Machine Learning Worksheet 1

Probability Theory

1 Basic Probability

Problem 1: A secret government agency has developed a scanner which determines whether a person is a terrorist. The scanner is fairly reliable; 95% of all scanned terrorists are identified as terrorists, and 95% of all upstanding citizens are identified as such. An informant tells the agency that exactly one passenger of 100 aboard an aeroplane in which you are seated is a terrorist. The agency decide to scan each passenger and the shifty looking man sitting next to you is tested as "TERRORIST". What are the chances that this man *is* a terrorist? Show your work!

The Bernoulli random variable T indicates if a person is a terrorist or not, i.e., from the informant's hint:

$$p(T = 1) = 0.01, \quad p(T = 0) = 0.99.$$

The Bernoulli random variable S indicates the terrorist scanner outcome, i.e., from the specifications:

$$p(S = 1 \mid T = 1) = 0.95 \Rightarrow p(S = 0 \mid T = 1) = 0.05,$$
$$p(S = 0 \mid T = 0) = 0.95 \Rightarrow p(S = 1 \mid T = 0) = 0.05.$$

We are interested in $p(T = 1 \mid S = 1)$, which we obtain by Bayes' rule:

$$p(T = 1 \mid S = 1) = \frac{p(S = 1 \mid T = 1)p(T = 1)}{p(S = 1 \mid T = 1)p(T = 1) + p(S = 1 \mid T = 0)p(T = 0)} = \frac{19}{118} \approx 0.16.$$

Note that in the denominator, we compute $p(S = 1)$ using the law of total probability.

Problem 2: A fair coin is tossed twice. Whenever it turns up heads, a red ball is placed into a box, otherwise a white ball. Afterwards, balls are drawn from the box three times in succession (replacing the drawn ball ever time). It is found that on all three occasions a red ball is drawn. What is the probability that both balls in the box are red? Show your work!

Denote by RRR the event that 3 red balls are drawn. Similarly, denote by rr the event that 2 red balls are placed in the box, rw the event that first a white and then a red ball are placed in the box, and wr and ww for the remaining two possibilities. Since the coin is fair, we know that

$$p(rr) = p(rw) = p(wr) = p(ww) = \frac{1}{4}.$$

Furthermore, by using independence of the draws,

$$p(RRR \mid rr) = 1, \quad p(RRR \mid rw) = p(RRR \mid wr) = \left(\frac{1}{2}\right)^3 = \frac{1}{8}, \quad p(RRR \mid ww) = 0.$$

Therefore, by Bayes' rule:

$$p(rr \mid RRR) = \frac{p(RRR \mid rr)p(rr)}{p(RRR)} = \frac{1/4}{5/16} = \frac{4}{5}$$

with

$$p(RRR) = p(RRR \mid rr)p(rr) + p(RRR \mid wr)p(wr) + p(RRR \mid rw)p(rw) + p(RRR \mid ww)p(ww).$$

Problem 3: A fair coin is flipped until heads shows up for the first time. What is the expected number of tails T and the expected number of heads H in any one run of this experiment? Show your work.

Hint: While there is a very short solution to this problem for people with a good intuition, the rest of us might need to look at the geometric series *and its properties. You may use them without proof.*

The expected number of heads $E[H]$ is 1, by definition of the problem.

Since the situation after getting tails once is equivalent to the opening situation, we can write down the recurrence relation

$$E[T] = \frac{1}{2}(1 + E[T]) + \frac{1}{2}0 \quad \Rightarrow E[T] = 1.$$

While this solution is short, it requires some good intuition. However, it can also be solved rigorously. To this end, let θ denote the probability of tails.

$$p(T = t) = \theta^t(1 - \theta)$$

The expected number of tails is

$$E[T] = \sum_{t=0}^{\infty} t\theta^t(1 - \theta) = (1 - \theta)\theta \sum_{t=1}^{\infty} t\theta^{t-1}.$$

Notice that the index has changed. The argument of the series on the rhs now looks like a derivative of θ^t, which would be the *geometric series*. In fact, since the geometric series converges absolutely (in this case simply because it is non-negative by definition), we are allowed to interchange the infinite sum and the derivative!

$$E[T] = (1 - \theta)\theta \sum_{t=1}^{\infty} \frac{\mathrm{d}}{\mathrm{d}\theta}\theta^t = (1 - \theta)\theta \frac{\mathrm{d}}{\mathrm{d}\theta} \sum_{t=0}^{\infty} \theta^t$$

Notice again the changed index. We know that the geometric series converges to $\frac{1}{1-\theta}$.

$$E[T] = (1 - \theta)\theta \frac{\mathrm{d}}{\mathrm{d}\theta} \frac{1}{1 - \theta} = (1 - \theta)\theta \frac{1}{(1 - \theta)^2} = \frac{\theta}{1 - \theta}$$

Inserting $\theta = 0.5$ for a fair coin yields $E[T] = 1$ as above.

Nota bene: This *derivative trick* can generally be very useful when calculating expectations over discrete variables with infinitely many values.

Problem 4: Calculate mean and variance of a uniform random variable X on the interval $[a, b], a < b$ with probability density function

$$p(x) = \begin{cases} \frac{1}{b-a} & a \le x \le b, \\ 0 & \text{elsewhere.} \end{cases}$$

$$E[X] = \int_{-\infty}^{+\infty} xp(x)\,\mathrm{d}x = \int_a^b \frac{x}{b-a}\,\mathrm{d}x = \frac{1}{b-a}\left[\frac{x^2}{2}\right]_a^b = 0.5 \cdot \frac{b^2-a^2}{b-a} = \frac{a+b}{2}$$

$$E[X^2] = \int_{-\infty}^{+\infty} x^2 p(x)\,\mathrm{d}x = \int_a^b \frac{x^2}{b-a}\,\mathrm{d}x = \frac{1}{b-a}\left[\frac{x^3}{3}\right]_a^b = \frac{a^2+ab+b^2}{3}$$

$$\mathrm{Var}[X] = E[X^2] - E[X]^2 = \frac{(a+b)^2 - ab}{3} - \left(\frac{a+b}{2}\right)^2 = \frac{(a-b)^2}{12}$$

Problem 5: Let X and Y be random variables with joint density $p(x, y)$. Prove the *tower properties*,

$$E[X] = E_Y[E_{X|Y}[X]],$$
$$\mathrm{Var}[X] = E_Y[\mathrm{Var}_{X|Y}[X]] + \mathrm{Var}_Y[E_{X|Y}[X]].$$

$E_{X|Y}[X]$ and $\mathrm{Var}_{X|Y}[X]$ denote the expectation and variance of X under the conditional density $p(x \mid y)$.

The first tower property is obtained by clever reordering (and, technically, Fubini's Theorem from first to second line):

$$E_Y[E_{X|Y}[X]] = \iint xp(x \mid y)\,\mathrm{d}x\,p(y)\,\mathrm{d}y = \iint xp(x \mid y)p(y)\,\mathrm{d}x\,\mathrm{d}y = \iint xp(x,y)\,\mathrm{d}x\,\mathrm{d}y$$
$$= \int x \int p(x,y)\,\mathrm{d}y\,\mathrm{d}x = \int xp(x)\,\mathrm{d}x = E[X]$$

If you have a hard time following these results, replace integrals with sums.

The second result is a little more tricky:

$$E_Y[\mathrm{Var}_{X|Y}[X]] = \int \mathrm{Var}_{X|Y}[X]p(y)\,\mathrm{d}y = \iint \big(x - E_{X|Y}[X]\big)^2 p(x \mid y)\,\mathrm{d}x\,p(y)\,\mathrm{d}y$$
$$= \underbrace{\int x^2 \int p(x,y)\,\mathrm{d}y\,\mathrm{d}x}_{E_X[X^2]} - \underbrace{2\iint xp(x \mid y)\,\mathrm{d}x E_{X|Y}[X]p(y)\,\mathrm{d}y}_{2\int E_{X|Y}[X]^2 p(y)\,\mathrm{d}y} + \underbrace{\int E_{X|Y}[X]^2 \int p(x,y)\,\mathrm{d}x\,\mathrm{d}y}_{\int E_{X|Y}[X]^2 p(y)\,\mathrm{d}y}$$
$$= E_X\left[X^2\right] - E_Y\left[E_{X|Y}[X]^2\right]$$

Wherever possible, we have exchanged the order of integration. This allows moving independent terms out of the inner integral. In the first term, we recognize a part of the classical variance

equation $\mathrm{Var}[X] = E_X\left[X^2\right] - E_X[X]^2$, which hints at the direction of this proof.

$$\begin{aligned}
\mathrm{Var}_Y[E_{X|Y}[X]] &= \int \left(E_{X|Y}[X] - E_Y[E_{X|Y}[X]]\right)^2 p(y)\,\mathrm{d}y \\
&= \int \left(E_{X|Y}[X] - E_X[X]]\right)^2 p(y)\,\mathrm{d}y \\
&= \int E_{X|Y}[X]^2 p(y)\,\mathrm{d}y - 2E_X[X]\int E_{X|Y}[X]p(y)\,\mathrm{d}y + E_X[X]^2\int p(y)\,\mathrm{d}y \\
&= E_Y[E^2_{X|Y}[X]] - 2E^2_X[X] + E^2_X[X] = E_Y[E^2_{X|Y}[X]] - E^2_X[X]
\end{aligned}$$

We have made use of the first tower property twice.

In total, we get:

$$E_Y[E_{X|Y}[X]] + \mathrm{Var}_Y[E_{X|Y}[X]] = E_X[X^2] - E^2_X[X] = \mathrm{Var}[X]$$

There is an alternative, shorter way, exploiting the fact that the tower property holds for general expectations:

$$\begin{aligned}
&E_Y[\mathrm{Var}_{X|Y}[X]] + \mathrm{Var}_Y[E_{X|Y}[X]] \\
=&E_Y[E_{X|Y}[X^2]] - E_Y[(E_{X|Y}[X])^2] + E_Y[(E_{X|Y}[X])^2] - E^2_Y[E_{X|Y}[X]] \\
=&E_Y[E_{X|Y}[X^2]] - E^2_X[X] \\
=&E_X[X^2] - E^2_X[X]
\end{aligned}$$

2 Probability Inequalities

Inequalities are useful for bounding quantities that might otherwise be hard to compute. A famous example is the Markov inequality

$$p(X > c) \leq \frac{E[X]}{c}$$

for a *non-negative* random variable X and a constant $c > 0$. From it, it is relatively easy to prove the Chebyshev inequality

$$p(|X - E[X]| > c) \leq \frac{\mathrm{Var}(X)}{c^2}$$

for arbitrary X with finite variance.

With the help of Chebyshev's inequality, one can prove (a weak version of) the *law of large numbers*, which roughly states that the empirical mean of n i.i.d. random variables X_i converges to the true mean for $n \to \infty$. More formally, for any $\epsilon > 0$

$$p\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - E[X_i]\right| > \epsilon\right) \to 0 \qquad \text{as } n \to \infty. \tag{1}$$

Problem 6: Prove eq. (1). You may assume that the X_i have finite variance $\mathrm{Var}[X_i]$. You may further use Markov's and Chebyshev's inequalities without proof.

(We highly recommend to practise your "proof skills" on them, though. The proofs are technical, but very short.)

Proof of the Markov inequality:

$$E[X] = \int_0^\infty p(x)x\,\mathrm{d}x = \int_0^c p(x)x\,\mathrm{d}x + \int_c^\infty p(x)x\,\mathrm{d}x \geq \int_c^\infty p(x)x\,\mathrm{d}x \geq c\int_c^\infty p(x)\,\mathrm{d}x = cp(X > c)$$

Proof of the Chebyshev inequality:

$$p(|X - \mathbb{E}[X]| > c) = p((X - \mathbb{E}[X])^2 > c^2) \leq \frac{\mathrm{Var}(X)}{c^2}$$

With these given, let $E[X]$ and $\mathrm{Var}[X]$ denote the shared mean and variance of the i.i.d. variables. $\mathrm{Var}[X]$ is finite (and thus exists) by the assignment.

First, we observe that

$$\left|\frac{1}{n}\sum_{i=1}^n X_i - E[X_i]\right| > \epsilon \Leftrightarrow \left|\sum_{i=1}^n X_i - E[X_i]\right| > n\epsilon.$$

Second, we define $Y = \sum_{i=1}^n X_i$ and observe that $E[Y] = \sum_{i=1}^n E[X_i] = nE[X]$.

$$p\left(\left|\frac{1}{n}\sum_{i=1}^n X_i - E[X_i]\right| > \epsilon\right) = p\left(|Y - E[Y]| > n\epsilon\right) \leq \frac{\mathrm{Var}[Y]}{n^2\epsilon^2} = \frac{\mathrm{Var}[X]}{n\epsilon^2} \to 0$$

The inequality is an application of Chebyshev's. In the last step, we used the fact that the X_i are independent and we can simply sum up their (identical) variances.