Solutions to the Exercises[*] on Bayesian Theory and Graphical Models

Laurenz Wiskott
Institut für Neuroinformatik
Ruhr-Universität Bochum, Germany, EU

4 February 2017

Contents

Several of my exercises (not necessarily on this topic) were inspired by papers and textbooks by other authors. Unfortunately, I did not document that well, because initially I did not intend to make the exercises publicly available, and now I cannot trace it back anymore. So I cannot give as much credit as I would like to. The concrete versions of the exercises are certainly my own work, though.

In cases where I reuse an exercise in different variants, references may be wrong for technical reasons.

[*] These exercises complement my corresponding lecture notes available at https://www.ini.rub.de/PEOPLE/wiskott/Teaching/Material/, where you can also find other teaching material such as programming exercises. The table of contents of the lecture notes is reproduced here to give an orientation when the exercises can be reasonably solved. For best learning effect I recommend to first seriously try to solve the exercises yourself before looking into the solutions.

1 Bayesian inference

1.1 Discrete random variables and basic Bayesian formalism

Joint probability

1.1.1 Exercise: Heads-tails-tails-heads

1. With four tosses of a fair coin, what is the probability to get exactly heads-tails-tails-heads, in this order?

   Solution: Each toss is independent of the others and the probability for each toss to get the desired result is $\frac{1}{2}$. Thus, the probability to get exactly heads-tails-tails-heads is $\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{16}$. This, by the way, holds for any concrete combination of length four.

2. With four tosses of a fair coin, what is the probability to get each heads and tails twice, regardless of the order?

   Solution: The probability for any particular combination of four times heads or tails is $\frac{1}{16}$, see above. Since there are six different ways to get heads and tails twice (namely tthh, thth, thht, htth, htht, hhtt), the probability to get any of these is $\frac{6}{16} = \frac{3}{8}$.

Total probability

1.1.2 Exercise: Election and Bridge

Three candidates run for an election as a major in a city. According to a public opinion poll their chances to win are 0.25, 0.35 und 0.40. The chances that they build a bridge after they have been elected are 0.60, 0.90 und 0.80. What is the probability that the bridge will be build after the election.

Solution: Let $C, c \in \{1, 2, 3\}$, be the random variable indicating the winning candidate and $B, b \in \{t, f\}$, the random variable indicating whether the bridge will be built. Then the total probability that the bridge will be built is

$$P(B = t) = \sum_{c=1}^{3} P(B = t|c)P(c) = 0.60 \times 0.25 + 0.90 \times 0.35 + 0.80 \times 0.40 = 0.785\,.$$

Bayes formula

1.1.3 Exercise: Bayes theorem in four variables

Consider four random variables A, B, C, and D. Given are the (marginal) joint probabilities for each pair of variables, i.e. probabilities of the form $P(A, B), P(A, C)$ etc., and the conditional probability $P(A, B|C, D)$.

Calculate $P(A, C|B, D)$.

Solution:

$$\begin{aligned} P(A, C|B, D) &= \frac{P(A, B, C, D)}{P(B, D)} && (1) \\ &= \frac{P(A, B|C, D)P(C, D)}{P(B, D)}\,. && (2) \end{aligned}$$

1.1.4 Exercise: Airport security

On an airport all passengers are checked carefully. Let T with $t \in \{0, 1\}$ be the random variable indicating whether somebody is a terrorist ($t = 1$) or not ($t = 0$) and A with $a \in \{0, 1\}$ be the variable indicating arrest. A terrorist shall be arrested with probability $P(A = 1|T = 1) = 0.98$, a non-terrorist with probability $P(A = 1|T = 0) = 0.001$. One in hundredthousand passengers is a terrorist, $P(T = 1) = 0.00001$. What is the probability that an arrested person actually is a terrorist?

Solution: This can be solved directly with the Bayesian theorem.

$$\begin{aligned} P(T = 1|A = 1) &= \frac{P(A = 1|T = 1)P(T = 1)}{P(A = 1)} && (1) \\ &= \frac{P(A = 1|T = 1)P(T = 1)}{P(A = 1|T = 1)P(T = 1) + P(A = 1|T = 0)P(T = 0)} && (2) \\ &= \frac{0.98 \times 0.00001}{0.98 \times 0.00001 + 0.001 \times (1 - 0.00001)} = 0.0097 && (3) \\ &\approx \frac{0.00001}{0.001} = 0.01 && (4) \end{aligned}$$

It is interesting that even though for any passenger it can be decided with high reliability (98% and 99.9%) whether (s)he is a terrorist or not, if somebody gets arrested as a terrorist, (s)he is still most likely not a terrorist (with a probability of 99%).

1.1.5 Exercise: Drug Test

A drug test (random variable T) has 1% false positives (i.e., 1% of those not taking drugs show positive in the test), and 5% false negatives (i.e., 5% of those taking drugs test negative). Suppose that 2% of those tested are taking drugs. Determine the probability that somebody who tests positive is actually taking drugs (random variable D).

Solution:

$T = p$ means Test positive,
$T = n$ means Test negative,
$D = p$ means person takes drug,
$D = n$ means person does not take drugs

We know:

$$\begin{aligned}
P(T = p|D = n) &= 0.01 \quad \text{(false positives)} && (1)\\
\text{(false negatives)} \quad P(T = n|D = p) &= 0.05 \Longrightarrow P(T = p|D = p) = 0.95 \quad \text{(true positives)} && (2)\\
P(D = p) &= 0.02 \Longrightarrow P(D = n) = 0.98 && (3)\\
& && (4)
\end{aligned}$$

We want to know the probability that somebody who tests positive is actually taking drugs:

$$P(D = p|T = p) = \frac{P(T = p|D = p)P(D = p)}{P(T = p)} \text{(Bayes theorem)} \tag{5}$$

We do not know $P(T = p)$:

$$P(T = p) = P(T = p|D = p)P(D = p) + P(T = p|D = n)P(D = n) \tag{6}$$

We get:

$$\begin{aligned}
P(D = p|T = p) &= \frac{P(T = p|D = p)P(D = p)}{P(T = p)} && (7)\\
&= \frac{P(T = p|D = p)P(D = p)}{P(T = p|D = p)P(D = p) + P(T = p|D = n)P(D = n)} && (8)\\
&= \frac{0.95 \cdot 0.02}{0.95 \cdot 0.02 + 0.01 \cdot 0.98} && (9)\\
&= 0.019/0.0288 \approx 0.66 && (10)
\end{aligned}$$

There is a chance of only two thirds that someone with a positive test is actually taking drugs.

An alternative way to solve this exercise is using decision trees. Let's assume there are 1000 people tested. What would the result look like?

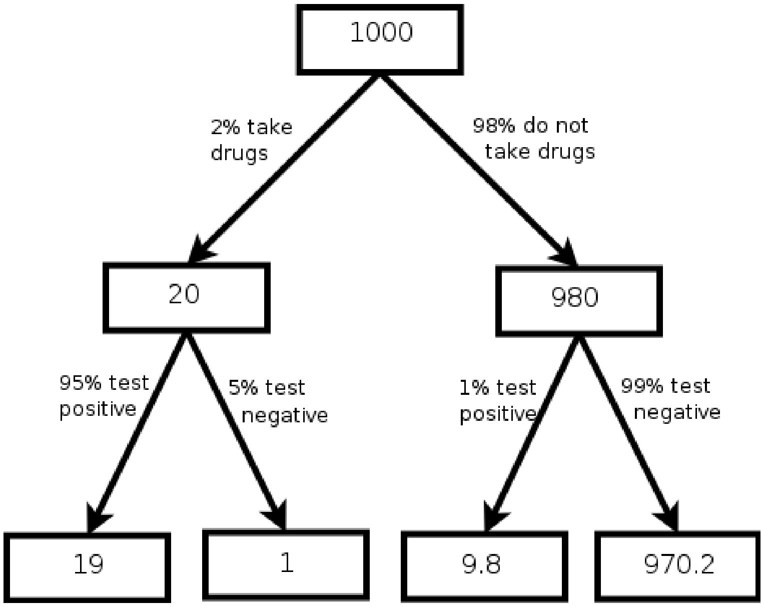


Figure: (Uknown, © unclear)

Now we can put this together in a contingency table:

	$D = p$	$D = n$	sum
$T = p$	19	9.8	28.8
$T = n$	1	970.2	971.2
sum	20	980	1000

To determine the probability that somebody who tests positive is actually taking drugs we have to calculate:

$$\frac{\text{taking drugs and positive test}}{\text{all positive test}} = \frac{19}{28.8} \approx 0.66 \tag{11}$$

1.1.6 Exercise: Oral Exam

In an oral exam you have to solve exactly one problem, which might be one of three types, A, B, or C, which will come up with probabilities 30%, 20%, and 50%, respectively. During your preparation you have solved 9 of 10 problems of type A, 2 of 10 problems of type B, and 6 of 10 problems of type C.

(a) What is the probability that you will solve the problem of the exam?

Solution: The probability to solve the problem of the exam is the probability of getting a problem of a certain type times the probability of solving such a problem, summed over all types. This is known as the total probability.

$$\begin{aligned} P(\text{solved}) &= P(\text{solved}|\text{A})P(\text{A}) + P(\text{solved}|\text{B})P(\text{B}) + P(\text{solved}|\text{C})P(\text{C}) && (1)\\ &= 9/10 \cdot 30\% + 2/10 \cdot 20\% + 6/10 \cdot 50\% && (2)\\ &= 27/100 + 4/100 + 30/100 = 61/100 = 0.61\,. && (3) \end{aligned}$$

(b) Given you have solved the problem, what is the probability that it was of type A?

Solution: For this to answer we need Bayes theorem.

$$
\begin{aligned}
P(\mathrm{A}|\mathrm{solved}) &= \frac{P(\mathrm{solved}|\mathrm{A})P(\mathrm{A})}{P(\mathrm{solved})} && (4)\\
&= \frac{9/10 \cdot 30\%}{61/100} = \frac{27/100}{61/100} = \frac{27}{61} = 0.442\ldots . && (5)\\
& && (6)
\end{aligned}
$$

So we see that given you have solved the problem, the *a posteriori* probability that the problem was of type A is greater than its *a priori* probability of 30%, because problems of type A are relatively easy to solve.

1.1.7 Exercise: Radar station

Consider a radar station monitoring air traffic. For simplicity we chunk time into periods of five minutes and assume that they are independent of each other. Within each five minute period, there may be an airplane flying over the radar station with probability 5%, or there is no airplane (we exclude the possibility that there are several airplanes). If there is an airplane, it will be detected by the radar with a probability of 99%. If there is no airplane, the radar will give a false alarm and detect a non-existent airplane with a probability of 10%.

1. How many airplanes fly over the radar station on average per day (24 hours)?

   Solution: There are $24 \times 12 = 288$ five-minute periods per day. In each period there is a probability of 5% for an airplane being present. Thus the average number of airplanes is $288 \times 5\% = 288 \times 0.05 = 14.4$.

2. How many false alarms (there is an alarm even though there is no airplane) and how many false no-alarms (there is no alarm even though there is an airplane) are there on average per day.

   Solution: On average there is no airplane in $288 - 14.4$ of the five-minute periods. This times the probability of 10% per period for a false alarm yields $(288 - 14.4) \times 10\% = 273.6 \times 0.1 = 27.36$ false alarms.

   On average there are 14.4 airplanes, each of which has a probability of 1% of getting missed. Thus the number of false no-alarms is $14.4 \times 1\% = 14.4 \times 0.01 = 0.144$.

3. If there is an alarm, what is the probability that there is indeed an airplane?

   Solution: For this question we need Bayes theorem.

$$
\begin{aligned}
& P(\mathrm{airplane}|\mathrm{alarm}) && (1)\\
=\; & \frac{P(\mathrm{alarm}|\mathrm{airplane})P(\mathrm{airplane})}{P(\mathrm{alarm})} && (2)\\
=\; & \frac{P(\mathrm{alarm}|\mathrm{airplane})P(\mathrm{airplane})}{P(\mathrm{alarm}|\mathrm{airplane})P(\mathrm{airplane}) + P(\mathrm{alarm}|\mathrm{no\ airplane})P(\mathrm{no\ airplane})} && (3)\\
=\; & \frac{0.99 \cdot 0.05}{0.99 \cdot 0.05 + 0.1 \cdot (1 - 0.05)} = 0.342... && (4)\\
\approx\; & \frac{0.05}{0.05 + 0.1} = 0.333\ldots . && (5)
\end{aligned}
$$

   It might be somewhat surprising that the probability of an airplane being present given an alarm is only 34% even though the detection of an airplane is so reliable (99%). The reason is that airplanes are not so frequent (only 5%) and the probability for an alarm given no airplane is relatively high (10%).

Miscellaneous

1.1.8 Exercise: Gambling machine

Imagine a simple gambling machine. It has two display fields that can light up in red or green. The first one lights up first with green being twice as frequent as red. The color of the second field depends on the first one. If the first color is red, green appears five times as often as red in the second field. If the first color is green, the two colors are equally likely.

A game costs 8€ and goes as follows. The player can tip right in the beginning on both colors, or he can tip the second color after he sees the first color, or he can tip not at all. He is allowed to decide on when he tips during the game. The payout for the three tip options is different of course, highest for tipping two colors and lowest for no tip at all.

1. To get a feeling for the question, first assume for simplicity that each color is equally likely and the second color is independent of the first one. How high must the payout for each of the three tip options be, if the tip is correct, to make the game just worth playing?

   Solution: If all colors are equally likely, then one would tip a two-color combination correctly with probability 1/4, the second color alone with 1/2, and no color with certainty. Thus the payout if the tip is correct must be a bit more than 32€, 16€, and 8€, respectively, to make the game worth playing.

2. Do the chances to win get better or worse if the colors are not equally likely anymore but have different probabilities and you know the probabilities? Does it matter whether the two fields are statistically independent or not?

   Solution: If the probabilities different, then some combinations are more frequent than others. If one systematically tips these more frequent combinations, the mean payout is increased. Thus, chances get better.

3. Given the payouts for the three tip options are 20€, 12€, and 7€. What is the optimal tip strategy and what is the mean payout?

   Solution: The solution to this question can be put in a table.

cost of one game	-8€			
best tip now	green-red or green-green			
mean payout for best tip now	+6/18·20€ = +6$\frac{2}{3}$€			
mean payout for best tip later	+1/3·10€ +2/3·7€ = +24/3€ = +8€			
prob. of first color	1/3 red		2/3 green	
best tip now	green		red or green	
mean payout for best tip now	+5/6·12€ = +10€		+3/6·12€ = +6€	
mean payout for no tip now	+7€		+7€	
prob. of second color given first	1/6 red	5/6 green	3/6 red	3/6 green
prob. of color combination	1/18 red-red	5/18 red-green	6/18 green-red	6/18 green-green
payout for no tip	+7€	+7€	+7€	+7€

   Thus the best strategy is not to tip initially and then tip green as the second color if red comes up as the first color. If green comes up as the first color, don't tip at all. The mean payout of this optimal strategy is 8€, which just cancels the costs of the game.

1.1.9 Exercise: Probability theory

1. A friend offers you a chance to win some money by betting on the pattern of heads and tails shown on two coins that he tosses hidden from view. Sometimes he is allowed to give you a hint as to the result, sometimes not. Calculate the following probabilities:

   (a) If your friend stays silent, the probability that the coins show TT.

   (b) If your friend stays silent, the probability that the coins show HT in any order.

   (c) If you friend tells you that at least one of the coins is an H, the probability that the coins show HH.

2. Your friend now invents a second game. This time he tosses a biased coin which can produce three different results. The coin shows an H with probability 0.375 and a T with probability 0.45. The rest of the time the coin shows neither an H nor a T but lands on its side. A round of the game consists of repeatedly tossing the coin until either an H or a T comes up, whereupon the round ends.

   (a) Calculate the probability the coin needs to be tossed more than three times before either an H or a T comes up and the round ends.

   (b) Your friend proposes that if a T comes up you have to pay him 8 Euros, while if an H comes up he has to pay you 10 Euros. Calculate the expectation value of your winnings per round when playing this game. Would you agree to play using these rules?

   (c) Assume you agree with your friend that if a T comes up you have to pay him 10 Euros. What is the minimum amount you should receive if an H comes up in order to give a positive expectation value for your winnings per round?

   (d) Your friend now produces a coin which is always either an H or a T. In other words, it cannot land on its side. He claims that using this new coin eliminates the need to re-toss the coin without changing the statistics of the game in any other way. Assuming this is true, what is the probability of getting an H and a T on this new coin?

Solution: Not available!

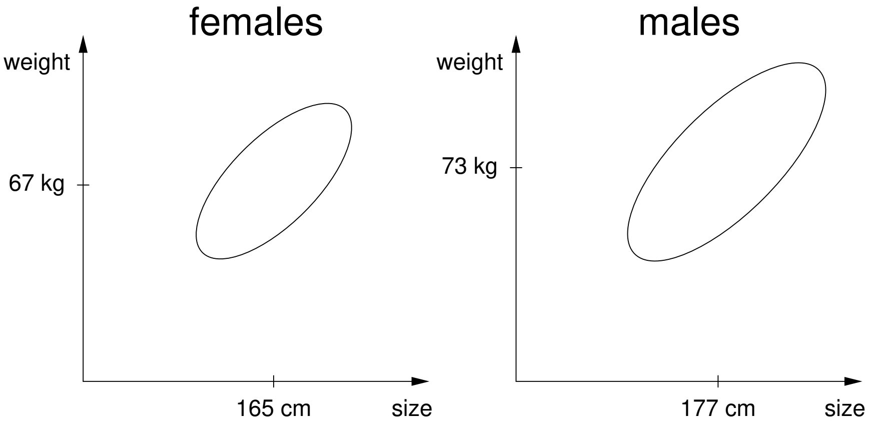
1.2 Partial evidence

1.3 Expectation values

1.4 Continuous random variables

1.4.1 Exercise: Probability densities

Let w, s, and G be random variables indicating body weight, size, and gender of a person. Let $p(w, s|G = f)$ and $p(w, s|G = m)$ be the conditional probability densities over weight and size for females (f) and males (m), respectively, in the shape of Gaussians tilted by 45°, see figure.



1. What is the probability $P(w = 50\text{kg}, s = 156\text{cm}|G = f)$?

   Solution: The probability that the weight is exactly 50kg and the size exactly 156cm is zero.

2. What is the probability $P(w \in [49\text{kg}, 51\text{kg}], s \in [154\text{cm}, 158\text{cm}]|G = f)$?

   Hint: You don't need to calculate a value here. Give an equation.

   Solution: This probability can be calculated by integrating the probability densities over the respective intervals.

$$P(w \in [49\text{kg}, 51\text{kg}], s \in [154\text{cm}, 158\text{cm}]|G = f) = \int_{49}^{51} \int_{154}^{158} p(w, s|G = f) \, \mathrm{d}s \, \mathrm{d}w \,. \tag{1}$$

3. Are weight and size statistically independent? Explain your statement.

   Solution: No, tall persons are typically heavier than short persons.

4. Can you derive variables that are statistically independent?

   Solution: Yes, for instance the weighted sum of weight and size in the principal direction of the Gaussians is statistically independent of the weighted sum orthogonal to that direction.

1.4.2 Exercise: Maximum likelihood estimate

Given N independent measurements $x_1, \ldots, x_N$. As a model of this data we assume the Gaußian distribution

$$p(x) := \frac{1}{\sqrt{2\pi\sigma^2}} \, \mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

1. Determine the probability density function $p(x_1, \ldots, x_N)$ of the set of data points $x_1, \ldots, x_N$ given the two parameters μ, σ^2.

   Solution: The probability density function is simply the product of the probabilities of the individual data points, which is given by the Gaußian.

$$\begin{aligned} p(x_1, \ldots, x_N) &= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \, \mathrm{e}^{-\frac{(x_i-\mu)^2}{2\sigma^2}} & (1) \\ &= \frac{1}{(\sqrt{2\pi\sigma^2})^N} \, \mathrm{e}^{-\frac{\sum_{i=1}^{N}(x_i-\mu)^2}{2\sigma^2}} \,. & (2) \end{aligned}$$

2. Determine the natural logarithm (i.e. ln) of the probability density function of the data given the parameters.

   Solution:

$$\begin{aligned} \ln(p(x_1, \ldots, x_N)) &= \ln\left(\frac{1}{(\sqrt{2\pi\sigma^2})^N} \, \mathrm{e}^{-\frac{\sum_{i=1}^{N}(x_i-\mu)^2}{2\sigma^2}} \right) & (3) \\ &= -\frac{\sum_{i=1}^{N}(x_i-\mu)^2}{2\sigma^2} - \frac{N}{2} \ln\left(2\pi\sigma^2\right) \,. & (4) \end{aligned}$$

3. Determine the optimal parameters of the model, i.e. the parameters that would maximize the probability density determined above. It is equivalent to maximize the logarithm of the pdf (since it is a strictly monotonically increasing function).

   Hint: Calculate the derivative wrt the parameters (i.e. $\frac{\partial}{\partial\mu}$ and $\frac{\partial}{\partial(\sigma^2)}$).

Solution: Taking the derivatives and setting them to zero yields

$$
\begin{aligned}
0 \overset{!}{=} \frac{\partial \ln(p(x_1, \ldots, x_N))}{\partial \mu} \quad &\overset{(4)}{=} \quad \frac{\sum_{i=1}^{N}(x_i - \mu)}{\sigma^2} && (5)\\
\iff \qquad \mu \quad &= \quad \frac{1}{N}\sum_{i=1}^{N} x_i\,, && (6)\\
0 \overset{!}{=} \frac{\partial \ln(p(x_1, \ldots, x_N))}{\partial (\sigma^2)} \quad &\overset{(4)}{=} \quad \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{2\sigma^4} - \frac{N}{2}\frac{2\pi}{2\pi\sigma^2} && (7)\\
\iff \qquad \sigma^2 \quad &= \quad \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2\,. && (8)
\end{aligned}
$$

Interestingly, μ becomes simply the mean and σ^2 the variance of the data.

Extra question: Why is often a factor of $1/(N-1)$ used instead of $1/N$ in the estimate of the variance?

1.4.3 Exercise: Medical diagnosis

Imagine you go to a doctor for a check up to determine your health status H, i.e. whether you are sick ($H = \text{sick}$) or well ($H = \text{well}$). The doctor takes a blood sample and measures a critical continuous variable B. The probability distribution of the variable depends on your health status and is denoted by $p(B|H)$, concrete values are denoted by b. The *a priori* probability for you to be sick (or well) is indicated by $P(H)$.

In the following always arrive at equations written entirely in terms of $p(B|H)$ and $P(H)$. B and H may, of course, be replaced by concrete values.

1. What is the probability that you are sick before you go to the doctor?

   Solution: If you do not know anything about B, the probability that you are sick is obviously the *a priori* probability $P(H = \text{sick})$.

2. If the doctor has determined the value of B, i.e. $B = b$, what is the probability that you are sick? In other words, determine $P(H = \text{sick}|B = b)$.

   Solution: We use Bayesian theory.

$$
\begin{aligned}
P(H = \text{sick}|B = b) \quad &= \quad \frac{p(b|\text{sick})P(\text{sick})}{p(b)} && (1)\\
&= \quad \frac{p(b|\text{sick})P(\text{sick})}{p(b|\text{sick})P(\text{sick}) + p(b|\text{well})P(\text{well})}\,. && (2)
\end{aligned}
$$

3. Assume the doctor diagnoses that you are sick if $P(H = \text{sick}|B = b) > 0.5$ and let D be the variable indicating the diagnosis. Given b, what is the probability that you are being diagnosed as being sick. In other words, determine $P(D = \text{sick}|B = b)$.

   Solution: For a given b one can calculate a concrete probability for being sick, i.e. $P(H = \text{sick}|B = b)$. Now since the diagnosis is deterministic we have

$$
P(D = \text{sick}|B = b) = \Theta(P(H = \text{sick}|B = b) - 0.5)\,, \tag{3}
$$

   with the Heaviside step function defined as $\Theta(x) = 0$ if $x < 0$, $\Theta(x) = 0.5$ if $x = 0$, and $\Theta(x) = 1$ otherwise.

4. Before having any concrete value b, e.g. before you go to the doctor, what is the probability that you will be diagnosed as being sick even though you are well? In other words, determine $P(D = \text{sick}|H = \text{well})$.

Solution: We use the rule for the total probability.

$$
\begin{aligned}
& P(D = \text{sick}|H = \text{well}) \\
= & \int P(D = \text{sick}|b)\, p(b|H = \text{well})\, \mathrm{d}b && (4) \\
\overset{(3)}{=} & \int \Theta(P(H = \text{sick}|b) - 0.5)\, p(b|H = \text{well})\, \mathrm{d}b && (5) \\
\overset{(2)}{=} & \int \Theta\left(\frac{p(b|\text{sick})P(\text{sick})}{p(b|\text{sick})P(\text{sick}) + p(b|\text{well})P(\text{well})} - 0.5\right) p(b|H = \text{well})\, \mathrm{d}b\,. && (6)
\end{aligned}
$$

Extra question: Here we have seen how to turn a continuous random variable into a descrete random variable. How do you describe the distribution of a continuous random variable that deterministically assumes a certain value?

1.4.4 Exercise: Bayesian analysis of a face recognition system

You run a company that sells a face recognition system called ’faceIt’. It consists of a camera and a software system that controls an entry door. If a person wants to identify himself to the system he gets a picture taken with the camera, the probe picture, and that picture is compared to a gallery of stored pictures, the gallery pictures. FaceIt gives a scalar score value between 0 and 1 for each comparison. If the probe picture and the gallery picture show the same person, the score is distributed like

$$p(s|\text{same}) = \alpha_s \exp(\lambda_s s)\,, \tag{1}$$

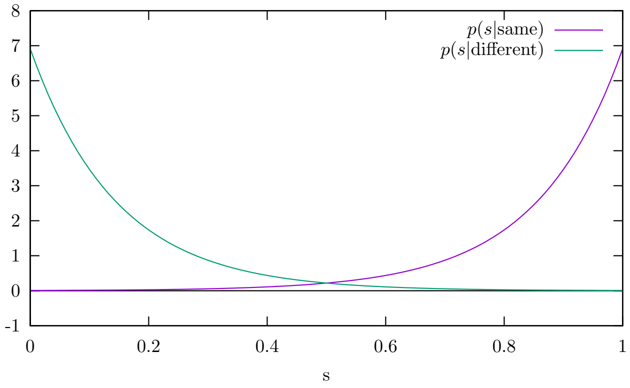
if they show different persons, the score is distributed like

$$p(s|\text{different}) = \alpha_d \exp(-\lambda_d s) \tag{2}$$

with some positive decay constants $\lambda_{\{s,d\}}$ and suitable normalization constants $\alpha_{\{s,d\}}$. All comparisons shall be independent of each other and the score depends only on whether the probe picture and the gallery picture show the same person or not.

1. Draw the score distributions and provide an intuition for why these pdfs might be reasonable.

   Solution:



   The pdfs are reasonable because if the persons are identical you get high probability densities for high scores and if the persons are identical you get high probability densities for low scores.

2. Determine the normalization constants $\alpha_{\{s,d\}}$ for given decay constants $\lambda_{\{s,d\}}$. First give general formulas and then calculate concrete values for $\lambda_s = \lambda_d = \ln(1000)$

Solution: The normalization constant can be derived from the condition that the probability density function integrated over the whole range should be 1.

$$\begin{aligned}
1 &\overset{!}{=} \int_0^1 p(s|\text{same})\,\mathrm{d}s && (3)\\
&= \int_0^1 \alpha_s \exp(\lambda_s s)\,\mathrm{d}s = \alpha_s \left[\frac{1}{\lambda_s}\exp(\lambda_s s)\right]_0^1 = \frac{\alpha_s}{\lambda_s}(\exp(\lambda_s) - 1) && (4)\\
\iff \qquad \alpha_s &= \frac{\lambda_s}{\exp(\lambda_s) - 1} = \frac{\ln(1000)}{1000 - 1} \approx 0.00691\,, && (5)\\
\text{and similarly} \qquad \alpha_d &= \frac{-\lambda_d}{\exp(-\lambda_d) - 1} = \frac{\ln(0.001)}{0.001 - 1} \approx 6.91\,. && (6)
\end{aligned}$$

With these constants we find that the two pdfs are actually mirrored versions of each other, since the exponents are the negative of each other and the normalization scales them to equal amplitude.

3. What is the probability that the score s is less (or greater) than a threshold θ if the probe picture and the gallery picture show the same person, and what if they show different persons. First give general formulas and then calculate concrete values for $\theta = 0.5$.

   Solution: This is straight forward. One simply has to integrate from the threshold to the upper or lower limit of the probability distribution.

$$\begin{aligned}
P(s < \theta|\text{same}) &= \int_0^\theta p(s|\text{same})\,\mathrm{d}s = \alpha_s \left[\frac{1}{\lambda_s}\exp(\lambda_s s)\right]_0^\theta && (7)\\
&= \frac{\alpha_s}{\lambda_s}(\exp(\lambda_s\theta) - 1) = \frac{\exp(\lambda_s\theta) - 1}{\exp(\lambda_s) - 1} \approx 0.031\,, && (8)\\
P(s > \theta|\text{same}) &= 1 - P(s < \theta|\text{same}) = \frac{(\exp(\lambda_s) - 1) - (\exp(\lambda_s\theta) - 1)}{\exp(\lambda_s) - 1} && (9)\\
&= \frac{\exp(\lambda_s) - \exp(\lambda_s\theta)}{\exp(\lambda_s) - 1} \approx 0.969\,, && (10)\\
P(s < \theta|\text{different}) &= \frac{\alpha_d}{-\lambda_d}(\exp(-\lambda_d\theta) - 1) = \frac{\exp(-\lambda_d\theta) - 1}{\exp(-\lambda_d) - 1} \approx 0.969\,, && (11)\\
P(s > \theta|\text{different}) &= 1 - P(s < \theta|\text{different}) = \frac{\exp(-\lambda_d) - \exp(-\lambda_d\theta)}{\exp(-\lambda_d) - 1} \approx 0.031\,. && (12)
\end{aligned}$$

   Notice that due to the finite probability densities it does not make any difference whether we write $<$ and $>$ or $<$ and $\geq$ or $\leq$ and $>$ in the probabilites on the lhs (left hand side).

4. Assume the gallery contains N pictures of N different persons (one picture per person). If N concrete score values $s_i, i = 1, ..., N$, are given and sorted to be in increasing order. What is the probability that gallery picture j shows the correct person? Assume that the probe person is actually in the gallery and that the *a priori* probability for all persons is the same. Give a general formula and calculate a concrete value for $N = 2$ and $s_1 = 0.3$ and $s_2 = 0.8$, and for $s_1 = 0.8$ and $s_2 = 0.9$ if $j = 2$.

Solution: This can be solved directly with Bayes' theorem.

$$\begin{aligned}
& P(\text{same}_j, \text{different}_{i\neq j}|s_1, ..., s_N) \qquad (13)\\
=\ & \frac{p(s_1, ..., s_N|\text{same}_j, \text{different}_{i\neq j})P(\text{same}_j, \text{different}_{i\neq j})}{p(s_1, ..., s_N)} \qquad (14)\\
=\ & \frac{p(s_j|\text{same}) \left(\prod_{i\neq j} p(s_i|\text{different})\right) (1/N)}{p(s_1, ..., s_N)} \qquad (15)\\
& \quad \text{(since the scores are independent of each other and}\\
& \quad \text{the } \textit{a priory} \text{ probability is the same for all gallery images)}\\
\overset{(1,2)}{=}\ & \frac{\alpha_s \exp(\lambda_s s_j) \left(\prod_{i\neq j} \alpha_d \exp(-\lambda_d s_i)\right) (1/N)}{p(s_1, ..., s_N)} \qquad (16)\\
=\ & \frac{\alpha_s \alpha_d^{N-1} \exp(\lambda_s s_j) \exp(\lambda_d(s_j - S))(1/N)}{p(s_1, ..., s_N)} \qquad (17)\\
& \quad \left(\text{with } S := \sum_i s_i\right) \qquad (18)\\
=\ & \frac{\alpha_s \alpha_d^{N-1} \exp((\lambda_s + \lambda_d)s_j) \exp(-\lambda_d S)(1/N)}{p(s_1, ..., s_N)} \qquad (19)\\
=\ & \frac{\alpha_s \alpha_d^{N-1} \exp((\lambda_s + \lambda_d)s_j) \exp(-\lambda_d S)(1/N)}{\sum_{j'} \alpha_s \alpha_d^{N-1} \exp((\lambda_s + \lambda_d)s_{j'}) \exp(-\lambda_d S)(1/N)} \qquad (20)\\
& \quad \text{(since } P \text{ must be normalized to 1)}\\
=\ & \frac{\exp((\lambda_s + \lambda_d)s_j)}{\sum_{j'} \exp((\lambda_s + \lambda_d)s_{j'})}\,. \qquad (21)
\end{aligned}$$

This is a neat formula. It is instructive to calculate the ratio between the probabilty that j is the correct gallery image and that k is the correct gallery image, which is

$$\begin{aligned}
\frac{P(\text{same}_j, \text{different}_{i\neq j}|s_1, ..., s_N)}{P(\text{same}_k, \text{different}_{i\neq k}|s_1, ..., s_N)} \overset{(21)}{=}\ & \frac{\exp((\lambda_s + \lambda_d)s_j)}{\exp((\lambda_s + \lambda_d)s_k)} \qquad (22)\\
=\ & \exp((\lambda_s + \lambda_d)(s_j - s_k))\,. \qquad (23)
\end{aligned}$$

We see that the ratio only depends on the difference between the score values but not on the values themselves. Thus, for the two examples given above it is clear that gallery image 2 is more likely the correct one in the first example even though its absolute score value is greater in the second example. We can verify this by calculating the actual probabilities.

$$\begin{aligned}
P(\text{different}_1, \text{same}_2|s_1 = 0.3, s_2 = 0.8) \overset{(21)}{=}\ & \frac{\exp(2\ln(1000)0.8)}{\exp(2\ln(1000)0.3) + \exp(2\ln(1000)0.8)} \qquad (24)\\
\approx\ & \frac{63096}{63159} \approx 0.9990\,, \qquad (25)\\
P(\text{different}_1, \text{same}_2|s_1 = 0.8, s_2 = 0.9) \overset{(21)}{=}\ & \frac{\exp(2\ln(1000)0.9)}{\exp(2\ln(1000)0.8) + \exp(2\ln(1000)0.9)} \qquad (26)\\
\approx\ & \frac{251189}{314284} \approx 0.7992\,. \qquad (27)
\end{aligned}$$

5. Without any given concrete score values, what is the probability that a probe picture of one of the persons in the gallery is recognized correctly if one simply picks the gallery picture with the highest score as the best guess for the person to be recognized. Give a general formula.

   Solution: The probability of correct recognition is the probability density that the correct gallery picture gets a certain score s', i.e. $p(s'|\text{same})$, times the probability that all the other gallery pictures get score below s', i.e. $P(s < s'|\text{different})^{(N-1)}$, integrated over all possible scores s'. Note that the integration turns the probability density $p(s'|\text{same})$ into a proper probability.

$$
\begin{aligned}
& P(\text{correct recognition}) && \qquad (28)\\
= \; & \int_0^1 p(s'|\text{same}) P(s < s'|\text{different})^{(N-1)} \, \mathrm{d}s' && \qquad (29)\\
= \; & \int_0^1 \alpha_s \exp(\lambda_s s') \left(\frac{\exp(-\lambda_d s') - 1}{\exp(-\lambda_d) - 1} \right)^{(N-1)} \mathrm{d}s' && \qquad (30)\\
= \; & \int_0^1 \frac{\lambda_s \exp(\lambda_s s')}{\exp(\lambda_s) - 1} \left(\frac{\exp(-\lambda_d s') - 1}{\exp(-\lambda_d) - 1} \right)^{(N-1)} \mathrm{d}s' \quad \text{(this is ok as a solution)} && \qquad (31)\\
= \; & \underbrace{\frac{\lambda_s}{(\exp(\lambda_s) - 1)(\exp(-\lambda_d) - 1)^{(N-1)}}}_{=:A} \int_0^1 \exp(\lambda_s s')(\exp(-\lambda_d s') - 1)^{(N-1)} \, \mathrm{d}s' && \qquad (32)\\
= \; & ... \quad \text{(one could simplify even further)} && \qquad (33)
\end{aligned}
$$

1.5 A joint as a product of conditionals

1.6 Marginalization

1.7 Application: Visual attention +

2 Inference in Bayesian networks

2.1 Graphical representation

2.2 Conditional (in)dependence

2.2.1 Exercise: Conditional independences

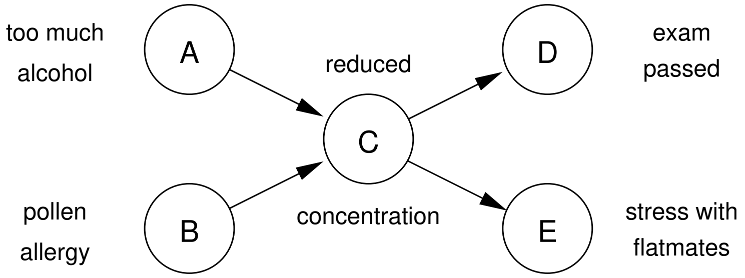
Consider the following scenario:

Imagine you are at a birthday party of a friend on Sunday and you have an exam on Monday. If you drink too much alcolhol at the birthday party, you most likely have problems concentrating the next day, which would reduce the probability that you pass the exam. Another consequence of the reduced concentration might be increased stress with your flatmates, because, e.g., you forget to turn off the radio or the stove. Lack of concentration might also be caused by your pollen allergy, which you suffer from on some summer days.

Consider the following random variables that can assume the values “true” or “false”: A: drinking too much alcolhol on Sunday; B: pollen allergy strikes; C: reduced concentration on Monday; D: you pass the exam; E: stress with your flatmates.

Search for conditional dependencies and independencies.

Solution: The corresponding Bayesian network is



It is obvious that A and B are conditionally dependent if C or any of its descendents, i.e. D or E, have received evidence. Furthermore, D (or E) is conditionally independent of A, B, and E (or D) if C is instantiated. D and E are conditionally dependent if C is not instantiated.

2.3 Inverting edges

2.4 d-Separation in Bayesian networks

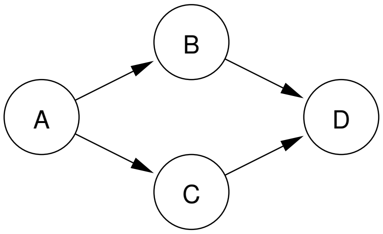
2.4.1 Exercise: Bayesian networks

Consider the factorized joint probability

$$P(\overline{A}, \underline{B}, \underline{C}, D) = P(D|B, C)P(C|A)P(B|A)P(A)\,. \tag{1}$$

1. Draw the corresponding Bayesian network.

   Solution: The factorization corresponds to the following network.



2. By which given and missing evidence can the two underscored variables be d-separated? Prove one d-separation also analytically.

   Solution: B and C are d-separated if A is instantiated and D has received no evidence.

$$\begin{aligned}
P(B, C|A) &= \frac{P(A, B, C)}{P(A)} && (2)\\
&= \frac{\sum_d P(A, B, C, d)}{P(A)} && (3)\\
&\overset{(1)}{=} \frac{\sum_d P(d|B, C)P(C|A)P(B|A)P(A)}{P(A)} && (4)\\
&= \underbrace{\left(\sum_d P(d|B, C)\right)}_{=1} P(C|A)P(B|A) && (5)\\
&= P(B|A)P(C|A)\,. && (6)
\end{aligned}$$

Alternatively one could have shown $P(B|A,C) = P(B|A)$, which is equivalent.

If D is also known, then the d-separation does not work.

$$
\begin{aligned}
P(B,C|A,D) &= \frac{P(A,B,C,D)}{P(A,D)} && (7)\\
&\overset{(1)}{=} \frac{P(D|B,C)P(C|A)P(B|A)P(A)}{P(D|A)P(A)} && (8)\\
&= \frac{P(D|B,C)P(C|A)P(B|A)}{P(D|A)} && (9)\\
&= \ ? && (10)
\end{aligned}
$$

There is no further simplification possible. Or do you find one?

3. Can the message passing algorithm be applied to the original network? If yes, write down which messages have to be passed how in order to compute the marginal probability of the overscored variable. Compare the result with the one that you get, when you calculate the marginal from the factorized joint probability.

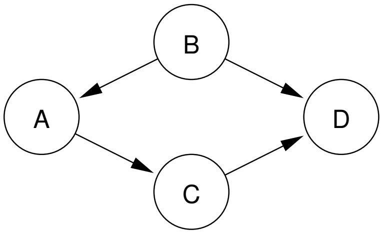
   Solution: Message passing does not work because D has two parants.

4. Invert as many edges as possible without adding new edges. Write down the new factorization of the joint probability in terms of conditionals and draw the corresponding Bayesian network.

   Solution: Only one edge can be inverted, for instance

$$
\begin{aligned}
P(A,B,C,D) &= P(D|B,C)P(C|A)P(B|A)P(A) && (11)\\
&= P(D|B,C)P(C|A)P(A|B)P(B)\,, && (12)
\end{aligned}
$$

which corresponds to the network



One could have also inverted $A \to C$.

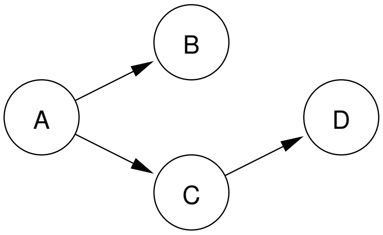
2.4.2 Exercise: Bayesian networks

Consider the factorized joint probability

$$P(A, \underline{B}, \overline{C}, \underline{D}) = P(D|C)P(C|A)P(B|A)P(A)\,. \qquad (1)$$

1. Draw the corresponding Bayesian network.

   Solution: Die Faktorisierung entspricht folgendem Netzwerk.



2. By which given and missing evidence can the two underscored variables be d-separated? Prove one d-separation also analytically.

   Solution: B und D werden d-separiert, wenn A bekannt oder C bekannt ist. Für bekanntes A erhalten wir

$$\begin{aligned} P(B,D|A) &= \frac{P(A,B,D)}{P(A)} && (2)\\ &= \frac{\sum_c P(A,B,c,D)}{P(A)} && (3)\\ &\overset{(1)}{=} \frac{\sum_c P(D|c)P(c|A)P(B|A)P(A)}{P(A)} && (4)\\ &= \underbrace{\left(\sum_c P(D|c)P(c|A)\right)}_{=P(D|A)} P(B|A) && (5)\\ &= P(B|A)P(D|A)\,. && (6) \end{aligned}$$

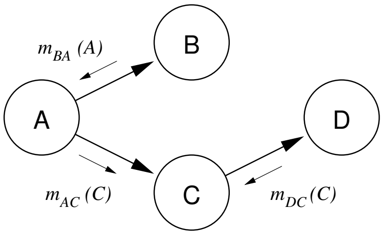
   Alternativ hätte man z.B. auch $P(B|A,D) = P(B|A)$ zeigen können, was äquivalent ist.

   Ist C bekannt, ergibt sich

$$\begin{aligned} P(B,D|C) &= \frac{P(B,C,D)}{P(C)} && (7)\\ &= \frac{\sum_a P(a,B,C,D)}{P(C)} && (8)\\ &\overset{(1)}{=} \frac{\sum_a P(D|C)P(C|a)P(B|a)P(a)}{P(C)} && (9)\\ &= \frac{\sum_a P(D|C)P(B|a)P(a|C)P(C)}{P(C)} && (10)\\ &= P(D|C) \underbrace{\left(\sum_a P(B|a)P(a|C)\right)}_{=P(B|C)} && (11)\\ &= P(B|C)P(D|C)\,. && (12) \end{aligned}$$

3. Can the message passing algorithm be applied to the original network? If yes, write down which messages have to be passed how in order to compute the marginal probability of the overscored variable. Compare the result with the one that you get, when you calculate the marginal from the factorized joint probability.

   Solution: The message- passing algorithm can be applied since no node has more than one parent.



We get

$$m_{BA}(A) := \sum_b P(b|A) = 1\,, \tag{13}$$

$$m_{AC}(C) := \sum_a \underbrace{m_{BA}(a)}_{=1} P(C|a)P(a) \tag{14}$$

$$= \sum_a P(C|a)P(a)\,, \tag{15}$$

$$m_{DC}(C) := \sum_d P(d|C) = 1\,, \tag{16}$$

$$P(C) = m_{AC}(C)\,\underbrace{m_{DC}(C)}_{=1} \tag{17}$$

$$= \sum_a P(C|a)P(a)\,. \tag{18}$$

Direct marginalization over the factorized joint probability leads to the same result.

$$P(C) = \sum_{a,b,d} P(a,b,C,d) \tag{19}$$

$$\overset{(1)}{=} \sum_{a,b,d} P(d|C)P(C|a)P(b|a)P(a) \tag{20}$$

$$= \underbrace{\left(\sum_d P(d|C)\right)}_{m_{DC}(C)=1} \underbrace{\sum_a P(C|a) \underbrace{\left(\sum_b P(b|a)\right)}_{m_{BA}(A)=1} P(a)}_{m_{AC}(C)} \tag{21}$$

$$= \sum_a P(C|a)P(a)\,. \tag{22}$$

Of course, this would have been more interesting if partial evidences were given.

4. Invert as many edges as possible (without adding new edges). Write down the new factorization of the joint probability in terms of conditionals and draw the corresponding Bayesian network.

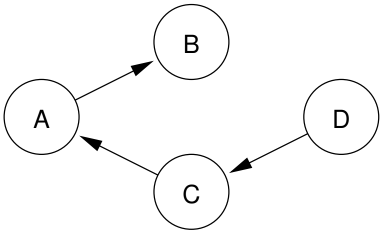
   Solution: Here at most two edges can be inverted

$$P(A,B,C,D) \overset{(1)}{=} P(D|C)P(C|A)P(B|A)P(A) \tag{23}$$

$$= P(D|C)P(B|A)P(A|C)P(C) \tag{24}$$

$$= P(B|A)P(A|C)P(C|D)P(D) \tag{25}$$

corresponding to the following network



2.5 Calculating marginals and message passing in Bayesian networks

2.6 Message passing with evidence

2.6.1 Exercise: Partial evidence in a Bayesian network

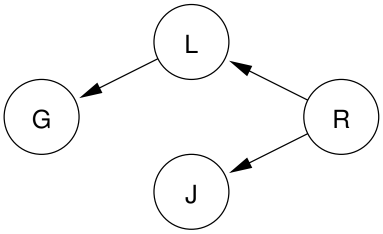
Imagine you are working in the human resources department of a big company and have to hire people. You get to screen a lot of applications and have arrived at the following model: Whether a person will perform well on the job (random variable J) depends on whether he is responsible and well organized (random variable R). The latter will, of course, also influence how well the person has learned during his studies (random variable L) and that in turn has had an impact on his grades (random variable G).

Assume random variable grade G has three possible values 0 (poor), 1 (medium), 2 (good) and all other random variables are binary with 0 indicating the negative and 1 the positive state. For simplicity, use probabilties of 0, 1/3 and 2/3 only, namely:

$P(G=0\|L=0)=2/3$	$P(L=0\|R=0)=2/3$	$P(J=0\|R=0)=2/3$	$P(R=0)=1/3$
$P(G=1\|L=0)=1/3$	$P(L=1\|R=0)=1/3$	$P(J=1\|R=0)=1/3$	
$P(G=2\|L=0)=0$			
$P(G=0\|L=1)=0$	$P(L=0\|R=1)=1/3$	$P(J=0\|R=1)=1/3$	$P(R=1)=2/3$
$P(G=1\|L=1)=1/3$	$P(L=1\|R=1)=2/3$	$P(J=1\|R=1)=2/3$	
$P(G=2\|L=1)=2/3$			

1. Draw a Baysian network for the model and write down the joint probability as a product of conditionals and priors.

   Solution: The Bayesian network is



   The joint is

$$P(G, L, J, R) = P(G|L)P(L|R)P(J|R)P(R)\,. \tag{1}$$

2. Consider $P(J=1)$. What does it mean? Calculate it along with all relevant messages.

   Solution: $P(J=1)$ is the *a priori* probability that any person performs well on the job and can be calculated as follows.

$$\begin{aligned}
P(J=1) &= \sum_{r,l,g} P(g|l)P(l|r)P(J=1|r)P(r) && (2)\\
&= \underbrace{\sum_r \underbrace{\left(\sum_l \underbrace{\left(\sum_g P(g|l)\right)}_{m_{GL}(l)=(\underline{1},\underline{1})} P(l|r)\right)}_{m_{LR}(r)=(\underline{1},\underline{1})} P(J=1|r)P(r)}_{m_{RJ}(J)=(\underline{1}*2/3*1/3+\underline{1}*1/3*2/3,\underline{1}*1/3*1/3+\underline{1}*2/3*2/3)=(\underline{4/9},\underline{5/9})} && (3)\\
&= \underline{5/9} \approx 0.5556\,. && (4)
\end{aligned}$$

3. Consider $P(J = 1|\mathcal{E}_G)$ with $\mathcal{E}_G = \{1, 2\}$. What does it mean? Calculate it along with all relevant messages.

   Solution: This is the probability that a person with medium or good grades performs well on the job. The way to calulate this is to first calcutate the unnormalized probability by dropping the combinations that are not possible anymore and then renormalizing this by the new partition function.

$$P(J=1)|\mathcal{E}_G$$

$$= \sum_{r,l,g\in\mathcal{E}_G} P(g|l)P(l|r)P(J=1|r)P(r) \tag{5}$$

$$= \underbrace{\sum_r \underbrace{\left(\sum_l \underbrace{\sum_{g\in\mathcal{E}_G} P(g|l)}_{m_{GL}(l)=(1/3+0,1/3+2/3)=(\underline{1/3},\underline{3/3})} P(l|r)\right)}_{m_{LR}(r)=(\underline{1/3}*2/3+\underline{3/3}*1/3,\underline{1/3}*1/3+\underline{3/3}*2/3)=(\underline{5/9},\underline{7/9})} P(J=1|r)P(r)}_{m_{RJ}(J)=(\underline{5/9}*2/3*1/3+\underline{7/9}*1/3*2/3,\underline{5/9}*1/3*1/3+\underline{7/9}*2/3*2/3)=(\underline{24/81},\underline{33/81})} \tag{6}$$

$$= \underline{33/81}\,, \tag{7}$$

$$Z = P(J=0)|\mathcal{E}_G + P(J=1)|\mathcal{E}_G \tag{8}$$

$$= \underline{24/81} + \underline{33/81} \tag{9}$$

$$= 57/81\,, \tag{10}$$

$$P(J=1|\mathcal{E}_G)$$

$$= \frac{1}{Z}\,P(J=1)|\mathcal{E}_G \tag{11}$$

$$= \frac{81}{57}\cdot\frac{33}{81} \tag{12}$$

$$= 4/7 \approx 0.5789\,. \tag{13}$$

   That is surprisingly little more than the *a priori* probability.

4. Consider the preceeding case and additionally assume you know the person is responsible and well organized. Calculate the corresponding probability.

   Solution:

There are two ways to calculate this. Firstly, one can do it the same way as above.

$$\begin{aligned}
&P(J=1|R=1)|\mathcal{E}_G \\
&\quad= \sum_{l,g\in\mathcal{E}_G} P(g|l)P(l|R=1)P(J=1|R=1)P(R=1) && (14)\\
&\quad= \underbrace{\underbrace{\left(\sum_l \underbrace{\sum_{g\in\mathcal{E}_G} P(g|l)}_{m_{GL}(l)=(1/3+0,1/3+2/3)=(\underline{1/3},\underline{3/3})} P(l|R=1)\right)}_{m_{LR}(r)=(\underline{1/3}*2/3+\underline{3/3}*1/3,\underline{1/3}*1/3+\underline{3/3}*2/3)=(\underline{5/9},\underline{7/9})} P(J=1|R=1)P(R=1)}_{m_{RJ}(J)=(\underline{7/9}*1/3*2/3,\underline{7/9}*2/3*2/3)=(\underline{14/81},\underline{28/81})} && (15)\\
&\quad= \underline{28/81}\,, && (16)\\
Z &\quad= P(J=0|R=1)|\mathcal{E}_G + P(J=1|R=1)|\mathcal{E}_G && (17)\\
&\quad= \underline{14/81} + \underline{28/81} && (18)\\
&\quad= 42/81\,, && (19)\\
&P(J=1|\mathcal{E}_G, R=1) \\
&\quad= \frac{1}{Z} P(J=1|R=1)|\mathcal{E}_G && (20)\\
&\quad= \frac{81}{42}\cdot\frac{28}{81} && (21)\\
&\quad= 2/3 \approx 0.6667\,. && (22)
\end{aligned}$$

Secondly, one can realize that J is conditionally independent of G given R and simply calculate

$$\begin{aligned}
P(J=1|\mathcal{E}_G, R=1) &= P(J=1|R=1) && (23)\\
&= 2/3 \approx 0.6667\,. && (24)
\end{aligned}$$

This now is significantly more than the *a priori* probability. This suggest that it might be much more important to figure out whether a person is responsible and well organized than whether he has good grades. The trouble is, that it is also much more difficult.

2.7 Application: Image super-resolution

2.7.1 Exercise: Everyday example

Make up an example from your everyday life or from the public context for which you can develop a Bayesian network. First draw a network with at least ten nodes. Then consider a subnetwork with at least three nodes that you can still treat explicitly. Define the values the different random variables can assume and the corresponding *a priori* and conditional probabilities. Try to calculate some non-trivial conditional probability with the help of Bayesian theory. For instance, you could invert an edge and calculate a conditional probability in the acausal direction, thereby determining the influence of a third variable. Be creative.

Solution: Not available!

3 Inference in Gibbsian networks

3.1 Introduction

3.2 Moral graph

3.3 Junction tree and message passing in a Gibbsian network

3.4 Most probable explanation

3.4.1 Exercise: Most probable explanation

Consider the network of Exercise 2.6.1. What is the most probable explanation for a person to perform well on the job?

Solution:

$$
\begin{aligned}
& \max_{r,l,g} P(g,l,J=1,r) && (1)\\
=\ & \max_{r,l,g} P(g|l)P(l|r)P(J=1|r)P(r) && (2)\\
=\ & \underbrace{\max_r \underbrace{\left(\max_l \underbrace{\left(\max_g P(g|l)\right)}_{e_{GL}(l) = (2/3 \text{ for } G{=}0|L{=}0,\, 2/3 \text{ for } G{=}1|L{=}1)} P(l|r)\right)}_{e_{LR}(r) = (2/3{*}2/3{=}4/9 \text{ for } L{=}0|R{=}0,\, 2/3{*}2/3{=}4/9 \text{ for } L{=}1|R{=}1)} P(J=1|r)P(r)}_{e_{RJ}(J) = (4/9{*}2/3{*}1/3{=}8/81 \text{ for } R{=}0|J{=}0 \text{ or } 4/9{*}1/3{*}2/3{=}8/81 \text{ for } R{=}1|J{=}0,\, 4/9{*}2/3{*}2/3{=}16/81 \text{ for } R{=}1|J{=}1)} && (3)\\
=\ & 16/81 \approx 0.1975\,. && (4)
\end{aligned}
$$

Notice that we have calculated $P(g,l,J=1,r)$ of the most probable state here and not $P(g,l,r|J=1)$, which might actually be more interesting. For the latter we would have to renormalize the probability given the evidence $J=1$, but this has not been asked for. We need the state itself, which we can get by tracing back the most probable state. $J=1$ is given. In the messages we can see that from that follows $R=1$, from that $L=1$, and from that finally $G=1$. Thus, the most probable explanation for somebody to perform well on the job is to assume that he is responsible and well organized, did learn well, and got good grades. It is quite clear that the person does not perform well on the job, because he got good grades, but assuming good grades is more plausible than assuming poor grades. This is the difference between causal and probabilistic reasoning.

3.5 Triangulated graphs

3.5.1 Exercise: Gibbsian network

For the Bayesian network given below solve the following exercises:

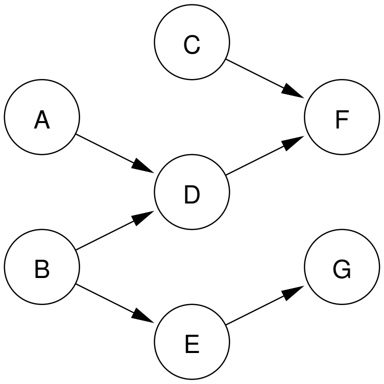
1. Write the joint probability as a product of conditional probabilities according to the network structure.
2. Draw the corresponding Gibbsian network.
3. Invent a Bayesian network that is consistent with the Gibbsian network but inconsistent with the original Bayesian network in that it differs from it in at least one conditional (in)dependence. Illustrate this difference.
4. List all cliques of the Gibbsian network and construct its junction tree. If necessary, use the triangulation method.

5. Define a potential function for each node of the junction tree, so that their product yields the joint probability. Make sure each potential function formally contains all variables of the node even if it does not explicitly depend on it. Provide suitable expressions for the potential functions in terms of conditional probabilities.

6. Give formulas for all messages in the network and simplify them as far as possible. Can you give a heuristics when a message equals 1.

7. Use the message passing algorithm for junction trees to compute the marginal probability $P(B)$. Verify that the result is correct.

Bayesian network:

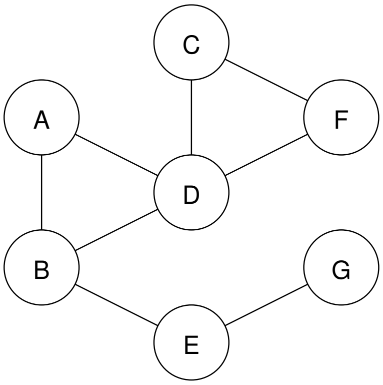


Solution:

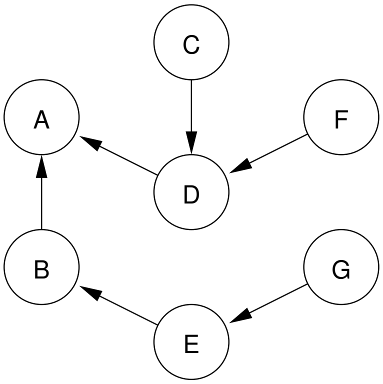
ad 1: From the structure of the network we derive the following factorization of the joint probability:

$$P(A,B,C,D,E,F,G) = P(G|E)P(E|B)P(F|C,D)P(C)P(D|A,B)P(B)P(A)\,. \tag{1}$$

ad 2: To draw the corresponding Gibbsian network we have to marry parents of common children and remove the arrow heads. This yields the following network.

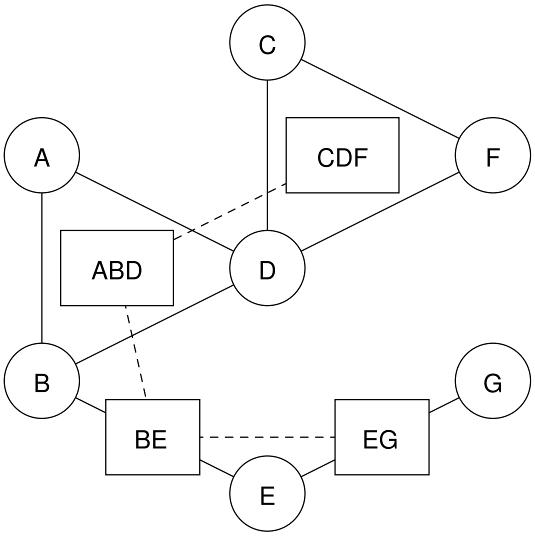


ad 3: There is usually a cheap solution, where one simply inverts an edge of the Bayesian network or adds one, and a more diffcult one, where one designs a completely new Bayesian network. The figure shows one of the latter type.



Extra question: How many different ways are there to create a Bayesian network consistent with the Gibbsian network?

ad 4: The cliques of the Gibbsian network are $\{A, B, D\}$, $\{C, D, F\}$, $\{B, E\}$, $\{E, G\}$. A possible junction tree looks as follows:



ad 5: One possible set of potential functions that reproduces the joint probability is

$$\phi(A, B, D) := P(D|A, B)P(B)P(A)\,, \tag{2}$$

$$\phi(C, D, F) := P(F|C, D)P(C)\,, \tag{3}$$

$$\phi(B, E) := P(E|B)\,, \tag{4}$$

$$\phi(E, G) := P(G|E)\,, \tag{5}$$

$$P(A, B, C, D, E, F, G) \overset{(1)}{=} P(G|E)P(E|B)P(F|C, D)P(C)P(D|A, B)P(B)P(A) \tag{6}$$

$$\overset{(2-5)}{=} \phi(A, B, D)\phi(C, D, F)\phi(B, E)\phi(E, G)\,. \tag{7}$$

ad 6: The messages are:

$$m_{CDF,ABD}(D) := \sum_{c,f} \phi(c,D,f) \tag{8}$$

$$\overset{(3)}{=} \sum_{c} \underbrace{\sum_{f} P(F|C,D)}_{=1} P(C) \;=\; 1 \tag{9}$$

$$m_{ABD,BE}(B) := \sum_{a,d} m_{CDF,ABD}(d)\, \phi(a,B,d) \tag{10}$$

$$\overset{(9,2)}{=} \underbrace{\sum_{a} P(a)}_{=1} \underbrace{\sum_{d} P(d|a,B)}_{=1} P(B) \;=\; P(B) \tag{11}$$

$$m_{BE,EG}(E) := \sum_{b} m_{ABD,BE}(b)\, \phi(b,E) \tag{12}$$

$$\overset{(11,4)}{=} \sum_{b} P(b) P(E|b)\,, \quad \text{(no simplification)} \tag{13}$$

$$m_{EG,BE}(E) := \sum_{g} \phi(E,g) \tag{14}$$

$$\overset{(5)}{=} \underbrace{\sum_{g} P(g|E)}_{=1} \;=\; 1 \tag{15}$$

$$m_{BE,ABD}(B) := \sum_{e} m_{EG,BE}(e)\, \phi(B,e) \tag{16}$$

$$\overset{(15,4)}{=} \sum_{e} P(e|B) \;=\; 1\,, \tag{17}$$

$$m_{ABD,CDF}(D) := \sum_{a,b} m_{BE,ABD}(b)\, \phi(a,b,D) \tag{18}$$

$$\overset{(17,2)}{=} \sum_{a,b} P(D|a,b) P(b) P(a)\,. \quad \text{(no simplification)} \tag{19}$$

Messages often equal 1 if they go against the arrows of the Bayesian network.

Extra question: What is the nature of the messages?

ad 7: To calculate the marginal probability $P(B)$, we have to multiply all incoming messages into a node containing B with its potential function and then sum over all random variables of that node, except B.

$$P(B) = \sum_{e} m_{ABD,BE}(B)\, m_{EG,BE}(e)\, \phi(B,e) \tag{20}$$

(this is the result, now follows the verification)

$$\overset{(11,15,4)}{=} \underbrace{\sum_{e} P(e|B)}_{=1} P(B) \;=\; P(B)\,. \tag{21}$$

3.5.2 Exercise: Gibbsian network

For the Bayesian network given below solve the following exercises:

1. Write the joint probability as a product of conditional probabilities according to the network structure.
2. Draw the corresponding Gibbsian network.

3. Invent a Bayesian network that is consistent with the Gibbsian network but inconsistent with the original Bayesian network in that it differs from it in at least one conditional (in)dependence. Illustrate this difference.

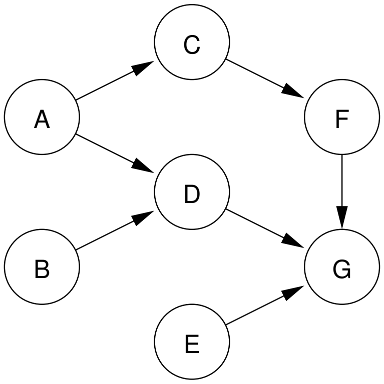
4. List all cliques of the Gibbsian network and construct its junction tree. If necessary, use the triangulation method.

5. Define a potential function for each node of the junction tree, so that their product yields the joint probability. Make sure each potential function formally contains all variables of the node even if it does not explicitly depend on it. Provide suitable expressions for the potential functions in terms of conditional probabilities.

6. Give formulas for all messages in the network and simplify them as far as possible. Can you give a heuristics when a message equals 1.

7. Use the message passing algorithm for junction trees to compute the marginal probability $P(B)$. Verify that the result is correct.

Bayesian network:

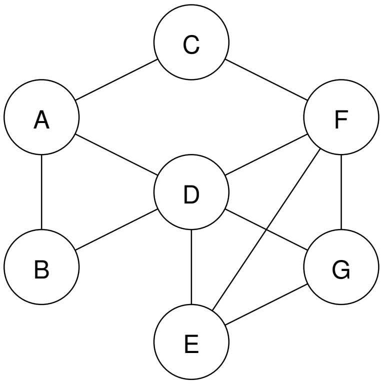


Solution:

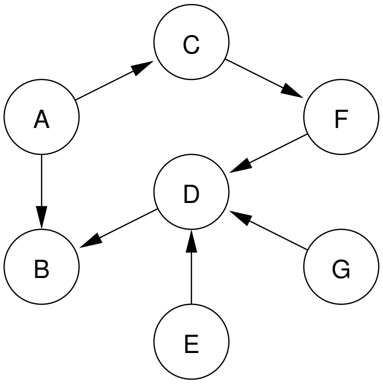
ad 1: From the structure of the network we derive the following factorization of the joint probability:

$$P(A,B,C,D,E,F,G) = P(G|D,E,F)P(E)P(D|A,B)P(B)P(F|C)P(C|A)P(A)\,. \tag{1}$$

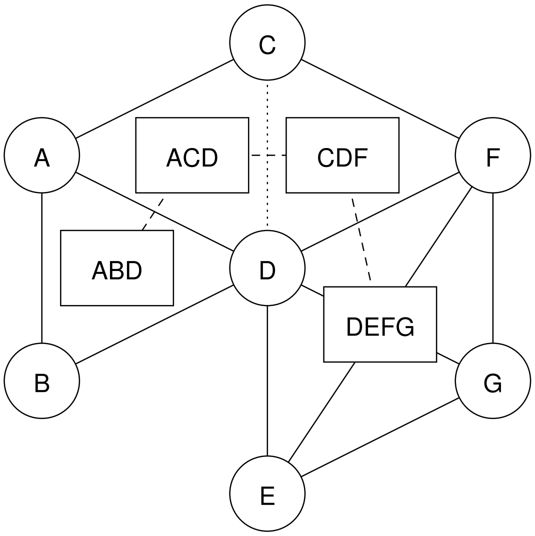
ad 2: To draw the corresponding Gibbsian network we have to marry parents of common children and remove the arrow heads. This yields the following network.



ad 3: There is usually a cheap solution, where one simply inverts an edge of the Bayesian network or adds one, and a more diffcult one, where one designs a completely new Bayesian network. The figure shows one of the latter type.



ad 4: The cliques of the Gibbsian network are $\{A, B, D\}$, $\{A, C\}$, $\{C, F\}$, $\{D, E, F, G\}$. A possible junction tree looks as follows:



ad 5: One possible set of potential functions that reproduces the joint probability is

$$\phi(A, B, D) := P(D|A, B)P(B)P(A)\,, \tag{2}$$

$$\phi(A, C, D) := P(C|A)\,, \tag{3}$$

$$\phi(C, D, F) := P(F|C)\,, \tag{4}$$

$$\phi(D, E, F, G) := P(G|D, E, F)P(E)\,, \tag{5}$$

$$P(A, B, C, D, E, F, G) \overset{(1)}{=} P(G|D, E, F)P(E)P(D|A, B)P(B)P(F|C)P(C|A)P(A) \tag{6}$$

$$\overset{(2\text{–}5)}{=} \phi(A, B, D)\phi(A, C, D)\phi(C, D, F)\phi(D, E, F, G)\,. \tag{7}$$

ad 6: The messages are:

$$\begin{aligned}
m_{ABD,ACD}(A,D) &:= \sum_b \phi(A,b,D) && (8)\\
&\overset{(2)}{=} \sum_b P(D|A,b)P(b)P(A)\,, \quad \text{(no simplification)} && (9)\\
m_{ACD,CDF}(C,D) &:= \sum_a m_{ABD,ACD}(a,D)\,\phi(a,C,D) && (10)\\
&\overset{(9,3)}{=} \sum_a \sum_b P(D|a,b)P(b)P(a)\,P(C|a)\,, \quad \text{(no simplification)} && (11)\\
m_{CDF,DEFG}(D,F) &:= \sum_c m_{ACD,CDF}(c,D)\,\phi(c,D,F) && (12)\\
&\overset{(11,4)}{=} \sum_c \sum_a \sum_b P(D|a,b)P(b)P(a)\,P(c|a)\,P(F|c)\,, && (13)\\
&\qquad \text{(no simplification)} &&\\
m_{DEFG,CDF}(D,F) &:= \sum_{e,g} \phi(D,e,F,g) && (14)\\
&\overset{(5)}{=} \sum_e \underbrace{\sum_g P(g|D,e,F)}_{=1}\,P(e) \;=\; 1\,, && (15)\\
m_{CDF,ACD}(C,D) &:= \sum_f m_{DEFG,CDF}(D,f)\,\phi(C,D,f) && (16)\\
&\overset{(15,4)}{=} \sum_f P(f|C) \;=\; 1\,, && (17)\\
m_{ACD,ABD}(A,D) &:= \sum_c m_{CDF,ACD}(c,D)\,\phi(A,c,D) && (18)\\
&\overset{(17,3)}{=} \sum_c P(c|A) \;=\; 1\,. && (19)
\end{aligned}$$

Messages often equal 1 if they go against the arrows of the Bayesian network.

ad 7: To calculate the marginal probability $P(B)$, we have to multiply all incoming messages into a node containing B with its potential function and then sum over all random variables of that node, except B.

$$\begin{aligned}
P(B) &= \sum_{a,d} m_{ACD,ABD}(a,d)\,\phi(a,B,d) && (20)\\
&\qquad \text{(this is the result, now follows the verification)} &&\\
&\overset{(19,2)}{=} \sum_{a,d} P(d|a,B)P(B)P(a) && (21)\\
&= \underbrace{\sum_a \underbrace{\sum_d P(d|a,B)}_{=1}\,P(a)}_{=1}\,P(B) \;=\; P(B)\,. && (22)
\end{aligned}$$

Extra question: What is the nature of the messages?

4 Approximate inference

4.1 Introduction

4.2 Gibbs sampling

4.2.1 Exercise: Gibbs sampling

Invent a simple example like in the lecture where you can do approximate inference by Gibbs sampling. Either write a little program or do it by hand, e.g., with dice. Compare the estimated with the true probabilities and illustrate your results. Draw a transition diagram.

Solution: Christian Bodenstein (WS'11) came up with the following example inspired by the card game Black Jack. Imagine two cards are drawn from a deck of 32 cards with 4 Aces, 16 cards with a value of 10 (King, Queen, Jack, 10), and 12 cards with value zero (9, 8, 7). The probabilities for the value of the first card are simply 4/32, 16/32, and 12/32. The probabilities for the second card depend on the first card and would be $(4-1)/(32-1) = 3/31$ for an Ace if the first card is an Ace already.

This scenario can be modeled with two random variables L (left card) and R (right card, drawn after the left one), each of which can assume three different values, namely A (Ace), B (value 10), and Z (zero). The Bayesian network would have two nodes with an arrow like $L \rightarrow R$. Viewed as a Gibbsian network with two nodes, one would define the potential function $\phi(L, R) := P(L, R)$. The probabilities $P(L)$, $P(R|L)$, and $P(L, R)$ are given in the table below.

probabilities					samples					
L	R	$P(L)$	$P(R\|L)$	$P(L,R)$	100	500	1000	10000	100000	1000000
A	A	0,125	0,097	0,0121	0,02	0,018	0,009	0,0130	0,0118	0,0120
A	B	0,125	0,516	0,0645	0,04	0,076	0,053	0,0607	0,0652	0,0641
A	Z	0,125	0,387	0,0484	0,06	0,032	0,053	0,0475	0,0510	0,0479
B	A	0,500	0,129	0,0645	0,07	0,060	0,053	0,0655	0,0626	0,0646
B	B	0,500	0,484	0,2419	0,34	0,226	0,234	0,2467	0,2419	0,2423
B	Z	0,500	0,387	0,1935	0,27	0,176	0,214	0,2045	0,1933	0,1938
Z	A	0,375	0,129	0,0484	0,05	0,042	0,042	0,0472	0,0471	0,0487
Z	B	0,375	0,516	0,1935	0,12	0,220	0,167	0,1878	0,1938	0,1934
Z	Z	0,375	0,355	0,1331	0,03	0,150	0,175	0,1271	0,1334	0,1332
accumulated squared error:					0,0323	0,00204	0,0032	0,00022	1,3E-05	6,9E-07

Gibbs sampling now works as follows: First a fair coin is tossed to determine which card to assign a new value to. Then a new value is assigned according to the corresponding conditional probability, either $P(L|R)$ if the left card is considered or $P(R|L)$ if the right card is considered. This process is repeated over and over again and the samples are collected. Finally the joint probabilities $P(L, R)$ are estimated, see right half of the table. The more samples are taken the more precise the estimates, at least in general. We see that from 500 to 1000 the estimate actually gets worse, but that is an accident and not systematic.

$P(R|L)$ is given in the table, but $P(L|R)$ has to be calculated with Bayes theorem,

$$P(L|R) = \frac{P(R|L)P(L)}{P(R)} \,. \tag{1}$$

For symmetry reasons one would expect that $P_{L|R}(l|r) = P_{R|L}(l|r)$ for all values of l and r, but who knows, you might want to check (I didn't). Likewise the sums over colums of $P(L, R)$ should add up to one, for the precise values as well as the estimated ones.

The figure below shows a transition diagram between the different states. Because the order of the cards does not really matter, $L = A, R = B$ and $L = B, R = A$ have been collapsed into one state AB and there is no state BA. If one does that, one has to be careful to double some of the transition probabilities. For instance there are two possibilites to get from two Aces to one Ace and a zero, either from AA to AZ or from AA to ZA. This has to be taken into account. A simple sanity check is to see whether the sum over all outgoing transitions per node is one. The little arcs at the boxes indicate transitions from a state into itself, they don't have arrowheads.

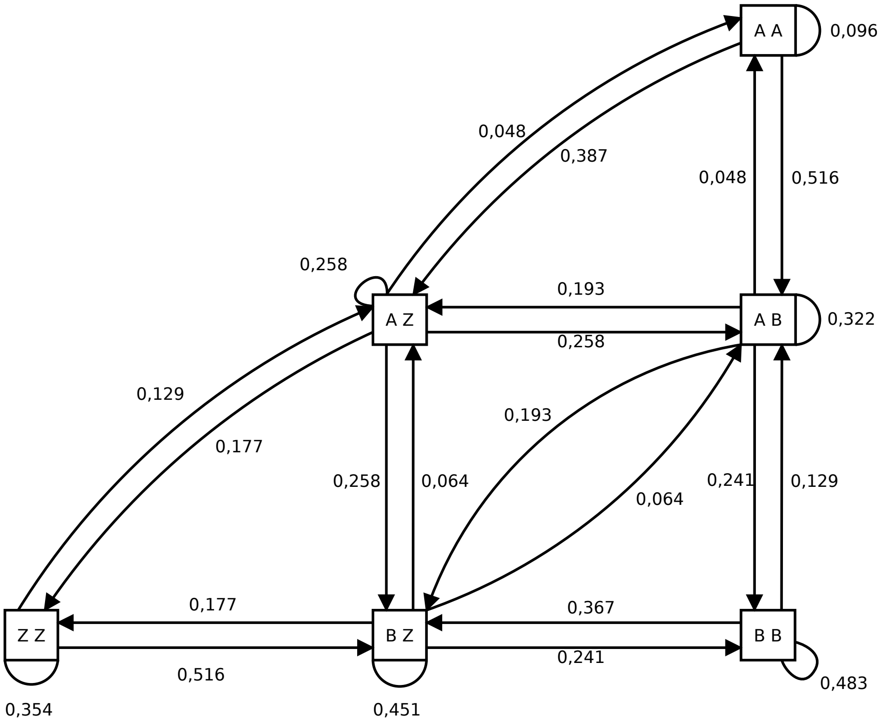


Figure: (Christian Bodenstein, 2016, © unclear)

Thanks to Christian Bodenstein for the table and the graph. A python script by him is available upon request.

See also the lecture notes for another example.

5 Bayesian learning for binary variables

5.0.1 Exercise: Kullback-Leibler divergence

A well known measure for the similarity between two probability density distributions $p(x)$ and $q(x)$ is the Kullback-Leibler divergence

$$D_{\mathrm{KL}}(p, q) := \int p(x) \ln\left(\frac{p(x)}{q(x)}\right) \mathrm{d}x\,. \tag{1}$$

1. Verify that

$$\ln(x) \leq x - 1 \tag{2}$$

for all $x > 0$ with equality if, and only if, $x = 1$.

Solution: First we show that for $x = 1$ we get

$$\ln(x) = 0 = (1 - 1) = x - 1\,. \tag{3}$$

Then we calculate the derivative of the difference

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(\ln(x) - (x - 1)\right) = \frac{1}{x} - 1\,. \tag{4}$$

Since $\ln(x) = x - 1$ for $x = 1$ and since the derivative of the difference is positive for $0 < x < 1$ and negative for $x > 1$, we conclude that $\ln(x) \leq x - 1$ for all $x > 0$.

2. Show that $D_{\mathrm{KL}}(p, q)$ is non-negative.

   Hint: Assume $p(x), q(x) > 0 \; \forall x$ for simplicity.

   Solution:

$$\begin{aligned} D_{\mathrm{KL}}(p,q) &= \int p(x) \ln\left(\frac{p(x)}{q(x)}\right) \mathrm{d}x && (5)\\ &= -\int p(x) \ln\left(\frac{q(x)}{p(x)}\right) \mathrm{d}x && (6)\\ &\overset{(2)}{\geq} -\int p(x) \left(\frac{q(x)}{p(x)} - 1\right) \mathrm{d}x \quad (\text{since } p(x) \geq 0) && (7)\\ &= -\int (q(x) - p(x))\, \mathrm{d}x && (8)\\ &= -\underbrace{\int q(x)\, \mathrm{d}x}_{=1} + \underbrace{\int p(x)\, \mathrm{d}x}_{=1} && (9)\\ &= 0\,. && (10) \end{aligned}$$

3. Show that $D_{\mathrm{KL}}(p, q) = 0$ only if $q(x) = p(x)$.

   Solution: This follows directly from the fact that in step (7) equality for any given x only holds if $q(x) = p(x)$. (This derivation is valid only up to differences between $q(x)$ and $p(x)$ on a set of measure zero.)

   Extra question: Is the Kullback-Leibler $D_{\mathrm{KL}}(p, q)$ divergence a good distance measure?

5.0.2 Exercise: Estimation of a probability density distribution

A scalar random variable x shall be distributed according to the probability density function $p_0(x)$. Assume you want to make an estimate $p(x)$ for this pdf based on N measurements x_i, $i \in \{1, ..., N\}$.

A suitable objective function is given by

$$F(p(x)) := \langle \ln(p(x)) \rangle_x := \int_{-\infty}^{\infty} p_0(x) \ln(p(x))\, \mathrm{d}x\,. \qquad (1)$$

Make the ansatz that the probability density function $p(x)$ is a Gaussian with mean μ and standard deviation σ,

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\,. \qquad (2)$$

1. Motivate on the basis of the Kullback-Leibler divergence (see exercise 5.0.1) that (1) is a good objective function. Should it be maximized or minimized?

   Solution: The goal obviously is to minimize the divergence between the true p_0 and the estimated distribution p. If we choose the roles of p and p_0 in the Kullback-Leibler divergence appropriately, we get

$$\begin{aligned} \text{minimize} \qquad D_{\mathrm{KL}}(p_0, p) &= \int p_0(x) \ln\left(\frac{p_0(x)}{p(x)}\right) \mathrm{d}x && (3)\\ &= \underbrace{\int p_0(x) \ln\left(p_0(x)\right)\, \mathrm{d}x}_{=\,\mathrm{const}} - \int p_0(x) \ln\left(p(x)\right)\, \mathrm{d}x && (4)\\ \Longleftrightarrow \qquad \text{maximize} \qquad & \int p_0(x) \ln\left(p(x)\right)\, \mathrm{d}x && (5)\\ &= \langle \ln(p(x)) \rangle_x && (6) \end{aligned}$$

with the averaging performed over the true distribution $p_0(x)$. This corresponds to (1).

2. Calculate the objective function $F(\mu, \sigma)$ as a function of μ and σ by substituting in (1) the averaging over $p_0(x)$ by the average over the measured data x_i and using the ansatz (2) for $p(x)$.

   Solution: Substituting yields

$$
\begin{align}
F(\mu,\sigma) &:= F(p(x)) \tag{7}\\
&\overset{(1)}{=} \langle \ln(p(x)) \rangle_{p_0} \tag{8}\\
&\approx \langle \ln(p(x_i)) \rangle_{i} \tag{9}\\
&:= \frac{1}{N}\sum_{i=1}^{N} \ln(p(x_i)) \tag{10}\\
&\overset{(2)}{=} \frac{1}{N}\sum_{i=1}^{N} \ln\left(\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)\right) \tag{11}\\
&= \ln\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) + \frac{1}{N}\sum_{i=1}^{N}\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right) \tag{12}\\
&= -\ln(2\pi)/2 - \ln(\sigma) - \frac{1}{N2\sigma^2}\sum_{i=1}^{N}(x_i-\mu)^2 \,. \tag{13}
\end{align}
$$

3. Calculate the gradient of the objective function $F(\mu, \sigma)$ wrt μ and σ.

   Solution: Taking the partial derivatives of $F(\mu, \sigma)$ wrt μ and σ yields

$$
\begin{align}
\frac{\partial F(\mu,\sigma)}{\partial \mu} &\overset{(13)}{=} \frac{1}{N2\sigma^2}\sum_{i=1}^{N} 2(x_i-\mu) \tag{14}\\
&= \frac{1}{\sigma^2}\left(\left(\frac{1}{N}\sum_{i=1}^{N} x_i\right)-\mu\right) \,, \tag{15}\\
\frac{\partial F(\mu,\sigma)}{\partial \sigma} &\overset{(13)}{=} -\frac{1}{\sigma} + \frac{1}{N\sigma^3}\sum_{i=1}^{N}(x_i-\mu)^2 \,. \tag{16}
\end{align}
$$

4. Determine the optimal values $\mu^\star$ and $\sigma^\star$ for μ and σ.

   Solution: A necessary condition for optimal values of μ and σ is that the derivatives vanish, i.e.

$$
\begin{align}
0 \overset{!}{=} \left.\frac{\partial F(\mu,\sigma)}{\partial \mu}\right|_{\mu^\star,\sigma^\star} &\overset{(15)}{=} \frac{1}{\sigma^{\star 2}}\left(\left(\frac{1}{N}\sum_{i=1}^{N} x_i\right)-\mu^\star\right) \tag{17}\\
\iff \qquad \mu^\star &= \frac{1}{N}\sum_{i=1}^{N} x_i \,, \quad \text{(for finite } \sigma^\star) \tag{18}\\
0 \overset{!}{=} \left.\frac{\partial F(\mu,\sigma)}{\partial \sigma}\right|_{\mu^\star,\sigma^\star} &\overset{(16)}{=} -\frac{1}{\sigma^\star} + \frac{1}{N\sigma^{\star 3}}\sum_{i=1}^{N}(x_i-\mu^\star)^2 \tag{19}\\
\iff \qquad \sigma^{\star 2} &= \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu^\star)^2 \,. \quad \text{(for finite } \sigma^\star) \tag{20}
\end{align}
$$

   Assuming $\sigma^\star$ has to be positive, these two equations uniquely determine the values of $\mu^\star$ and $\sigma^\star$ to be the mean and the variance of the data distribution, as one would expect.

   Extra question: Does the result make sense?

   Extra question: Why does one sometimes use $1/(N-1)$ instead of $1/N$ for estimating the standard deviation?

5.1 Levels of Bayesian learning

5.2 A simple example

5.3 The Beta distribution

5.3.1 Exercise: Beta-distribution

The beta-distribution is defined on $[0, 1]$ as

$$\beta(\theta; H, T) := \frac{\theta^{H-1}(1-\theta)^{T-1}}{B(H,T)}, \tag{1}$$

with the normalizing factor $B(H, T)$ chosen such that

$$\int_0^1 \beta(\theta; H, T)\,\mathrm{d}\theta = 1 \tag{2}$$

and given by the beta-function, which can be written in terms of the Γ-function:

$$B(H,T) = \frac{\Gamma(H)\Gamma(T)}{\Gamma(H+T)} \tag{3}$$

The Γ-function is a generalization of the factorial to real numbers. For positive integer numbers a the relation $\Gamma(a + 1) = a!$ holds; for real a it fulfills the recursion relation

$$\Gamma(a+1) = a\Gamma(a)\,. \tag{4}$$

1. Prove that the moments $\mu_r := \langle\theta^r\rangle_\beta$ of the beta-distribution are given by

$$\langle\theta^r\rangle_\beta = \frac{\Gamma(H+r)\Gamma(H+T)}{\Gamma(H+r+T)\Gamma(H)}. \tag{5}$$

Solution: The moments of the beta-distribution can be computed as

$$\begin{aligned}
\langle\theta^r\rangle_\beta &= \int_0^1 \theta^r\,\beta(\theta; H, T)\,\mathrm{d}\theta && (6)\\
&\overset{(1)}{=} \int_0^1 \theta^r\,\frac{\theta^{H-1}(1-\theta)^{T-1}}{B(H,T)}\,\mathrm{d}\theta && (7)\\
&= \frac{B(H+r,T)}{B(H,T)} \int_0^1 \frac{\theta^{(H+r)-1}(1-\theta)^{T-1}}{B(H+r,T)}\,\mathrm{d}\theta && (8)\\
&\overset{(1)}{=} \frac{B(H+r,T)}{B(H,T)} \underbrace{\int_0^1 \beta(\theta; H+r, T)\,\mathrm{d}\theta}_{\overset{(2)}{=}1} && (9)\\
&\overset{(3)}{=} \frac{\Gamma(H+r)\Gamma(T)}{\Gamma(H+r+T)} \Bigg/ \frac{\Gamma(H)\Gamma(T)}{\Gamma(H+T)} && (10)\\
&= \frac{\Gamma(H+r)\Gamma(H+T)}{\Gamma(H+r+T)\Gamma(H)}\,. && (11)
\end{aligned}$$

2. Show that the mean of the beta-distribution is given by

$$\langle\theta\rangle_\beta = \frac{H}{H+T}\,. \tag{12}$$

Solution: This follows directly from the first part of the exercise.

$$\begin{aligned}
\langle\theta\rangle_\beta &\overset{(5)}{=} \frac{\Gamma(H+1)\Gamma(H+T)}{\Gamma(H+1+T)\Gamma(H)} && (13)\\
&\overset{(4)}{=} \frac{H\Gamma(H)\Gamma(H+T)}{(H+T)\Gamma(H+T)\Gamma(H)} && (14)\\
&= \frac{H}{H+T}\,. && (15)
\end{aligned}$$

6 Learning in Bayesian networks

6.1 Breaking down the learning problem

6.2 Learning with complete data

6.3 Learning with incomplete data

6.3.1 Exercise: Parameter estimation

Consider a Bayesian network consisting of N binary random variables. The task is to learn the probability distribution of these variables. How many parameters must be estimated in the worst case?

Solution: In the worst case the joint probability cannot be factorized in conditional probabilities in any non-trivial way. In that case, a probability must be learned for each of the 2^N combinations of values. This number is only reduced by one due to the normalization condition, resulting in $2^N - 1$.

In the solution above, we have avoided thinking about the worst case Bayesian network at all. We have simply argued from the worst case distribution. One trivial way one can factorize any distribution is like

$$\begin{aligned}
P(A,B,C,D,E) &= \frac{P(A,B,C,D,E)}{P(B,C,D,E)}\frac{P(B,C,D,E)}{P(C,D,E)}\frac{P(C,D,E)}{P(D,E)}\frac{P(D,E)}{P(E)}P(E) && (1)\\
&= \underbrace{P(A|B,C,D,E)}_{\text{16 parameters}}\underbrace{P(B|C,D,E)}_{\text{8 parameters}}\underbrace{P(C|D,E)}_{\text{4 parameters}}\underbrace{P(D|E)}_{\text{2 parameters}}\underbrace{P(E)}_{\text{1 parameters}}\quad . && (2)
\end{aligned}$$

The nodes of the corresponding Bayesian network can be linearly ordered such that all units are connected to later ones but not to earlier ones. The number of parameters of each conditional probability depends on the number d of random variables the considered variable depends upon and reads 2^d. It is relatively easy to see that adding up all parameters leads to $2^N - 1$, as found above. (Jörn Buchwald, WS'10)

Extra question: How many parameters do you need in the best case?