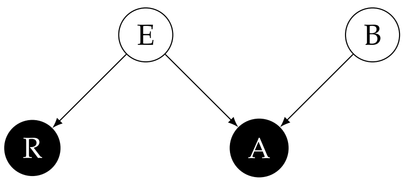
Exercise Sheet #1

Solution

October 30, 2018

1 Inference in a directed graphical model

We are given the following graph:



A = the alarm was triggered
E = there was an earthquake
B = there was a break-in
R = an announcement is made on the radio

and the following probabilities:

$p(E = 1) = 10^{-3}$	$p(B = 1) = 10^{-3}$
$p(A = 1 \mid B = 0, E = 0) = 0.001$	$p(A = 1 \mid B = 1, E = 0) = 0.99001$
$p(A = 1 \mid B = 0, E = 1) = 0.01099$	$p(A = 1 \mid B = 1, E = 1) = 0.9901099$

In order to compute the probability that there was an earthquake, after hearing about the alarm, but before hearing the radio report, we need to compute

$$p(E = 1 \mid A = 1) \overset{\text{Bayes' Theorem}}{=} \frac{p(A = 1 \mid E = 1)p(E = 1)}{p(A = 1)} \,. \tag{1}$$

While the probability $p(E = 1)$ is directly given to be 0.001, we have to compute the other two occurring probabilities separately.

$$
\begin{aligned}
p(A=1|E=1) &= p(A=1|B=0,E=1)p(B=0) + p(A=1|B=1,E=1)p(B=1) \\
&= 0.01099 \cdot (1-0.001) + 0.9901099 \cdot 0.001 \\
&= 0.01197
\end{aligned} \tag{2}
$$

$$
\begin{aligned}
p(A=1) = &\quad p(A=1|B=0,E=0)p(B=0)p(E=0) \\
&+ p(A=1|B=0,E=1)p(B=0)p(E=1) \\
&+ p(A=1|B=1,E=0)p(B=1)p(E=0) \\
&+ p(A=1|B=1,E=1)p(B=1)p(E=1) \\
= &\quad 0.001 \cdot (1-0.001) \cdot (1-0.001) \\
&+ 0.01099 \cdot (1-0.001) \cdot 0.001 \\
&+ 0.99001 \cdot 0.001 \cdot (1-0.001) \\
&+ 0.9901099 \cdot 0.001 \cdot 0.001 \\
\approx &\ 0.002
\end{aligned} \tag{3}
$$

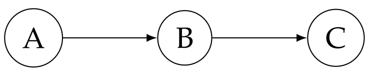
Plugging (2) and (3) into (1), we get

$$
p(E=1|A=1) \overset{\text{Bayes' Theorem}}{=} \frac{p(A=1|E=1)p(E=1)}{p(A=1)} = \frac{0.01197 \cdot 0.001}{0.002} = 0.00599\,.
$$

Which means that the probability that there was an earthquake after hearing the alarm, but before hearing the radio report is about $0.6\,\%$.

2 Independence in three-node directed graphs

(a)



We need to show that $A \perp\!\!\!\perp C \mid B$ which means $p(A,C|B) = p(A|B)p(C|B)$.

From the graph, we can read the factorization $p(A,B,C) = p(C|B)p(B|A)p(A)$. Therefore,

$$
p(A,C|B) = \frac{p(A,B,C)}{p(B)} \underset{\text{Factorization}}{=} \frac{p(C|B)p(B|A)p(A)}{p(B)} \underset{\text{on } p(B|A)}{\overset{\text{Bayes' Theorem}}{=}} \frac{p(C|B)p(A|B)\cancel{p(B)}\,\cancel{p(A)}}{\cancel{p(B)}\,\cancel{p(A)}} = p(A|B)p(C|B)\,.
$$

(b)

$$\begin{tikzcd} A & B \arrow[l] & C \arrow[l] \end{tikzcd}$$

We need to show that $A \perp\!\!\!\perp C \mid B$ which means $p(A, C|B) = p(A|B)p(C|B)$.

From the graph, we can read the factorization $p(A, B, C) = p(A|B)p(B|C)p(C)$. Therefore,

$$p(A, C|B) = \frac{p(A, B, C)}{p(B)} \underset{\text{Factorization}}{=} \frac{p(A|B)p(B|C)p(C)}{p(B)} \underset{\text{on } p(B|C)}{\overset{\text{Bayes' Theorem}}{=}} \frac{p(A|B)p(C|B)\cancel{p(B)}\,\cancel{p(C)}}{\cancel{p(B)}\,\cancel{p(C)}} = p(A|B)p(C|B)\,.$$

(c)

$$\begin{tikzcd} A & B \arrow[l] \arrow[r] & C \end{tikzcd}$$

We need to show that $A \perp\!\!\!\perp C \mid B$ which means $p(A, C|B) = p(A|B)p(C|B)$.

From the graph, we can read the factorization $p(A, B, C) = p(A|B)p(C|B)p(B)$. Therefore,

$$p(A, C|B) = \frac{p(A, B, C)}{p(B)} \underset{\text{Factorization}}{=} \frac{p(A|B)p(C|B)\cancel{p(B)}}{\cancel{p(B)}} = p(A|B)p(C|B)\,.$$

(d)

$$\begin{tikzcd} A \arrow[r] & B & C \arrow[l] \end{tikzcd}$$

We need to show that $A \perp\!\!\!\perp C$ which means $p(A, C) = p(A)p(C)$.

From the graph, we can read the factorization $p(A, B, C) = p(B|A, C)p(A)p(C)$. Therefore,

$$p(A, C) = \frac{p(A, B, C)}{p(B|A, C)} \underset{\text{Factorization}}{=} \frac{\cancel{p(B|A, C)}\,p(A)p(C)}{\cancel{p(B|A, C)}} = p(A)p(C)\,.$$

We follow the example of inferring the probability of wearing glasses from slides 24–31 from Lecture 2 (Probabilities over Continuous Variables) of the lecture 'Probabilistic Inference and Learning'. For both sellers, i.e. for $i \in \{1, 2\}$, we model the probability for seller i to receive a negative review by a *random variable* (RV) Y_i. Moreover, we model the reviews as binary random variables. Recall that the first and second seller have received $k_1 := 100$ and $k_2 := 2$ reviews respectively. We denote the (binary) reviews of the first seller by $\{X_1^{(j)}\}_{j=1}^{100}$. Analogously, we denote the (binary) reviews of the second seller by $\{X_2^{(j)}\}_{j=1}^2$. Since we are interested in the *posterior* over θ_1 and θ_2, we need to apply Bayes' rule. To this end, we need priors $p(\theta_1)$ and $p(\theta_2)$ which are assumed uniform. Hence, for $i \in \{1,2\}$,

$$\begin{aligned} p(\theta_i) &= \mathbb{1}_{\theta_i \in [0,1]} \\ &= \frac{\theta_i^{1-1}(1-\theta_i)^{1-1}}{\int_0^1 \theta_i^{1-1}(1-\theta_i)^{1-1}\, \mathrm{d}\theta_i} \\ &= f_{\mathrm{beta},a=1,b=1}(\theta_i), \end{aligned} \tag{4}$$

where $f_{\mathrm{beta},a=1,b=1}$ denotes the *probability density function* of the beta distribution with parameters $a = 1$ and $b = 1$. Now, the number of negative reviews

$$N_i = \sum_{j=1}^{k_i} X_i^{(j)} \tag{5}$$

is—as a sum of i.i.d. random variables $X_i^{(j)}$ with Bernoulli(θ_i) distribution—distributed according to a binomial distribution B(k_i,θ_i). Hence, the likelihood of n_i is given by

$$p(n_i|\theta_i) = \theta_i^{n_i}\,(1-\theta_i)^{k_i - n_i}\,. \tag{6}$$

Now, we incorporate the data. We observed that, for seller 1, 10 reviews were negative and, for seller 2, 0 reviews were negative. Hence, $n_1 = 10$ and $n_2 = 0$. The (marginal) posteriors on θ_i are therefore (as in the lecture) given by

$$\begin{aligned} p(\theta_i|D_i) &= p(\theta_i|N_i = n_i) \\ &= \frac{\theta_i^{n_i+1-1}\,(1-\theta_i)^{(k_i-n_i)+1-1}}{B(1+n_i, 1+k_i-n_i)} \\ &= \frac{\theta_i^{n_i}\,(1-\theta_i)^{k_i-n_i}}{B(1+n_i, 1+k_i-n_i)} \\ &= f_{\mathrm{beta},a=1+n_i,b=1+k_i-n_i}(\theta_i). \end{aligned} \tag{7}$$

Now, to compute the joint posterior over (θ_1, θ_2), first note that, for $i \in \{0,1\}$,

$$p(D_i|(\theta_1,\theta_2)) = p(D_i|\theta_i) \tag{8}$$

because the likelihood of the data on seller i does only depend on the reliability θ_i of seller i and not on the reliability of the other seller. Moreover, recall that, by the assumption that 'the data is independent of each other',

$$\begin{aligned} p(D_1, D_2|\theta_1,\theta_2) &= p(D_1|\theta_1,\theta_2)p(D_1|\theta_1,\theta_2) \\ &\overset{(8)}{=} p(D_1|\theta_1)p(D_2|\theta_2), \end{aligned} \tag{9}$$

and that, by the assumption that 'the reliabilities are independent of reach other',

$$p(\theta_1, \theta_2) = p(\theta_1)p(\theta_2). \tag{10}$$

Hence, by Bayes' rule, we can compute that

$$\begin{aligned}
p((\theta_1,\theta_2)|D_1,D_2) &= \frac{\overbrace{p(D_1,D_2|\theta_1,\theta_2)}^{p(D_1|\theta_1)p(D_2|\theta_2),\text{ by (9)}} \cdot \overbrace{p(\theta_1,\theta_2)}^{p(\theta_1)p(\theta_2),\text{ by (10)}}}{\int_0^1\int_0^1 \underbrace{p(D_1,D_2|\theta_1,\theta_2)}_{p(D_1|\theta_1)p(D_2|\theta_2),\text{ by (9)}} \underbrace{p(\theta_1,\theta_2)}_{p(\theta_1)p(\theta_2),\text{ by (10)}} \,\mathrm{d}\theta_1\,\mathrm{d}\theta_2} \\
&= \frac{p(D_1|\theta_1)p(\theta_1)}{\int_0^1 p(D_1|\theta_1)p(\theta_1)\,\mathrm{d}\theta_1} \cdot \frac{p(D_2|\theta_2)p(\theta_2)}{\int_0^1 p(D_2|\theta_2)p(\theta_2)\,\mathrm{d}\theta_2} \\
&= p(\theta_1|D_1)\cdot p(\theta_2|D_2) \\
&\overset{(7)}{=} f_{\text{beta},a=1+n_1,b=1+k_1-n_1}(\theta_1) f_{\text{beta},a=1+n_2,b=1+k_2-n_2}(\theta_2) \\
&= f_{\text{beta},a=11,b=91}(\theta_1) f_{\text{beta},a=1,b=3}(\theta_2),
\end{aligned} \tag{11}$$

which means that the joint posterior factorizes into the marginal posteriors $f_{\text{beta},a=11,b=91}$ and $f_{\text{beta},a=1,b=3}$. To conclude we compute,

$$\begin{aligned}
p(\theta_1 > \theta_2|D_1,D_2) &= \int_0^1\int_0^1 \mathbb{1}_{\theta_1\geq\theta_2}\,\mathrm{d}p((\theta_1,\theta_2)|D_1,D_2) \\
&= \int_0^1\int_{\theta_2}^1 p((\theta_1,\theta_2)|D_1,D_2)\,\mathrm{d}\theta_1\,\mathrm{d}\theta_2 \\
&= \int_0^1\int_{\theta_2}^1 p(\theta_1|D_1)p(\theta_2|D_2)\,\mathrm{d}\theta_1\,\mathrm{d}\theta_2 \\
&= \int_0^1 p(\theta_2|D_2)\left[\int_{\theta_2}^1 p(\theta_1|D_1)\,\mathrm{d}\theta_1\right]\,\mathrm{d}\theta_2 \\
&= \int_0^1 p(\theta_2|D_2)\left[1-\int_0^{\theta_2} p(\theta_1|D_1)\right]\,\mathrm{d}\theta_2 \\
&= \int_0^1 p(\theta_2|D_2)\left[1-F_{\text{beta},a=11,b=91}(\theta_2)\right]\,\mathrm{d}\theta_2 \\
&= \mathbb{E}_{\theta_2\sim p(\theta|D_2)}\left[1-F_{\text{beta},a=11,b=91}(\theta_2)\right],
\end{aligned} \tag{12}$$

where $F_{\text{beta},a=11,b=91}(\theta_2)$ denotes the *cumulative distribution function* (CDF) of the beta distribution with parameters $a = 11$ and $b = 91$. This integral can be computed with your favorite numerical integration algorithms. Here are three solutions with approximately the same computational cost (0.4 s on iMac with 4GHz Intel Core i7). Example code can be found in the python file 01_Ex3.py.

Method 1 (Brute Force Sampling): Using the first line of (12), simply sample a large amount N of random vectors $(\xi_1^{(n)}, \xi_2^{(n)}) \sim \text{beta}(a = 11, b = 91) \times \text{beta}(a = 1, b = 3)$

and return the ratio of samples with $\xi_1^{(n)} > \xi_2^{(n)}$, i.e.

$$
\begin{aligned}
p(\theta_1 > \theta_2 | D_1, D_2) &\approx \frac{\left|\left\{n : \xi_1^{(n)} > \xi_2^{(n)}\right\}\right|}{N} \\
&\approx 0.298 \qquad \text{(example run, after } N = 1000 \text{ samples).} \qquad (13)
\end{aligned}
$$

<u>*Method 2 (Smart Sampling):*</u> Using the last line of (12), we can sample a large amount N of random variables $\xi_2^{(n)} \sim \text{beta}(a = 1, b = 3)$ and compute the average of $1 - F_{\text{beta},a=11,b=91}(\xi_2)$, i.e.

$$
\begin{aligned}
p(\theta_1 > \theta_2 | D_1, D_2) &\approx \frac{1}{N} \sum_{n=1}^{N} 1 - F_{\text{beta},a=11,b=91}\left(\xi_2^{(n)}\right) \qquad (14) \\
&\approx 0.283 \qquad \text{(example run, after } N = 1000 \text{ samples).} \qquad (15)
\end{aligned}
$$

<u>*Method 3 (Numerical integration):*</u> Using the second last line of (12), we derive that

$$
\begin{aligned}
p(\theta_1 > \theta_2 | D_1, D_2) &= \int_0^1 p(\theta_2 | D_2) \left[1 - F_{\text{beta},a=11,b=91}(\theta_2)\right] \, \mathrm{d}\theta_2 \qquad (16) \\
&\overset{(7)}{=} \int_0^1 \underbrace{f_{\text{beta},a=1,b=3}(\theta_2)[1 - F_{\text{beta},a=11,b=91}(\theta_2)]}_{=:g(\theta_2)} \, \mathrm{d}\theta_2 \qquad (17) \\
&\approx 0.2874 \qquad (\pm 9 * 10^{-11}). \qquad (18)
\end{aligned}
$$

which we integrated with the numerical integration library scipy.integrate. Note that $9 * 10^{-11}$ is an estimate of the absolute error computed by scipy.integrate. Hence, this is more exact than the sampling-based approaches from above. (This is the smartest solution.)

<u>*Take-away:*</u> Perhaps surprisingly, you are more likely to have a negative experience at seller 2, despite their 100% positive reviews. *Sample-size matters! Don't be overconfident in statistics based on just a small number of data points!*

3 The Poisson Distribution

We can re-write the binomial distribution

$$
p_b(r|f, N) = \binom{N}{r} f^r (1-f)^{N-r} = \frac{N \cdot (N-1) \dots (N - (r-1))}{r!} f^r (1-f)^{N-r}
$$

if we consider the limit of this for $N \to \infty$, with a sequence f_N such that $\lim_{N\to\infty} N f_N = \lambda$, we get

$$
\begin{aligned}
\lim_{N\to\infty} p_b(r|f_N, N) &= \lim_{N\to\infty} \frac{N\cdot(N-1)\dots(N-(r-1))}{r!} f_N^r (1-f_N)^{N-r} \\
&\overset{\lambda=\lim_{N\to\infty} N f_N}{=} \lim_{N\to\infty} \frac{N\cdot(N-1)\dots(N-(r-1))}{r!} \left(\frac{\lambda}{N}\right)^r \left(1-\frac{\lambda}{N}\right)^{N-r} \\
&= \lim_{N\to\infty} \frac{N\cdot(N-1)\dots(N-(r-1))}{N^r} \frac{\lambda^r}{r!} \left(1-\frac{\lambda}{N}\right)^N \left(1-\frac{\lambda}{N}\right)^{-r} .
\end{aligned}
$$

With the following three limits

$$
\begin{aligned}
\lim_{N\to\infty} \left(1-\frac{\lambda}{N}\right)^N &= e^{-\lambda} \\
\lim_{N\to\infty} \left(1-\frac{\lambda}{N}\right)^{-r} &= 1 \\
\lim_{N\to\infty} \frac{N\cdot(N-1)\dots(N-(r-1))}{N^r} = \lim_{N\to\infty} \frac{N^r + \mathcal{O}(N^{r-1})}{N^r} &= 1
\end{aligned}
$$

we get

$$
\lim_{N\to\infty} p_b(r|f_N, N) = e^{-\lambda} \frac{\lambda^r}{r!} \qquad \square
$$