Machine Learning 521289S
Spring 2019
Exercise 1: Basic Probability and Statistics & Bayesian Classification

Here the upper case $P(\cdot)$ is used to denote the *probability mass function* (of a discrete random variable), and lower case $p(\cdot)$ to denote a *probability distribution* (of a continuous random variable). Vectors are written in bold lowercase letters ($\mathbf{x}$) and matrices in bold uppercase letters ($\mathbf{X}$). Random variables are written in upper case letters and their values in lower case letters ($X = x$, $C = 1$).

Solutions

1. *Conditional Probability*

   (a) To solve this, we must first clearly define the events that are taking place. Let us denote the event "the woman lives to age 60" by A and the event "the woman lives to age 80" by B. Now the question we have been asked is given that "the woman lives to age 60" what is the probability that "the woman lives to age 80". This is exactly what the conditional probability tells us, and it can be written as $\Pr(B|A)$. Now, to use the formula that defines the conditional probability, we must find out $\Pr(A, B)$ and $\Pr(B)$. The latter is clearly known and the former $\Pr(A, B)$ means that "the woman lives to age 60 **and** the woman lives to age 80" which is, in fact, just the event $B =$"the woman lives to age 80". Thus,

$$\begin{aligned}\Pr(B|A) &= \frac{\Pr(B, A)}{\Pr(A)} \\ &= \frac{\Pr(B)}{\Pr(A)} \\ &= \frac{0.57062}{0.89835} \\ &= 0.635\,19 \\ &\approx 63.5\ \%\end{aligned}$$

   which is the desired result.

   Another way to look at this problem is to define two random variables X and Y so that $X = 1$ if the woman lives up to age 60, and 0 if not. Similarly, we let $Y = 1$ if the woman lives up to age 80, and 0 if not. Clearly, $P(X = 1) = 0.89835$, and so forth. Now, the event A can be restated as $X = 1$ and the event B as $Y = 1$. This allows us to use the *probability mass functions* (PMFs) to write the conditional probability as

$$\begin{aligned}P(y|x) &= \frac{P(y, x)}{P(x)} \\ &= \frac{P(y)}{P(x)}\end{aligned}$$

   arriving at the same solution as above. When the random variables at play are clear from the context, one often just writes out the probabilities using PMFs omitting the tedious definition of variables.

(b) This problem motivates the Bayes' theorem. Just like previously, let us define two random variables H and R for Mary having a better hand and raising such that 0 denotes the case in which it does not happen and 1 denotes the case in which it happens. We are given the probabilities $P(H=1)=0.04$, $P(R=1|H=1)=0.9$, and $P(R=1|H=0)=0.1$.

The task is now to find $P(H=1|R=1)$. We may use the Bayes formula from Question 2 to write

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

In the above, we may use the marginalisation technique backwards (cf. the solution of 4) and then product rule (cf. the solution of 2 a) to get

$$\begin{aligned} P(R=1) &= \sum_{h\in\{0,1\}} P(R=1|H=h)\,P(H=h) \\ &= P(R=1|H=0)\,P(H=0)+P(R=1|H=1)\,P(H=1) \\ &= 0.1\times 0.96+0.9\times 0.04=0.132\,. \end{aligned}$$

Hence,

$$\begin{aligned} P(H=1|R=1) &= \frac{0.9\times 0.04}{0.132} \\ &= 0.272\,73 \\ &\approx 27.3\ \%. \end{aligned}$$

2. *The Bayes' Theorem*

(a) First, we recall that the probability of A given B is by definition

$$\Pr(A|B)=\frac{\Pr(A,B)}{\Pr(B)},$$

Now we use the conditional probability formula in the product form to the event $A\cap B=B\cap A$ in the nominator, i.e. $Pr(A,B)=Pr(B,A)=Pr(B|A)Pr(A)$. Consequently, we get the final result

$$\Pr(A|B)=\frac{\Pr(B|A)\Pr(A)}{\Pr(B)}.$$

(b) Let Ω be partitioned as $\Omega=\cup_{i=1}^{n}A_i$, where $A_i\cap A_j=\emptyset$ whenever $i\neq j$. As the events A_i are mutually disjoint, the *Law of Total probability* states that

$$\Pr(B)=\sum_{k=1}^{n}\Pr(B,A_k)\,.$$

Now, by knowing that $P(B,A_k)=P(B|A_k)\,P(A_k)$ for all k, so we get

$$\Pr(B)=\sum_{k=1}^{n}\Pr(B|A_k)\Pr(A_k)\,.$$

Therefore, the Bayes' Theorem $P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B)}$ may be rewritten in the *alternative form*

$$\Pr(A_i|B) = \frac{\Pr(B|A_i)\Pr(A_i)}{\sum_{k=1}^{n}\Pr(B|A_k)\Pr(A_k)},$$

as required.

3. *A Bayesian Classification Task*

(a) Let's sketch the class conditional probabilities $p(x|1)$ and $p(x|2)$ to the same coordinate frame as a function of x.

Now we know that

$$\begin{aligned} p(x|1) &= \frac{1}{\sqrt{2\pi}}e^{-\frac{(x-2)^2}{2}}, \\ p(x|2) &= \frac{1}{2\sqrt{2\pi}}e^{-\frac{(x-4)^2}{2\times 4}}. \end{aligned}$$

The standard normal distribution has the PDF

$$f(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}.$$

Hence, we see that

$$\begin{aligned} p(x|1) &= \frac{1}{\sigma_1}f\left(\frac{x-\mu_1}{\sigma_1}\right) = f(x-2) \\ p(x|2) &= \frac{1}{\sigma_2}f\left(\frac{x-\mu_2}{\sigma_2}\right) = \frac{1}{2}f\left(\frac{x-4}{2}\right) \end{aligned}$$

So by knowing only a few values of the standard normal distribution, we may sketch the class conditional distribution. Note that since f is symmetric, we only need values from the other side of the standard distribution. The formulas also tell us that the peak height of the PDF for the second class is half of the peak height for the first class and that the distribution for the second class is "two times wider" than that of the first one. Table 1 shows some tabulated values. Using it, we know that the peak values are $p(\mu_1|1) = p(2|1) = f(0) \approx 0.4$ and $p(\mu_2|2) = p(4|2) = \frac{1}{2}p(\mu_1|1) \approx 0.2$. One standard deviation away, we have $p(\mu_1 \pm \sigma_1|1) = p(2 \pm 1|1) = f(\pm 1) \approx 0.24$ and $p(\mu_2 \pm \sigma_2|2) = p(4 \pm 2|2) = \frac{1}{2}p(\mu_1 \pm \sigma_1|1) \approx 0.12$.

The distributions are depicted in Figure 1.

(b) Now

$$\begin{aligned} P(1|x) &= \frac{p(x|1)P(1)}{p(x|1)P(1) + p(x|2)P(2)} \\ &= \frac{p(x|1)\cdot\frac{1}{2}}{p(x|1)\cdot\frac{1}{2} + p(x|2)\cdot\frac{1}{2}} \\ &= \frac{p(x|1)}{p(x|1) + p(x|2)}. \end{aligned}$$

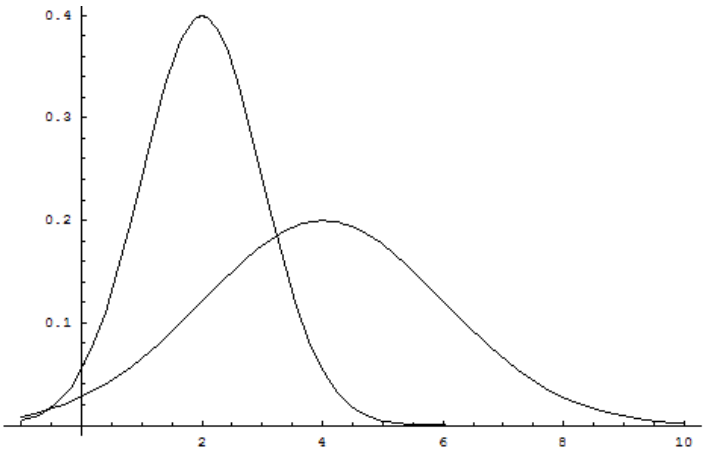

Figure 1: The likelihood distributions for both classes.

Let us now substitute the PDFs

$$\begin{aligned}
P(1|x) &= \frac{\frac{1}{\sqrt{2\pi}}e^{-\frac{(x-2)^2}{2}}}{\frac{1}{\sqrt{2\pi}}e^{-\frac{(x-2)^2}{2}} + \frac{1}{2\sqrt{2\pi}}e^{-\frac{(x-4)^2}{2\times 4}}} \\
&= \frac{e^{-\frac{(x-2)^2}{2}}}{e^{-\frac{(x-2)^2}{2}} + \frac{1}{2}e^{-\frac{(x-4)^2}{2\times 4}}} \\
&= \frac{2e^{-\frac{x^2-4x+4}{2}}}{2e^{-\frac{x^2-4x+4}{2}} + e^{-\frac{x^2-8x+16}{8}}} \\
&= \frac{2e^{-\frac{x^2}{2}+2x-2}}{2e^{-\frac{x^2}{2}+2x-2} + e^{-\frac{x^2}{8}+x-2}} \\
&= \frac{2}{2 + e^{-\frac{x^2}{8}+x-2-\left(-\frac{x^2}{2}+2x-2\right)}} \\
&= \frac{2}{2 + e^{\frac{3}{8}x\left(x-\frac{8}{3}\right)}}
\end{aligned}$$

How do we sketch this? We know that $\frac{3}{8}x\left(x-\frac{8}{3}\right)$ represents a parabola that opens upwards. It crosses the x-axis at $x_0 = 0$ and $x_1 = \frac{8}{3} = 2\frac{2}{3}$. It is symmetric

Table 1: The PDF of the standard normal distribution evaluated at a few points

x	$f(x)$	$f(x)$ approx.
0	$\frac{1}{\sqrt{2\pi}}$	0.4
0.5	$\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{8}}$	0.35
1	$\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}}$	0.24
1.5	$\frac{1}{\sqrt{2\pi}}e^{-\frac{9}{8}}$	0.13
2	$\frac{1}{\sqrt{2\pi}}e^{-2}$	0.05

with respect to the peak (the minimum) which is in the middle of the roots at $x_{peak} = \frac{0+\frac{8}{3}}{2} = \frac{8}{6} = \frac{4}{3}$. The exponent amplifies the parabola non-linearly, but the general shape stays the same as the exponent function is a monotonically increasing function. Finally, as the exponent is in the denominator, the scaled parabola we discussed in the previous step is inverted and the peak becomes the maximum, $P\left(1|\frac{4}{3}\right) \approx 0,8$. As x approaches positive or negative infinity, the exponent approaches also. Hence, $\lim_{x\to\pm\infty} P(1|x) = 0$.

Since $P(1|x) + P(2|x) = 1$, we know that $P(2|x) = 1 - P(1|x)$. Figure 2 shows the resulting posteriors that we may now sketch.

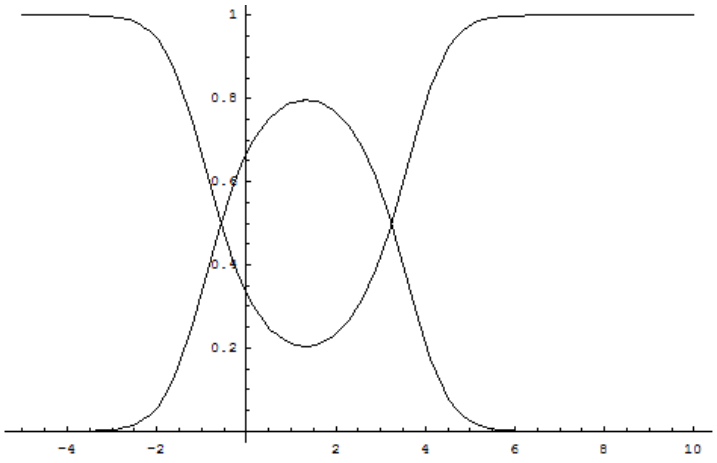

Figure 2: The posterior probabilities.

(c) Let us first compare the information presented in Figures 1 and 2. Since the prior probabilities are the same, classification is changed at the point in which the likelihoods are equal, i.e. at the intersections of the PDFs in Figure 1 sketched in the part (a). On the other hand, if the priors were different, the heights of the PDFs should be scaled accordingly before looking for the intersection. Figure 2 conveys the same information as Figure 1 but now in the posterior distribution form and gives us the same intersection point(s).

Note that the posterior distributions are not true PDFs (as a function of the continuous variable x). Too see this, consider, for example, $\int_{-\infty}^{\infty} P(2|x)\,dx = \infty \neq 1$. However, for each $x \in \mathbb{R}$ it holds that $\sum_{c=1}^{2} P(c|x) = 1$. So for a given x, the posterior distributions can be considered as *probability mass functions* (with respect to the discrete class variable) .

Finally, we may read Figure 2 to obtain the *Bayes decision rule*: If $-0.6 \leq x \leq 3.2$, then decide that the sample is from the class 1, otherwise assign it to the class 2. Consequently, the classification is to class 1 when the feature has the value 3.

(d) The classification thresholds occur when $P(1|x) = P(2|x)$. Now,

$$\begin{aligned}
P(1|x) &= P(2|x) \\
\frac{p(x|1)\,P(1)}{p(x)} &= \frac{p(x|2)\,P(2)}{p(x)} \\
p(x|1) &= p(x|2) \\
\frac{1}{\sqrt{2\pi}} e^{-\frac{(x-2)^2}{2}} &= \frac{1}{2\sqrt{2\pi}} e^{-\frac{(x-4)^2}{8}} \\
e^{-\frac{(x-2)^2}{2}} &= \frac{1}{2} e^{-\frac{(x-4)^2}{8}} \\
-\frac{x^2-4x+4}{2} &= \ln\frac{1}{2} - \frac{x^2-8x+16}{8} \\
-4x^2+16x-16 &= 8\ln\frac{1}{2} - x^2 + 8x - 16 \\
-4x^2+16x &= 8\ln\frac{1}{2} - x^2 + 8x \\
3x^2 - 8x - 8\ln 2 &= 0.
\end{aligned}$$

This can be solved with the formula for quadratic equation. We get

$$\begin{aligned}
x &= \frac{-(-8) \pm \sqrt{(-8)^2 - 4\times 3 \times (-8\ln 2)}}{2\times 3} \\
&= \frac{8 \pm \sqrt{64 + 96\ln 2}}{6} \\
&= \frac{8 \pm \sqrt{16\,(4 + 6\ln 2)}}{6} \\
&= \frac{8 \pm 4\sqrt{4 + 6\ln 2}}{2\times 3} \\
&= \frac{4 \pm 2\sqrt{4 + 6\ln 2}}{3} \\
&\approx \begin{cases} -0.57 \\ 3.2 \end{cases}.
\end{aligned}$$

If we are talking about the weight of a fish, negative weights do not make much sense. In practice, the normal distribution often only approximates the true distribution. Therefore, we may discard the negative value case and classify the fish to the class 2 when the weight is more than 3.2 units.

(e) The probability of error is

$$\begin{aligned}
\Pr(\text{“error”}) &= \int_{-\infty}^{\infty} P(\text{“error”}, x)\,dx \\
&= \int_{-\infty}^{\infty} P(\text{“error”}|x)\,p(x)\,dx
\end{aligned}$$

Our Bayes decision rule uses the optimal threshold at $x^* = 3.2$. Hence, using the Bayes decision rule when $x \le x^*$, an error is made only if the class is 2.

The probability of this happening is $P(2|x)$. Similarly, an error is made with probability $P(1|x)$ when $x > x^*$. Therefore, we may continue to write

$$\begin{aligned}
\Pr(\text{“error”}) &= \int_{-\infty}^{x^*} P(2|x)\, p(x)\, dx + \int_{x^*}^{\infty} P(1|x)\, p(x)\, dx \\
&= \int_{-\infty}^{x^*} P(2|x)\, p(x)\, dx + \int_{x^*}^{\infty} P(1|x)\, p(x)\, dx \\
&= \int_{-\infty}^{x^*} \frac{p(x|2)\, P(2)}{p(x)} p(x)\, dx + \int_{x^*}^{\infty} \frac{p(x|1)\, P(1)}{p(x)} p(x)\, dx \\
&= P(2) \int_{-\infty}^{x^*} p(x|2)\, dx + P(1) \int_{x^*}^{\infty} p(x|1)\, dx.
\end{aligned}$$

Figure 3 — a remake of Figure 1 — shows the area that gives us the probability of making an error according to this derivation.

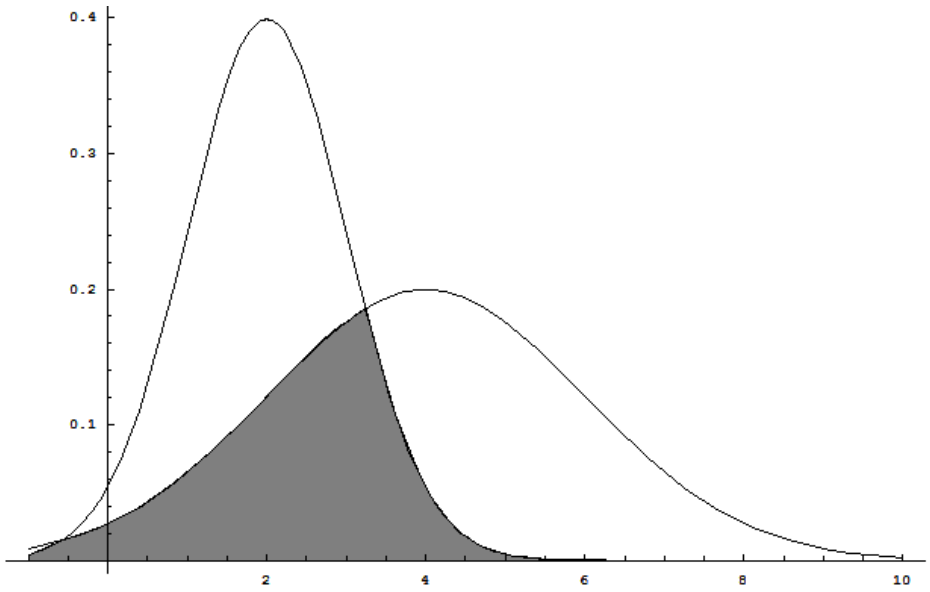

Figure 3: The grey area represents the probability of an error.

To calculate the actual probability of error, we use a table shown in Figure 4 that lists the values of the cumulative density function of a standardised normal distribution. Let us denote this CDF by

$$F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{y^2}{2}}\, dy.$$

Using the CDF we may write

$$\begin{aligned}
\Pr(\text{“error”}) &= P(2)\,F\left(\frac{x^\star-\mu_2}{\sigma_2}\right)+P(1)\left(1-F\left(\frac{x^\star-\mu_1}{\sigma_1}\right)\right)\\
&= \frac{1}{2}F\left(\frac{3.2-4}{2}\right)+\frac{1}{2}\left(1-F\left(\frac{3.2-2}{1}\right)\right)\\
&= \frac{1}{2}F(-0.4)+\frac{1}{2}(1-F(1.2))\\
&= \frac{1}{2}(1-F(0.4))+\frac{1}{2}(1-F(1.2))\\
&= 0.5\times(1-0.6554)+0.5\times(1-0.8849)\\
&= 0.229\,85\\
&\approx 23\ \%.
\end{aligned}$$

where we have used the values of the table shown in Table 2.

Table 2: Tabularised values of the CDF of the standard normal distribution $N(0,1)$. Probability content from $-\infty$ to Z

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

4. *Marginal Distributions and Statistical Independence*

First, let's use Table 1 to estimate estimate the joint *probability mass function* $P(c,s) = \Pr(\text{“}C = c \text{ and } S = s\text{”})$:

$$\begin{aligned} P(0,0) &= \frac{40}{60} = \frac{2}{3}, \\ P(0,1) &= \frac{10}{60} = \frac{1}{6}, \\ P(1,0) &= \frac{7}{60}, \\ P(1,1) &= \frac{3}{60} = \frac{1}{20}. \end{aligned}$$

Remember that the marginalisation of a discrete random variable is performed by summing over the unwanted variables, i.e.

$$P(c) = \sum_{s \in \{0,1\}} P(c,s)$$

Now, the task is easy.

(1) Let us handle the cancer case first. When $c = 0$, $P(c) = P(0) = P(0,0) + P(0,1) = \frac{2}{3} + \frac{1}{6} = \frac{5}{6}$. Similarly for $c = 1$, we get $P(c) = P(1) = P(1,0) + P(1,1) = \frac{7}{60} + \frac{1}{20} = \frac{1}{6}$. Thus,

$$P(c) = \begin{cases} \frac{5}{6} & \text{when } c = 0 \\ \frac{1}{6} & \text{when } c = 1 \end{cases}.$$

Of course, we could have calculated the latter probability directly from the first one, because with only two possibilities

$$P(1) = 1 - P(0).$$

(2) For the smoking habit, we get

$$P(s) = \begin{cases} \frac{2}{3} + \frac{7}{60} = \frac{47}{60} & \text{when } s = 0 \\ \frac{13}{60} & \text{when } s = 1 \end{cases}.$$

Note that with this short notation, it is easy to confuse $P(0) = \frac{5}{6}$ for cancer with $P(0) = \frac{47}{60}$ for smoking. In this kind of a scenario, it would be better to note the random variable for which the marginal distribution is calculated in a subscript. For example, $P_C(0)$ or $P_S(1)$. Another commonly used notation for the marginal distributions is $P(C = 0)$ or $P(S = 1)$ and $P(C = 1, S = 1)$ for the joint distribution.