Machine Learning 521289S
Spring 2019
Exercise 3: Bayesian Networks

Here each node represents a random variable denoted with uppercase letter having associated discrete states in lowercase letters, e.g. node A has states $\{a_1, a_2, ..., a_n\}$. As befofe the upper case $P(\cdot)$ is used to denote the *probability mass function* of a discrete random variable

Solutions

1. *Bayes Nets.*

(a) Let us enumerate some of all the possible instances where the conditional independence assumptions in the Figure 1 can be used. For example, $P(C|X, A) = P(C|X)$ because there is no arrow between the nodes A and C. Similarly, $P(D|B) = P(D)$ and so on.

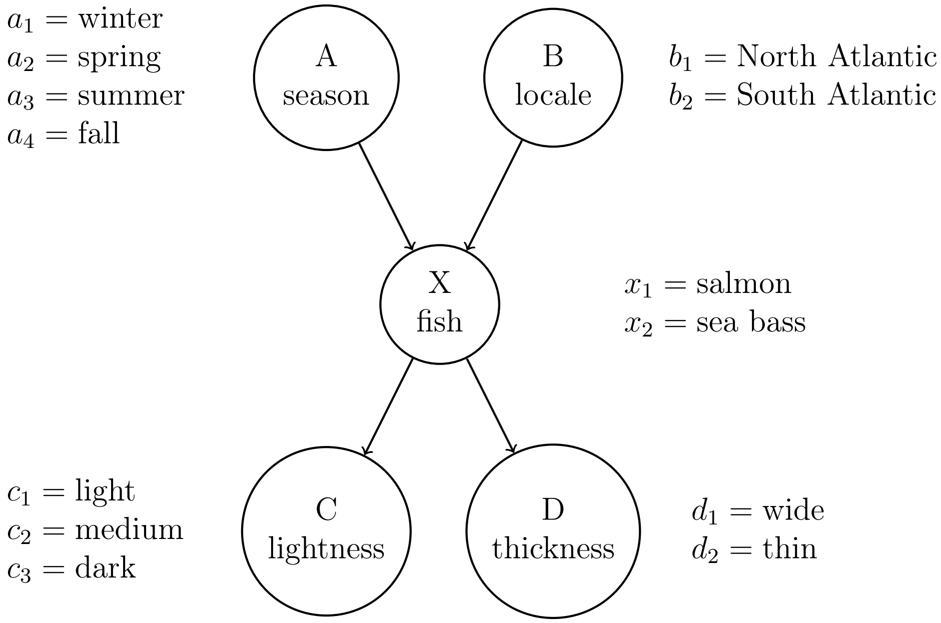

Figure 1: A belief network for fish. Adapted from the Duda, Hart & Stork, "Pattern classification," 2001.

These conditional independencies can be used to find a simplified formula for the joint probability mass distribution $P(A, B, X, C, D)$. Using the "product rule" (conditional probability), we may write

$$\begin{aligned}
P(A,B,X,C,D) &= P(D,A,B,X,C) \\
&= P(D|A,B,X,C)P(A,B,X,C) \\
&= P(D|A,B,X,C)P(C,A,B,X) \\
&= P(D|A,B,X,C)P(C|A,B,X)P(A,B,X) \\
&= P(D|A,B,X,C)P(C|A,B,X)P(X,A,B) \\
&= P(D|A,B,X,C)P(C|A,B,X)P(X|A,B)P(A,B) \\
&= P(D|A,B,X,C)P(C|A,B,X)P(X|A,B)P(A)P(B) \quad (1)
\end{aligned}$$

Now the conditional independencies tell us that $P(D|A,B,X,C) = P(D|X)$, and $P(C|A,B,X) = P(C|X)$. Hence, we may simplify (1) as

$$\begin{aligned}
P(A,B,X,C,D) &= P(D|X)P(C|X)P(X|A,B)P(A)P(B) \\
&= P(A)P(B)P(X|A,B)P(C|X)P(D|X). \qquad (2)
\end{aligned}$$

In the last step (2), we have just reordered the terms in the product for convenience.

(b) In this part, we may use (2) to calculate

$$\begin{aligned}
P(a_3,b_1,x_2,c_3,d_2) &= P(a_3)P(b_1)P(x_2|a_3,b_1)P(c_3|x_2)P(d_2|x_2) \\
&= 0.25 \times 0.6 \times 0.6 \times 0.5 \times 0.4 \\
&= 0.018 \approx 0.02. \qquad (3)
\end{aligned}$$

(c) In order to calculate $P(X)$, we use the marginalization trick, and then the conditional probability formula and the independence assumption to write

$$\begin{aligned}
P(x_1) &= \sum_{a,b} P(x_1,a,b) \\
&= \sum_{a,b} P(x_1|a,b)P(a,b) \\
&= \sum_a \sum_b P(x_1|a,b)P(a)P(b) \qquad (4)
\end{aligned}$$

Now, (4) is in the desired format where all the probabilities can be taken from the supplied tables. By reordering the terms in the sum and factoring out common terms in the product, however, we can reduce the number of calculations needed.

$$\begin{aligned}
P(x_1) &= \sum_a \sum_b P(x_1|a,b)P(a)P(b) \\
&= \sum_b P(b) \sum_a P(x_1|a,b)P(a) \qquad (5)
\end{aligned}$$

Note also that $P(a_i) = 0.25\ \forall i \in \{1,\ldots,4\}$. Let us denote this value by $p_a := 0.25$.

Consequently, (5) can be expressed as

$$\begin{aligned} P(x_1) &= p_a \sum_b P(b) \sum_a P(x_1|a,b) && (6) \\ &= p_a \times \{P(b_1) \times [P(x_1|a_1,b_1) + P(x_1|a_2,b_1) + P(x_1|a_3,b_1) + P(x_1|a_4,b_1)] \\ &\quad + P(b_2) \times [P(x_1|a_1,b_2) + P(x_1|a_2,b_2) + P(x_1|a_3,b_2) + P(x_1|a_4,b_2)]\} \\ &= 0.25 \times \{0.6 \times (0.5+0.6+0.4+0.2) + 0.4 \times (0.7+0.8+0.1+0.3)\} \\ &= 0.445. && (7) \end{aligned}$$

Since there are only two possible fish species, we have

$$P(x_2) = 1 - P(x_1) = 1 - 0.445 = 0.555. \tag{8}$$

(d) Now it is known that $B = b_2$. In this part, we have to calculate $P(x_1|b_2)$ and $P(x_2|b_2)$. Using the conditional probability formula, we proceed as follows.

$$\begin{aligned} P(x_1|b_2) &= \frac{P(x_1,b_2)}{P(b_2)} \\ &= \frac{\sum_a P(x_1,a,b_2)}{P(b_2)} \\ &= \frac{\sum_a P(x_1|a,b_2)P(a,b_2)}{P(b_2)} \\ &= \frac{\sum_a P(x_1|a,b_2)P(a)P(b_2)}{P(b_2)} \\ &= \sum_a P(x_1|a,b_2)P(a) \\ &= p_a \times \sum_a P(x_1|a,b_2) && (9) \end{aligned}$$

Note that the sum is exactly the same as the term in (6). Thus,

$$P(x_1|b_2) = 0.25 \times (0.7+0.8+0.1+0.3) = 0.475. \tag{10}$$

Using the same argument as in (8), we get

$$P(x_2|b_2) = 1 - P(x_1|b_2) = 1 - 0.475 = 0.525. \tag{11}$$

(e) Now, the question is, what is $P(x_1|a_4,b_2)$? The result can be red directly from the probability tables for this exercise, i.e.

$$P(x_1|a_4,b_2) = 0.3. \tag{12}$$

and

$$P(x_2|a_4,b_2) = 0.7. \tag{13}$$

2. *More Bayes Nets.*

The network models the dependencies (causality, conditional probabilities, arrows) between cloudy sky (C) and sprinkler (S), cloudy sky and rain (R), sprinkler and wet grass (W), and rain and wet grass. The network contains four two-class (true=T or false=F) discrete nodes. Using the conditional probability formula $P(A,B) = P(A|B)P(B)$, we have the joint probability

$$\begin{aligned} P(c,r,s,w) &= P(w,r,s,c) \\ &= P(w|r,s,c)\,P(r,s,c) \\ &= P(w|r,s,c)\,P(r|s,c)\,P(s,c) \\ &= P(w|r,s,c)\,P(r|s,c)\,P(s|c)\,P(c)\,. \end{aligned}$$

Please note that now we may deduce from the supplied figure that $P(r|s,c) = P(r|c)$, as the rain node depends directly only on the cloudy sky node. Similarly, $P(w|r,s,c) = P(w|r,s)$, as the wet grass node depends directly only on the rain and sprinkler nodes. Thus, we have

$$\begin{aligned} P(c,s,r,w) &= P(w|r,s)\,P(r|c)\,P(s|c)\,P(c) \\ &= P(c)\,P(s|c)\,P(r|c)\,P(w|r,s)\,. \qquad (14) \end{aligned}$$

The probabilities needed are shown in the given tables.

(a) The probability that it is cloudy (without further evidence) is

$$P(c_T) = 0.5$$

which may be directly red from the figure. Note that for brevity, we use the notation $P(c_T)$ as a shorthand for $P(C=T) = P_C(T)$.

(b) Again from the tables, we see directly that the probability that it does not rain given that it is cloudy is

$$P(r_F|c_T) = 0.2.$$

(c) To solve the probability that the grass is wet (without further evidence) requires a bit more calculation. First, remember that we get the marginal distribution for the wet node by considering all the possibilities for other nodes and summing up the probabilities of these mutually exclusive events, i.e. $P(w) = \sum_{c,r,s} P(c,s,r,w)\,.$ Now, as the grass is wet, we have

$$\begin{aligned} P(w_T) &= \sum_{c,r,s} P(c,s,r,w_T) \\ &= \sum_{c\in\{T,F\}} \sum_{r\in\{T,F\}} \sum_{s\in\{T,F\}} P(c,s,r,w_T) \qquad (15) \end{aligned}$$

Here, we could now utilise the conditional independencies of the variables by substituting (14) into (15) writing it as

$$P(w_T) = \sum_{c\in\{T,F\}} P(c) \sum_{s\in\{T,F\}} P(s|c) \sum_{r\in\{T,F\}} P(r|c)\,P(w_T|r,s)$$

in order to reduce the number of operations it takes to evaluate the sum. However, it is somewhat easier and more illustrative to still work with the full joint distribution. Hence, we just expand (15) as

$$\begin{aligned}
P\left(w_T\right) &= P\left(c_T, s_T, r_T, w_T\right) \\
&\quad +P\left(c_F, s_T, r_T, w_T\right) \\
&\quad +P\left(c_T, s_F, r_T, w_T\right) \\
&\quad +P\left(c_F, s_F, r_T, w_T\right) \\
&\quad +P\left(c_T, s_T, r_F, w_T\right) \\
&\quad +P\left(c_F, s_T, r_F, w_T\right) \\
&\quad +P\left(c_T, s_F, r_F, w_T\right) \\
&\quad +P\left(c_F, s_F, r_F, w_T\right).
\end{aligned}$$

Now we use (14) to get

$$\begin{aligned}
P\left(w_T\right) &= P\left(c_T\right) P\left(s_T|c_T\right) P\left(r_T|c_T\right) P\left(w_T|r_T, s_T\right) \\
&\quad +P\left(c_F\right) P\left(s_T|c_F\right) P\left(r_T|c_F\right) P\left(w_T|r_T, s_T\right) \\
&\quad +P\left(c_T\right) P\left(s_F|c_T\right) P\left(r_T|c_T\right) P\left(w_T|r_T, s_F\right) \\
&\quad +P\left(c_F\right) P\left(s_F|c_F\right) P\left(r_T|c_F\right) P\left(w_T|r_T, s_F\right) \\
&\quad +P\left(c_T\right) P\left(s_T|c_T\right) P\left(r_F|c_T\right) P\left(w_T|r_F, s_T\right) \\
&\quad +P\left(c_F\right) P\left(s_T|c_F\right) P\left(r_F|c_F\right) P\left(w_T|r_F, s_T\right) \\
&\quad +P\left(c_T\right) P\left(s_F|c_T\right) P\left(r_F|c_T\right) P\left(w_T|r_F, s_F\right) \\
&\quad +P\left(c_F\right) P\left(s_F|c_F\right) P\left(r_F|c_F\right) P\left(w_T|r_F, s_F\right).
\end{aligned}$$

Finally, we may use the values from the tables to get

$$\begin{aligned}
P\left(w_T\right) &= 0,5 \times 0,1 \times 0,8 \times 0,99 \\
&\quad +0,5 \times 0,5 \times 0,2 \times 0,99 \\
&\quad +0,5 \times 0,9 \times 0,8 \times 0,9 \\
&\quad +0,5 \times 0,5 \times 0,2 \times 0,9 \\
&\quad +0,5 \times 0,1 \times 0,2 \times 0,9 \\
&\quad +0,5 \times 0,5 \times 0,8 \times 0,9 \\
&\quad +0,5 \times 0,9 \times 0,2 \times 0 \\
&\quad +0,5 \times 0,5 \times 0,8 \times 0 \\
&= 0,6471.
\end{aligned}$$

Hence, $P\left(w_T\right) \approx 0,65$.

(d) The probability that the sprinkler was on with the evidence that the grass is wet is

$$P\left(s_T|w_T\right) = \frac{P\left(s_T, w_T\right)}{P\left(w_T\right)}.$$

Here,

$$
\begin{aligned}
P\left(s_T, w_T\right) &= \sum_{c,r} P\left(c, s_T, r, w_T\right) \\
&= \sum_{c\in\{T,F\}} \sum_{r\in\{T,F\}} P\left(c, s_T, r, w_T\right) \\
&= P\left(c_T, s_T, r_T, w_T\right) \\
&\quad +P\left(c_F, s_T, r_T, w_T\right) \\
&\quad +P\left(c_T, s_T, r_F, w_T\right) \\
&\quad +P\left(c_F, s_T, r_F, w_T\right)
\end{aligned}
$$

Using Equation 14 again, we get

$$
\begin{aligned}
P\left(s_T, w_T\right) &= P\left(c_T\right) P\left(s_T|c_T\right) P\left(r_T|c_T\right) P\left(w_T|r_T, s_T\right) \\
&\quad +P\left(c_F\right) P\left(s_T|c_F\right) P\left(r_T|c_F\right) P\left(w_T|r_T, s_T\right) \\
&\quad +P\left(c_T\right) P\left(s_T|c_T\right) P\left(r_F|c_T\right) P\left(w_T|r_F, s_T\right) \\
&\quad +P\left(c_F\right) P\left(s_T|c_F\right) P\left(r_F|c_F\right) P\left(w_T|r_F, s_T\right) \\
&= 0,5 \times 0,1 \times 0,8 \times 0,99 \\
&\quad +0,5 \times 0,5 \times 0,2 \times 0,99 \\
&\quad +0,5 \times 0,1 \times 0,2 \times 0,9 \\
&\quad +0,5 \times 0,5 \times 0,8 \times 0,9 \\
&= 0,2781.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
P\left(s_T|w_T\right) &= \frac{P\left(s_T, w_T\right)}{P\left(w_T\right)} \\
&= \frac{0.2781}{0.6471} \\
&= 0.42976 \\
&\approx 0.43.
\end{aligned}
$$

(e) The probability that it rains given that the sprinkler is on is

$$
P\left(r_T|s_T\right) = \frac{P\left(r_T, s_T\right)}{P\left(s_T\right)} = \alpha P\left(s_T, r_T\right). \tag{16}
$$

where $\alpha := P\left(s_T\right)^{-1}$. Let us this time use the factorisation of the sum to reduce the number of operations, although the notation becomes a bit cumbersome.

$$\begin{aligned}
P\left(s_T, r_T\right) &= \sum_{c,w} P\left(c, s_T, r_T, w\right) \\
&= \sum_{c,w} P\left(c\right) P\left(s_T|c\right) P\left(r|c\right) P\left(w|r_T, s_T\right) \\
&= \sum_{c\in\{T,F\}} \left(P\left(c\right) P\left(s_T|c\right) P\left(r_T|c\right) \underbrace{\sum_{w\in\{T,F\}} P\left(w|r_T, s_T\right)}_{=1} \right) \\
&= \sum_{c\in\{T,F\}} P\left(c\right) P\left(s_T|c\right) P\left(r_T|c\right) \\
&= P\left(c_T\right) P\left(s_T|c_T\right) P\left(r_T|c_T\right) + P\left(c_F\right) P\left(s_T|c_F\right) P\left(r_T|c_F\right) \\
&= 0.5 \times 0.1 \times 0.8 + 0.5 \times 0.5 \times 0.2 \\
&= 0.09
\end{aligned}$$

Similarly, we calculate

$$\begin{aligned}
P\left(s_T, r_F\right) &= P\left(c_T\right) P\left(s_T|c_T\right) P\left(r_F|c_T\right) + P\left(c_F\right) P\left(s_T|c_F\right) P\left(r_F|c_F\right) \\
&= 0.5 \times 0.1 \times 0.2 + 0.5 \times 0.5 \times 0.8 \\
&= 0.21
\end{aligned}$$

We can now calculate the scaling parameter α, since

$$P\left(r_T|s_T\right) + P\left(r_F|s_T\right) = 1$$

it follows from (16) that

$$\begin{aligned}
\alpha \times \left[P\left(s_T, r_T\right) + P\left(s_T, r_F\right)\right] &= 1 \\
\alpha &= \frac{1}{P\left(s_T, r_T\right) + P\left(s_T, r_F\right)}.
\end{aligned}$$

Hence,

$$\alpha = \frac{1}{0.09 + 0.21} \approx 3.333.$$

Finally, we get that

$$\begin{aligned}
P\left(r_T|s_T\right) &= \alpha\ P\left(s_T, r_T\right) \\
&= 3.333 \times 0.09 \\
&\approx 0.30.
\end{aligned}$$

3. *Naive Bayes Nets.*

   Let's denote the nodes: Play Tennis=P, Outlook=O, Temperature=T, Humidity=H and Wind=W. Using the naive Bayes assumption of conditional independence[1] we have the net shown in Figure 2.

[1] See for example the handout (or the old lecture notes page 53, in Finnish).

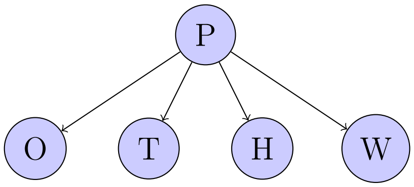

Figure 2: The naive Bayes net.

Furthermore, the probability of playing tennis given the observations is

$$P\left(p|o,t,h,w\right)=\alpha P\left(p\right)P\left(o|p\right)P\left(t|p\right)P\left(h|p\right)P\left(w|p\right).$$

We are interested in the case (Outlook=Sunny, Temperature=Cool, Humidity=High, Wind=Strong). Consequently,

$$P\left(p|o_{sunny},t_{cool},h_{high},w_{strong}\right)=\alpha P\left(p\right)P\left(o_{sunny}|p\right)P\left(t_{cool}|p\right)P\left(h_{high}|p\right)P\left(w_{strong}|p\right),$$

where $p\in\{yes,no\}$.

From the supplied table, we may estimate the priors

$$\begin{aligned}P\left(p_{yes}\right)&=\frac{9}{14}=0,64,\\P\left(p_{no}\right)&=\frac{5}{14}=0,36.\end{aligned}$$

Similarly, we may estimate the conditional probabilities

$$\begin{aligned}P\left(o_{sunny}|p_{yes}\right)&=\frac{2}{9}=0,22,\\P\left(o_{sunny}|p_{no}\right)&=\frac{3}{5}=0,60,\\P\left(t_{cool}|p_{yes}\right)&=\frac{3}{9}=0,33,\\P\left(t_{cool}|p_{no}\right)&=\frac{1}{5}=0,20,\\P\left(h_{high}|p_{yes}\right)&=\frac{3}{9}=0,33,\\P\left(h_{high}|p_{no}\right)&=\frac{4}{5}=0,80,\\P\left(w_{strong}|p_{yes}\right)&=\frac{3}{9}=0,33,\\P\left(w_{strong}|p_{no}\right)&=\frac{3}{5}=0,60.\end{aligned}$$

Now, omitting the scaling factor α for brevity

$$\begin{aligned}&P\left(p_{yes}\right)P\left(o_{sunny}|p_{yes}\right)P\left(t_{cool}|p_{yes}\right)P\left(h_{high}|p_{yes}\right)P\left(w_{strong}|p_{yes}\right)\\=&\frac{9}{14}\times\frac{2}{9}\times\frac{3}{9}\times\frac{3}{9}\times\frac{3}{9}\\=&\frac{1}{189}\\\approx&0.0053\end{aligned}$$

and

$$\begin{aligned}
& P\left(p_{no}\right) P\left(o_{sunny}|p_{no}\right) P\left(t_{cool}|p_{no}\right) P\left(h_{high}|p_{no}\right) P\left(w_{strong}|p_{no}\right) \\
= & \frac{5}{14} \times \frac{3}{5} \times \frac{1}{5} \times \frac{4}{5} \times \frac{3}{5} \\
= & \frac{18}{875} \\
\approx & 0.0206.
\end{aligned}$$

Because the latter probability is higher, the naive Bayes classifier decides that it **is not** an appropriate whether to play. Note that the we do not have to calculate the actual value for α even if we wanted to know the actual probabilities, as they sum up to one and are

$$\begin{aligned}
P\left(p_{yes}|o_{sunny}, t_{cool}, h_{high}, w_{strong}\right) &= \frac{0.0053}{0.0053 + 0.0206} \\
&= 0.204\,63 \\
&\approx 0.20
\end{aligned}$$

and

$$\begin{aligned}
P\left(p_{no}|o_{sunny}, t_{cool}, h_{high}, w_{strong}\right) &= \frac{0.0206}{0,0053 + 0.0206} \\
&= 0.795\,37 \\
&\approx 0.80.
\end{aligned}$$

4. *Hidden Markov Models.*

   (a) Here, we are actually interested in selecting the internal model that is most likely given the observations, i.e. the one that maximises

$$\begin{aligned}
& P\left(h_1, h_2, \ldots, h_N|o_1, o_2, \ldots, o_N\right) \\
= & \frac{P\left(o, o_2, \ldots, o_N|h_1, h_2, \ldots, h_N\right) P\left(h_1, h_2, \ldots, h_N\right)}{P\left(o_1, o_2, \ldots, o_N\right)}.
\end{aligned}$$

   In the above, we have used the Bayes formula. Note also that the term in the denominator does not depend on the choice of the internal state sequence and may be omitted in maximisation. From the given figure, we may see that the prior probability may be factored as

$$\begin{aligned}
P\left(h_1, h_2, \ldots, h_N\right) &= P\left(h_N, h_{N-1}, \ldots, h_1\right) \\
&= P\left(h_N|h_{N-1}, \ldots, h_1\right) P\left(h_{N-1}, \ldots, h_1\right) \\
&= P\left(h_N|h_{N-1}\right) P\left(h_{N-1}, \ldots, h_1\right) \\
&= P\left(h_N|h_{N-1}\right) P\left(h_{N-1}|h_{N-2}, \ldots, h_1\right) P\left(h_{N-2}, \ldots, h_1\right) \\
&= \ldots \\
&= P\left(h_1\right) \prod_{i=2}^{N} P\left(h_i|h_{i-1}\right)
\end{aligned}$$

It is also clear from the figure that the conditional independence allows us to simplify the likelihood as

$$\begin{aligned}
& P\left(o_1, o_2, \ldots, o_N | h_1, h_2, \ldots, h_N\right) \\
= \; & P\left(o_1 | h_1, h_2, \ldots, h_N\right) P\left(o_2 | h_1, h_2, \ldots, h_N\right) \cdots P\left(o_N | h_1, h_2, \ldots, h_N\right) \\
= \; & P\left(o_1 | h_1\right) P\left(o_2 | h_2\right) \cdots P\left(o_N | h_N\right) \\
= \; & \prod_{i=i}^{k} P\left(o_i | h_i\right).
\end{aligned}$$

Here in the first equality, we have used the fact that the observed nodes are conditionally independent. The second equality follows from the fact that the outputs do not depend on any other node than the hidden node they are connected to. Consequently, we have found out that

$$\begin{aligned}
P\left(h_1, h_2, \ldots, h_N | o_1, o_2, \ldots, o_N\right) &= \alpha \; P\left(h_1\right) \prod_{i=2}^{N} P\left(h_i | h_{i-1}\right) \prod_{i=i}^{N} P\left(o_i | h_i\right) \\
&= \alpha \; P\left(h_1\right) \; P\left(o_1 | h_1\right) \prod_{i=2}^{N} P\left(h_i | h_{i-1}\right) \; P\left(o_i | h_i\right),
\end{aligned}$$

where $\alpha = P\left(o_1, o_2, \ldots, o_N\right)^{-1}$.

(b) In a HMM with 15 time steps there would be $5^{15} = 30\,517\,578\,125$ combinations to be tested assuming that each hidden node could have 5 states when using a brute force approach. Further, if we calculated 1000 posteriors per second, it would take

$$\frac{30\,517\,578\,125}{1000\frac{1}{s} \times 60 \; \frac{s}{\min} \times 60 \; \frac{\min}{h} \times 24\frac{h}{d} \times 365\frac{d}{a}} \approx 0.967\,71 \; a$$

that is almost a year to select the optimal combination. This of course, is infeasible. There is a better way to find the best combination. It is called the *Viterbi algorithm* (1967) named after its inventor. It is a dynamic programming algorithm, and a very efficient one. Everyone is encouraged to google a bit more on it.