this will also be the MAP solution. If not, we take the mean equal ta a or $-a$, whichever is nearer. This procedure produces the MAP mean. Given this mean, call it m', we can simply get the MAP variance v' as follows: $v' = \frac{1}{N}\sum_{i-1}^{N}()$
(c) If the mean is in the interval $[-a, a]$, this implies that $(l+u)/2$ should be in this interval as well. Now, the ML estimates for l and u are the minimum and the maximum over all x_i respectively. If $(l+u)/2$ is in the interval, we are done. If not, we need to change l and/or u. Now, to keep the likelihood as large as possible, we need to keep the interval between l and u as small as possible, but it should contain all data. If $(l+u)/2 > a$, we can fix this by decreasing $l+u$, but u cannot be decreased, as this would mean data points will fall outside of the fitting interval. So, we need to decrease l and the least amount d we need to decrease it with should be such that $(l-d+u)/2 = a$, so $d = l + u - 2a$. A similar argument can be made in case $(l+u)/2 < a$.

Exercise 4.8
(a) I would say that in all of the above, except for the Haldane prior, the bias increases and the variance reduces. All except Haldane clearly limit the variability of our estimates, though some do it more explicit than others. For Haldane it is maybe not that clear cut. As Haldane concentrates

around 0 and 1 it may have little variability, but the actual variance for the estimate q would probably go up when using Haldane... Maybe you can do the experiment? :)

Exercise 4.9
(a) $\theta = 1$.
(b) $\theta = 1$.
(c) Taking the same prior as in the question before, we find $p(Y^* = +1|Y = +1) = \int p(Y^* = +1|\theta)p(\theta|Y = +1)d\theta$. Now realise that $p(Y^* = +1|\theta) = \theta$, and rewrite $p(\theta|Y = +1)$ with Bayes' rule. Use $P(Y = +1) = \int p(Y = +1|\theta)p(\theta)d\theta$, and you find $p(Y^* = +1|Y = +1) = \frac{2}{3}$ and so $p(Y^* = -1|Y = +1) = \frac{1}{3}$.
(d) +1, +1, +1.
(e) When $c < 1/2$ it is better to start guessing $Y^* = -1$. ML and MAP will stick to +1 no matter what c is.
(f) ML: +1, MAP: +1, predictive: 1/3.

Exercise 4.10
No.

Exercise 4.11
(a) 25. There are 27 possible directed graphs, but two have a cycle.
(b) I would say eleven (11). 1 when there are no arrows; 3 when there is one arrow; 3 for chains and forks in the graph, since many of these two-arrow BNs model exactly the same conditional independencies; 3 for colliders; 1 for the fully connected graphs.

Exercise 4.12
(a) 6.
(b) Of course you can! If only by checking these solutions... The answer is $N!$.

Exercise 4.13
(a) 242
(b) No additional independences were assumed, so still 242.
(c) $C, BandC, B, D$

Exercise 4.14
(a) A general solution to such questions with BNs is that one expressed what is asked for in terms of the full probability. From there one can start simplifying and possibly use shortcuts based on conditional independence properties. In some cases, one can see a more direct and quicker solution.
For this particular question, we have $p(e) = \sum_l \sum_s p(e, l, s) = \sum_l \sum_s p(l)p(s)p(e|l, s)$ (where $\sum_x$ means that one sums over all values for x). The third expression only contains probabilities that are basically given in the text and so we can work out the sum computationally based on the values provided.
(b) Let us first do the second one. $p(l|e) = \frac{\sum_s p(e,l,s)}{\sum_s \sum_l p(e,l,s)}$. The denominator, we already determined in the previous exercise and $\sum_s p(e, l, s) = p(l) \sum_s p(s)p(e|l, s)$. This last expression can again be worked out further on the basis of the given numbers in the exercise.
For the first problem we could again take the same route: $p(e|l) = \frac{\sum_s p(e,l,s)}{\sum_s \sum_e p(e,l,s)}$ and then simplify based on the DAG. A bit more direct way is to realize that $p(e|l) = \sum_s p(s, e|l) = \sum_s p(s)p(e|l, s)$.

Exercise 4.15
It's a product of nonnegatives, so it is nonnegative. Now integrating over all variables in a clever way, starting at the pdfs that are not conditional (or are conditioned on the empty set $\emptyset$), it shouldn't be too hard to show that the integral is 1.

Exercise 4.16
Note that $p(a, c|b) = p(a|b, c)p(c|b) = p(a|b)p(c|b)$, where the second equality follows from the assumption. Combining this with $p(c|a, b) = p(a, c|b)/p(a|b)$, we see that $p(c|a, b) = p(a|b)p(c|b)/p(a|b) = p(c|b)$. Both statement are actually just different ways of expressing that

$a \perp c|b$.

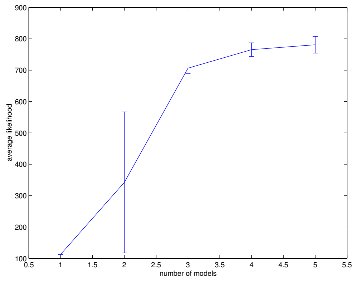
Exercise 4.17
(a) There are simply (much) less parameters to be estimated in the former case, which will lead to an overall better model fit with the same amount of data.
(b) The reduction in variance can easily outweigh the (small) increase in bias, which overall leads all in all to improvement.

Exercise 4.18
(a) At the finest level, there are 14 clusters. These small clusters are organised in three larger clusters.
(b) This is an artificial problem. However, the desired number of clusters will usually depend on the application. If the small clusters have a physical meaning, such as e.g. individual species while the three larger clusters represent animal kingdoms, then the specific problem which the biologist is studying will determine the appropriate number of clusters, i.e. the level of detail considered.

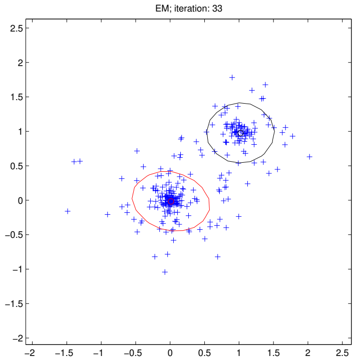
Exercise 4.19
(a) There are no vertical stems that are distinctly longer than the other. The tree grows gradually, not in leaps and bounds.
(c) In terms of the lengths of the stems there is no difference.

Exercise 4.20
(a) These lengths correspond to the distances (single, average or complete linkage) between the two clusters being joined by the bridge consisting of two stems and a vertical bar.
(b) The closer together the samples in a cluster and the further apart the clusters, the better the clustering.
(c) Long stems correspond to large distances between clusters and therefore a potential point to cut the dendrogram.
(d) Yes, in the single linkage dendrogram, there are basically two lengths of vertical stems: those indicating the joining of the smallest clusters, and those indicating the joining of the larger three clusters. This is less pronounced with complete linkage.
(e) The average linkage dendrogram is roughly similar to the complete linkage dendrogram.

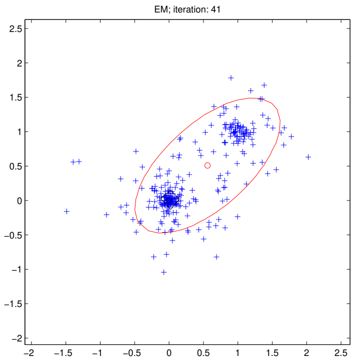
Exercise 4.21
(a) `'circular'`: diagonal covariance matrix with *equal* variances on the diagonal
`'aligned'`: diagonal covariance matrix with *unequal* variances on the diagonal
`'gauss'`: full covariance matrix, i.e. no constraints on the entries.
(b) The optimal k should be 2.
(c) The log-likelihood vs. the number of models is depicted in figure (a), the desired solution with two models in (b) and two frequently occurring undesirable solutions with two models in figures (c) and (d). These undesirable solutions are the reason why the variance is so large at two clusters and why the curve does not flatten at two, but at three clusters.



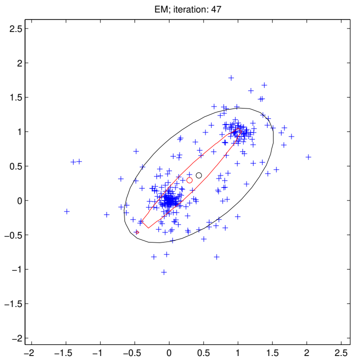
(a)



(b)



(c)



(d)

Exercise 4.22

(a) A large jump in the fusion graph implies that two clusters were merged that were, relative to the other cluster distances, quite far apart.

Exercise 4.23

(a) At $g = 3$.

(b) The fusion graph is ambiguous about the exact number of clusters: either two or three clusters are associated with large fusion jumps of roughly the same magnitude. This is due to the complete linkage distance: distance between furthest samples of clusters at $(0, 0)$ and $(1, 1)$ (which are merged to result from a total of two clusters) is roughly half the distance between the distance between the furthest samples in the cluster at $(0, 0)$ and $(2, 2)$, which are merged to result in a single cluster. In single linkage this does not occur, since the minimal distances between the clusters at $(0, 0)$ and $(1, 1)$ is the same as the minimal distance between the clusters at $(1, 1)$ and $(2, 2)$.

Exercise 4.24

(a) There are two pronounced jumps, at 3 and 14 clusters.

Exercise 4.25

(a) At three clusters, since the largest fusion jump occurs at $g = 3$.

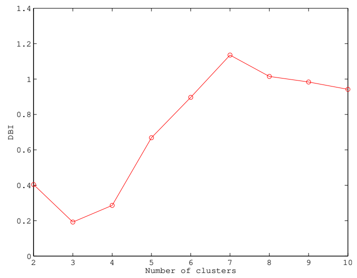
(b) No, three outliers constitute the members of two of the three clusters with all the other samples in the third cluster.

(c) While the size of the fusion jump is not very convincing (not much larger than the other jumps) the resulting clustering is better. This stems from the fact that complete linkage is less prone to outliers than single linkage.

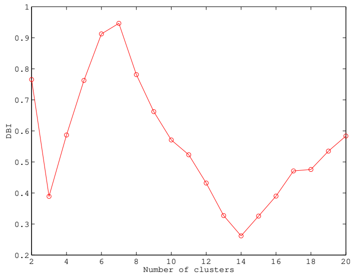
Exercise 4.26

(b) It should have a minimum at three clusters.

(c) The DBI curve is shown in the figure below, with a clear minimum at three clusters.



(d) There is a pronounced minimum at 3 clusters and a slightly larger minimum at 16 clusters. The first minimum (3 clusters) corresponds to the situation where the smaller clusters are grouped in three large clusters (four in top-left, four in top-right and six at the bottom) while the minimum at 16 corresponds to the fine cluster structure in the data. The peak at sixteen is more pronounced since the ratio of the maximal within-scatter to the minimal distance between any pair of these clusters is smaller than for the three cluster configuration.


DBI
Number of clusters