Example sheet 0

Prerequisites

IB Data Science—DJW—2020/2021

Data Science builds on the probability theory you learnt in IA *Introduction to Probability*. It also relies on basic calculus from IA *Maths for NST*. This example sheet reviews the material that you need to know. Please look through, and make sure you remember how to answer these questions! Solutions are provided.

For supervisors: this example sheet is not intended for supervision.

Rules of probability (IA Probability lectures 1, 2)

Understand what is meant by *sample space*, written Ω, and know that $\mathbb{P}(\Omega) = 1$. Be able to reason about probabilities of events with Venn diagrams. Know the core definitions and laws, both in the standard form and in the conditional form …

Conditional probability:

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A, B)}{\mathbb{P}(B)} \text{ if } \mathbb{P}(B) > 0 \qquad \mathbb{P}(A \mid B, C) = \frac{\mathbb{P}(A, B \mid C)}{\mathbb{P}(B \mid C)} \text{ if } \mathbb{P}(B \mid C) > 0$$

If A and B are independent; or if A and B are conditionally independent given C:

$$\mathbb{P}(A, B) = \mathbb{P}(A)\,\mathbb{P}(B) \qquad \mathbb{P}(A, B \mid C) = \mathbb{P}(A \mid C)\,\mathbb{P}(B \mid C)$$
$$\mathbb{P}(A \mid B) = \mathbb{P}(A) \qquad \mathbb{P}(A \mid B, C) = \mathbb{P}(A \mid C)$$

Sum rule, and the law of total probability, for events $\{B_1, B_2, \dots\}$ that partition the sample space (i.e. for events that are mutually exclusive and where $\bigcup_i B_i = \Omega$):

$$\mathbb{P}(A) = \sum_i \mathbb{P}(A, B_i) \qquad \mathbb{P}(A \mid C) = \sum_i \mathbb{P}(A, B_i \mid C)$$
$$= \sum_i \mathbb{P}(A \mid B_i)\,\mathbb{P}(B_i) \qquad = \sum_i \mathbb{P}(A \mid B_i, C)\,\mathbb{P}(B_i \mid C)$$

Bayes's rule:

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(B \mid A)\,\mathbb{P}(A)}{\mathbb{P}(B)} \qquad \mathbb{P}(A \mid B, C) = \frac{\mathbb{P}(B \mid A, C)\,\mathbb{P}(A \mid C)}{\mathbb{P}(B \mid C)}$$

when $\mathbb{P}(B) > 0$ or $\mathbb{P}(B \mid C) > 0$ respectively.

Question 1 (Elementary probability). A card is drawn at random from a pack. Event A is 'the card is an ace', event B is 'the card is a spade', event C is 'the card is either an ace, or a king, or a queen, or a jack, or a 10'. Compute the probability that the card has (i) one of these properties, (ii) all of these properties.

Question 2 (Elementary probability). A biased die has probabilities p, $2p$, $3p$, $4p$, $5p$, $6p$ of throwing 1, 2, 3, 4, 5, 6 respectively. Find p. What is the probability of throwing an even number?

Question 3 (Independence). We roll a die twice, and get the answers X and Y. Assume the two rolls are independent. Consider the events $E = \{X = 1\}$, $F = \{Y = 6\}$, and $G = \{X + Y = 7\}$. (i) Are E and F independent? (ii) Are E and G independent? (iii) Are E, F, and G independent?

Question 4 (Law of total probability). There are three boxes each containing a different number of lightbulbs. The first box has 10 bulbs of which 4 are dead, the second has 6 bulbs of which 1 is dead, and the third has 8 bulbs of which 3 are dead. What is the probability of a dead bulb being selected when a bulb is chosen at random from one of the three boxes?

Question 5 (Conditional probability). A deck of 52 playing cards is randomly divided into 4 piles of 13 cards each. What is the probability that each pile has exactly one ace?

Question 6 (Advanced elementary probability). Players A and B roll a six-sided die in turn. If a player rolls 1 or 2 that player wins and the game ends; if a player rolls 3 the other player wins and the game ends; otherwise the turn passes to the other player. A has the first roll. What is the probability (i) that B gets a first throw and wins on it? (ii) that A wins before A's second throw? (iii) that A wins, if the game is played until there is a winner?

Question 7 (Bayes's rule). A screening test is 94% effective in detecting COVID, when the person has the disease. The test yields a 'false positive' for 1% of healthy persons tested. Suppose 0.4% of the population has the disease. (i) What is the probability you actually have COVID, if you test positive? (ii) What is the probability you actually have COVID, if you test negative?

Question 8 (Elementary probability). Derive Bayes's rule from the other three core definitions and laws of probability. Explain carefully which of the three you are using at each step.

Random variables (IA Probability lectures 3, 4)

Know what a random variable is. Understand what is meant by discrete and continuous random variables. Describe them using a cumulative distribution function, and using a probability mass function (pmf) or probability density function (pdf) as appropriate. Know that density functions integrate to 1. Know what is meant by independent random variables. (Marginals and joint distributions from lecture 7 are also important, and they will be studied further in this course.)

Be aware of the following random variables and their uses: Bernoulli, Binomial, Poisson, Geometric; Exponential, Gaussian, Uniform.

Know what is meant by expectation, variance, and standard deviation of a random variable. Be familiar with the linearity of expectation, and the implications for variance …

$$\begin{aligned}
&\mathbb{E}(aX + b) = a(\mathbb{E}\,X) + b \\
&\mathbb{E}(X + Y) = (\mathbb{E}\,X) + (\mathbb{E}\,Y) \\
&\operatorname{Var}(aX + b) = a^2(\operatorname{Var} X) \\
&\operatorname{Var}(X + Y) = (\operatorname{Var} X) + (\operatorname{Var} Y) \quad \text{if } X, Y \text{ independent}
\end{aligned}$$

and with the Law of the Unconscious Statistician

$$\mathbb{E}\,h(X) = \sum_x h(x)\,\mathrm{pmf}(x) \quad \text{or} \quad \int_x h(x)\,\mathrm{pdf}(x)\,dx$$

Question 9 (Expectation). Let X be a random variable. Show that $\operatorname{Var} X = \mathbb{E}\, X^2 - (\mathbb{E}\, X)^2$.

Note: by convention, $\mathbb{E}$ is taken to have lower precedence than multiplication and power, and higher precedence than addition and subtraction. So the expression of interest is $(\mathbb{E}(X^2)) - (\mathbb{E}(X))^2$.

Question 10 (cdf and pdf). Derive the mean and variance of the Exponential distribution, which has density

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

What is the probability that X takes a value in excess of two standard deviations from the mean?

Question 11 (Law of the Unconscious Statistician). The University bus arrives at the Computer Lab bus stop at 7am, 7:15am, and so on at 15 minute intervals. You arrive at the bus stop at a time uniformly distributed in the interval [1pm, 1:20pm]. Let W be the length of time you wait. Find the expected value of W.

Optimization (IA Maths for NST)

Be able to optimize a function of one or more variables, by differentiating and looking for stationary points.

Question 12. A coal bunker is to be constructed on the side of a house. Assuming that it is a cuboid of given volume V, find the shape that minimizes the external surface area.

Question 13. *More examples of finding stationary points. These two specific cases crop up again and again in data science. You should be able to solve them blindfolded.*

(a) Find the value of $p \in [0, 1]$ that maximizes $p^a(1-p)^b$, where a and b are both positive.

(b) Find the value of $p \in [0, 1]$ that maximizes $\log\big(p^a(1-p)^b\big)$, where a and b are both positive.

(c) Find the values of $\mu \in \mathbb{R}$ and $\sigma > 0$ that jointly maximize

$$\Big(\frac{1}{\sqrt{2\pi\sigma^2}}\Big)^n \exp\Big(-\sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2}\Big)$$

(d) Find the values of $\mu \in \mathbb{R}$ and $\rho > 0$ that jointly maximize

$$\Big(\frac{1}{\sqrt{2\pi\rho}}\Big)^n \exp\Big(-\sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\rho}\Big)$$

What do you notice about the solutions to (a) *versus* (b), and about the solutions to (c) *versus* (d)?

numpy vectorized syntax

Be able to translate `for` loops into numpy vectorized operations. Be able to use `np.linspace`, `x @ y`, `np.default_rng`, `np.where`, and vector indexing notation.

Question 14. Replace the `for` loop with vectorized operations. Here `size` is an integer.

```
def rgalaxies(size, p=[0.28, 0.54, 0.18], μ=[9740, 21300, 15000], σ=[340, 1700, 10600]):
    res = []
    for _ in range(size):
        cluster = np.random.choice([0,1,2], p=p)
        μi,σi = μ[cluster], σ[cluster]
        x = np.random.normal(loc=μi, scale=σi)
        res.append(x)
    return res
```

Question 15. Given a vector $[x_1, \ldots, x_n]$, and parameters p, μ, and σ as in question 14, write numpy vectorized code to compute

$$\sum_{i=1}^{n} \log\Big[\sum_{k=1}^{3} p_k \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{-(x_i-\mu_k)^2/2\sigma_k^2}\Big]$$

Your code should be vectorized over $i = 1, \ldots, n$, but it need not be vectorized over $k = 1, 2, 3$. (Vectorizing over both i and k needs numpy's 'broadcast semantics', which are frequently needed in deep learning, but are overkill for IB Data Science.)