70015 Mathematics for Machine Learning: Exercises

Mark van der Wilk, Yingzhen Li*
{m.vdwilk,yingzhen.li}@imperial.ac.uk

October 5, 2022

Contents

1 Notation of probabilities

In this course we will use the notation for probabilities that is common in machine learning. The main advantage is that this notation is shorter, although it does leave certain things implicit. We include this to reduce confusion.

Consider a probability space $(\Omega, \mathcal{E}, \mathbb{P})$ with sample space Ω (all possible outcomes of a random procedure), event space $\mathcal{E}$ (the set of all sets of outcomes that we assign a probability to), and probability function $\mathbb{P} : \mathcal{E} \to [0, 1]$ (a function that assigns a probability to an event), with a random variable $X : \Omega \to \mathbb{R}^D$.

- With $\mathbb{P}(E)$ we denote the probability of an event $E \in \mathcal{E}$, where E is a set of outcomes.
- Following the usual convention, we use the same notation when considering random variables, e.g. $\mathbb{P}(X < 2)$ is short for $\mathbb{P}(\{s \in \Omega : X(s) < 2\})$ (see §6.1 in `50008` *Probability & Statistics*).
- We usually work directly with random variables, and specify all properties using a probability mass function (pmf) or probability density function (pdf). For a specific outcome of the random variable α, we write:

$$\mathbb{P}(X = \alpha) = p_X(\alpha) \qquad \text{for a pmf } p_X(\cdot)\,, \tag{1}$$

$$\mathbb{P}(X \in [a, b]) = \int_a^b p_X(\alpha) d\alpha \qquad \text{for a pdf } p_X(\cdot) \text{ with } \alpha \in \mathbb{R}\,, \tag{2}$$

$$\mathbb{P}(X \in A) = \int_A p_X(\alpha) d\alpha \qquad \text{for a pdf } p_X(\cdot) \text{ with } \alpha \in \mathbb{R}^D\,. \tag{3}$$

- Sometimes we may write vectors in boldface, i.e. $\mathbf{x} \in \mathbb{R}^D$. We won't always though, so keep track of how we define variables!

*Many thanks to teaching assistants Carles Balsells Rodas, and Alex Spies for their solutions and improvements to the document.

- We generally denote outcomes of random variables without referring explicitly to the random variable itself. For example, when we refer to an outcome $\mathbf{x}$, we implicitly know there is a random variable that can take this value. We usually denote this as the capital, for example here X.
- Sometimes we abuse notation, and drop the random variable when denoting distributions when the argument of the function identifies it, e.g. $p(\mathbf{x}) = p_X(\mathbf{x})$.
- If we want to be explicit about the random variable that we are evaluating the density/mass of, I will write e.g. $p_{X,Y}(\mathbf{x}, \mathbf{y}) = p_{X|Y}(\mathbf{x}|\mathbf{y})p_Y(\mathbf{y})$.
- Expectations can be denoted in two ways:

$$\mathbb{E}_X[f(X)] \qquad \text{to emphasise that X is random, if it is clear what its distribution is}\,, \tag{4}$$

$$\mathbb{E}_{p(\mathbf{x})}[f(\mathbf{x})] \qquad \text{to emphasise that we will be integrating over the distribution } p(\mathbf{x})\,. \tag{5}$$

  In both cases this corresponds to the integral $\int p(\mathbf{x})f(\mathbf{x})\mathrm{d}\mathbf{x}$.
- Often, densities and pmfs can be discussed in exactly the same way, if we think of the density of a discrete RV as a sum of delta functions. I.e. $p(\mathbf{x}) = \sum_o \delta(\mathbf{x} - \mathbf{x}_o)p_o$, where $\{\mathbf{x}_o\}$ is the set of discrete possible outcomes that X can take, and p_o are their corresponding probabilities. This allows us to write an expectation as an integral, regardless of whether the RV is continuous or discrete, because for discrete RVs we get:

$$\mathbb{E}_{p(\mathbf{x})}[f(\mathbf{x})] = \int p(\mathbf{x})f(\mathbf{x})\mathrm{d}\mathbf{x} = \int \sum_o \delta(\mathbf{x} - \mathbf{x}_o)p_o f(\mathbf{x})\mathrm{d}\mathbf{x} = \sum_o f(\mathbf{x}_o)p_o\,. \tag{6}$$

  (A delta function has the property that $\int_A \delta(\mathbf{x})\mathrm{d}\mathbf{x}$ is 1 if $0 \in A$, and 0 otherwise. Linearity of integrals still holds. It can often be seen as the limit of a Gaussian distribution with zero variance.)

2 Formula Sheet

- Gaussian probability density function (pdf) with input $\mathbf{x} \in \mathbb{R}^D$, denoted as $\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{D}{2}}|\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\intercal \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right)\,. \tag{7}$$

- For a joint Gaussian density

$$p\left(\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix}\right) = \mathcal{N}\left(\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix}; \begin{bmatrix}\mathbf{m}_\mathbf{x}\\ \mathbf{m}_\mathbf{y}\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_{\mathbf{xx}} & \boldsymbol{\Sigma}_{\mathbf{xy}}\\ \boldsymbol{\Sigma}_{\mathbf{yx}} & \boldsymbol{\Sigma}_{\mathbf{yy}}\end{bmatrix}\right)\,, \tag{8}$$

  we have the conditional density

$$p(\mathbf{x}\,|\,\mathbf{y}) = \mathcal{N}\big(\mathbf{x}; \quad \mathbf{m}_\mathbf{x} + \boldsymbol{\Sigma}_{\mathbf{xy}}\boldsymbol{\Sigma}_{\mathbf{yy}}^{-1}(\mathbf{y} - \mathbf{m}_\mathbf{y}), \quad \boldsymbol{\Sigma}_{\mathbf{xx}} - \boldsymbol{\Sigma}_{\mathbf{xy}}\boldsymbol{\Sigma}_{\mathbf{yy}}^{-1}\boldsymbol{\Sigma}_{\mathbf{yx}}\big)\,. \tag{9}$$

3 Warm-up Exercises

To start, here are some exercises which test knowledge which is assumed in the course.

3.1 Probability Theory

We assume that you are familiar with probability theory up to the Computing 2nd year `50008` *Probability & Statistics* course. Here are some questions to serve as a refresher. Students who are not familiar with this background should refer to the notes of `50008` *Probability & Statistics* or relevant chapters of [Deisenroth et al., 2020]. **We recommend you look at these questions when/before the course starts**. If you need a refresher, or if you do not know the notation, refer to the `50008` *Probability & Statistics* notes, or discuss with a TA.

Question 1 (Set Theory and Probability)**.** Using the three axioms of probability show that

a. Write down the sample space of a dice. In your notation, use the set A to denote the event of a 3 or 4 occurring. What is the complement of A, denoted $\neg A$?

b. For a problem about lengths, we have a sample space $\Omega = [0, 1]$. For $A = (0.3, 0.4]$, what is $\neg A$?

c. $\mathbb{P}(\neg A) = 1 - \mathbb{P}(A)$

d. $\mathbb{P}(\varnothing) = 0$, where $\varnothing$ is the empty set

e. $0 \leq \mathbb{P}(A) \leq 1$

f. $A \subseteq B \implies \mathbb{P}(A) \leq \mathbb{P}(B)$
Hint: Consider the following definition. $B \backslash A = \{x \in B : x \notin A\}$

g. $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$

h. (*) if $\{A_i\}_{i=1}^{\infty} \subseteq \Omega$ and $A_i \subseteq A_{i+1} \forall i$ then:

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \lim_{i \longrightarrow \infty} \mathbb{P}(A_i)$$

Hint: Use axiom 3. *: The emphasis of this course isn't on these kinds of details, even though this should be doable with 1st-year calculus.

i. For two mutually exclusive events A, B, what is $\mathbb{P}(A \cup B)$?

See Deisenroth et al. [2020, §6.1.2] for a general overview, and §4, §§5.1-5.4 of `50008` *Probability & Statistics* for more details.

Question 2 (Independent events)**. Independent events don't come up as much as independent random variables, so it's ok to just follow this answer, rather than spending lots of time on it.** When tossing two coins (where we care about the order), we have a sample space $\Omega = \{HH, HT, TH, TT\}$.

a. What outcomes are contained in the event that corresponds to the the first coin being heads? We denote the event E_{1H}, and others similarly.

b. If you assume that all outcomes have equal probability, show that E_{1H} and E_{2T} are independent.

c. If you assume that E_{1H} and E_{2H} are independent and 0.5 each, show that all outomes must have equal probability.

See §5.3.3 in `50008` *Probability & Statistics*.

Question 3 (Random Variables)**.** Consider throwing two fair dice.

a. What is the sample space for all outcomes that you can get from throwing two dice? We specify the probability of each outcome to be the same.

b. Define two random variables A, B which map the outcome to the face value on each die respectively. Find the probability mass function for A from the probability on outcomes. The answer will work from the definition of a random variable, but you will probably intuitively get the right answer as well.

c. Show that A and B are independent.

d. Define the random variable $C = A + B$. Derive the probability mass function of C.

See §6 of `50008` *Probability & Statistics*.

Question 4 (Continuous Random Variables)**.** Consider the random variable X with a probability density $p(x) = C \cdot x$ when $x \in [0, 1]$ and 0 elsewhere.

a. Calculate C.

b. Calculate $\mathbb{P}(0.3 \leq X \leq 0.75)$.

c. Calculate $\mathbb{P}(X \in [0.3, 0.75] \cup [0.8, 0.9])$.

d. Calculate $\mathbb{E}_X[X]$, $\mathbb{E}_X\left[X^2\right]$, $\mathbb{V}_X[X]$.

Check your answers by performing numerical integration, e.g. in Python.

See §6.3, §7 of `50008` *Probability & Statistics* or Deisenroth et al. [2020, §6.2.2].

Question 5 (Joint Discrete Random Variables)**.** Consider two random variables A, C, where A is the outcome of one die, and C gives the sum of A and the sum of another die B.

a. From intuition, write a table of $\mathbb{P}(C = c|A = a)$, which we use to denote the probability of C taking the value c, if we know that A has taken the value a.

b. Write a table of $\mathbb{P}(C = c, A = a)$. To help you think it through, consider a tree of outcomes that can occur. This helps illustrate independence between outcomes, which helps you figure out when you can multiply probabilities.

c. From the values in the table $\mathbb{P}(C = c, A = a)$ find $\mathbb{P}(2 \leq C \leq 4)$ and $\mathbb{P}(2 \leq C \leq 4, 2 \leq A \leq 4)$.

We will cover conditional probability more later, but for now just think it through.

Question 6 (Multivariate Integration)**.** Consider two continuous random variables X, Y with joint density $p(x, y) = C \cdot (x^2 + xy)$.

a. Find C.

b. Find $\mathbb{P}(0.3 \leq X \leq 0.5)$.

c. Find $\mathbb{P}(X < Y)$. Perform the integration twice in both orders, once integrating over x first, once by integrating over y first.

d. **Bonus:** Convince yourself that you know how to do this for $p(x, y, z) = C \cdot (x^2 + xyz)$ as well.

Check your answers by performing numerical integration, e.g. in Python.

Question 7 (Statistics)**.** Recall the following statistical terminology.

a. What is a statistic?

b. What is an estimator?

c. What is a consistent estimator?

3.2 Linear Algebra

Question 8 (Dot product)**.** Compute $\mathbf{x}^\top \mathbf{y}$ where $\mathbf{x} = (1, -2, 5, -1)^\top$ and $\mathbf{y} = (0, 4, -3, 7)^\top$.

Question 9 (Matrix product)**.** Compute $\mathbf{y} = A\mathbf{x}$ as well as the ℓ_2 norm of $\mathbf{x}$ and $\mathbf{y}$, where

$$A = \begin{pmatrix} -1 & 4 & 7 & 2 \\ 3 & -2 & -1 & 0 \\ 5 & 3 & 0 & -1 \end{pmatrix}, \quad \mathbf{x} = (-3, 2, 1, 3)^\top.$$

Question 10 (Basis)**.** Which of the following set of vectors are basis for $\mathbb{R}^2$?

1. $\{(1, 1), (1, 0)\}$
2. $\{(2, 4), (3, -1)\}$
3. $\{(1, -1), (0, 2), (2, 1)\}$
4. $\{(2, -1), (-2, 1)\}$
5. $\{(0, 3)\}$

Question 11 (Span of vectors)**.** Which of the following points are within the span of $\{(-1, 0, 2), (3, 1, 0)\}$?

1. $(0, 1, 1)$
2. $(1, 1, 4)$
3. $(2, 1, 1)$

4. $(-3, 4, 2)$
5. $(0, 0, 0)$

Question 12 (Rotation matrix in $\mathbb{R}^2$)**.** What is the 2×2 matrix that rotates all the non-zero vectors in $\mathbb{R}^2$ by $45°$ counter-clockwise?

Question 13 (Linear equations)**.** Given the following system of linear equations:

$$\begin{aligned} x + 2y &= 2 \\ 3x + 2y + 4z &= 5 \\ -2x + y - 2z &= -1 \end{aligned}$$

Answer the following questions:

a Writing this system in a matrix form $A\mathbf{x} = \mathbf{b}$ with $\mathbf{x} = (x, y, z)^\top$. What are A and $\mathbf{b}$?

b Solve this system, or show that the solution does not exist.

c What is the rank of A?

4 Lecture 1: Vectors & Probability

Question 14 (Vector notation)**.** We define the probability density on the vector $\mathbf{x} \in \mathbb{R}^3$ with all elements $0 \le x_k \le 1$ as

$$p(\mathbf{x}) = \frac{1}{C}(x_1^2 + x_1 x_2 + x_2^2 + 2x_2 x_3)\,. \tag{10}$$

Put this into notation that only uses $\mathbf{x}$ as a single whole vector.

Question 15 (Vector independence)**.** Consider the density on $\mathbf{x} \in \mathbb{R}^4$ with all elements $0 \le x_k \le 1$ as

$$p(\mathbf{x}) = \mathbf{x}^\intercal \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \end{bmatrix} \mathbf{x}\,. \tag{11}$$

Now divide up $\mathbf{x}$ into two sub vectors $\mathbf{y} = [x_2, x_4]^\intercal$ and $\mathbf{z} = [x_1, x_3]^\intercal$. Show that $\mathbf{y} \perp\!\!\!\perp \mathbf{z}$, i.e. that they are independent.

Question 16 (Noise conditional independence)**.** Consider the probability of the data in linear regression, for a fixed setting of the parameters $\boldsymbol{\theta}$ and given inputs $X \in \mathbb{R}^{N \times D}$ where $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$:

$$p(\mathbf{y}|\boldsymbol{\theta}, X) = \mathcal{N}\big(\mathbf{y}; \mathbf{x}, \sigma^2 \mathrm{I}\big) \tag{12}$$

Show that all y_ns are independent, for a fixed setting of the parameters $\boldsymbol{\theta}$ and given inputs X.

Question 17 (Maximum likelihood revision)**.** For a Gaussian distribution with mean μ and variance σ^2.

a. Derive the probability distribution for N iid draws.

b. Derive the maximum likelihood estimator for the mean μ.

Question 18 (Maximum likelihood and minimum loss)**.** Show that the solution to the Maximum Likelihood estimator for linear regression is the same as the minimum squared loss estimator.

5 Warm-up Exercises Answers

5.1 Warm-up Exercises

Question 1 – Set Theory and Probability

a. Sets can contain anything, so we can choose a representation using abstract symbols $\Omega = \{$⚀, ⚁, ⚂, ⚃, ⚄, ⚅$\}$. Alternatively, we can represent each of the outcomes as a number $\Omega = \{1, 2, 3, 4, 5, 6\}$. finish

b. $P(\neg A) = 1 - P(A)$

Let us consider a collection of events contained in the sample space $\{A_1, \dots, A_N\} \subseteq \Omega$. Let us select the first i events (where $i \leq N$) and denote them as A. The rest of them will be the complementary set, denoted as $\neg A$.

$$A = \{A_1, \dots, A_i\} \neg A = \{A_{i+1}, \dots, A_N\} A \cap \neg A = \varnothing$$

Using axiom (2), we have

$$P(A_1 \cup A_2 \cup \cdots \cup A_N) = P(\Omega) = 1$$

And using axiom (3), we have

$$P(A_1, \dots, A_i) + P(A_{i+1}, \dots, A_N) = P(A_1 \cup A_2 \cup \cdots \cup A_N)$$

$$P(A) + P(\neg A) = 1$$

Thus

$$P(\neg A) = 1 - P(A)$$

I think this can be simplified? Can we not just say $\Omega = A \cup \neg A$ and then combine axioms 2 and 3?

c. $P(\varnothing) = 0$, where $\varnothing$ is the empty set

We can just consider the sample space, Ω, where its complementary is the empty set $\varnothing$. Using the previous property and axiom 2, we have.

$$P(\Omega) = 1$$

$$P(\varnothing) = P(\neg\Omega) = 1 - P(\Omega) = 1 - 1 = 0$$

$$P(\varnothing) = 0$$

d. $0 \leq P(A) \leq 1$

Here we can also use property (a) and the first axiom. Consider any event A.

$$P(A) \geq 0$$

$$P(\neg A) = 1 - P(A) \geq 0$$

$$1 \geq P(A)$$

We can join the previous inequalities and obtain the following.

$$0 \leq P(A) \leq 1$$

e. $A \subseteq B \implies P(A) \leq P(B)$

Hint: Consider the following definition. $B \backslash A = \{x \in B : x \notin A\}$

We can construct B as the union of two disjoint sets.

$$B = B \backslash A \cup A$$

where $B \backslash A \cap A = \varnothing$ by definition of $B \backslash A$. Let us use axiom 3 and 1.

$$P(B) = P(B \backslash A) + P(A) \geq P(A)$$

where by axiom 1, we have $P(B \backslash A) \geq 0$. Thus

$$P(A) \leq P(B)$$

f. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Let us define the union $(A \cup B)$ in terms of two disjoint sets.

$$(A \cup B) = A \cup B \backslash A$$

where $A \cap B \backslash A = \varnothing$. Using axiom 3 we have

$$P(A \cup B) = P(A) + P(B \backslash A)$$

To calculate $P(B\backslash A)$, let us define B in terms of A, and the union of two disjoint sets.

$$B = (B \cap A) \cup (B\backslash A)$$

where $(B \cap A) \cap (B\backslash A) = \varnothing$ by definition. Using also axiom 3, we have.

$$P(B) = P(B \cap A) + P(B\backslash A)$$

$$P(B\backslash A) = P(B) - P(B \cap A)$$

Therefore, the probability of $(A \cup B)$ is the following

$$P(A \cup B) = P(A) + P(B\backslash A) = P(A) + P(B) - P(B \cap A)$$

g. (*) if $\{A_i\}_{i=1}^{\infty} \subseteq \Omega$ and $A_{i-1} \subseteq A_i \quad \forall i > 0$ then:

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \lim_{i \to \infty} P(A_i)$$

Hint: Use axiom 3.

Let us define the following

$$A := \bigcup_{i=1}^{\infty} A_i$$

We would like to write A in terms of disjoint sets so as to use axiom 3.

$$A_{i-1} \subseteq A_i \quad \forall i > 0 \implies A = \bigcup_{i=1}^{\infty} A_i\backslash A_{i-1}$$

The previous expression holds if we have $A_0 = \varnothing$. Notice this new expression can be regarded as starting with A_1 and adding the information form $A_2, A_3, \ldots$ which is not previously considered (e.g $A_2\backslash A_1, A_3\backslash A_2, \ldots$). Since this construction is a union of disjoint sets, we now can use axiom 3.

$$P(A) = P\left(\bigcup_{i=1}^{\infty} A_i\backslash A_{i-1}\right) = \sum_{i=1}^{\infty} P(A_i\backslash A_{i-1})$$

The infinite summation is in fact defined as a limit.

$$P(A) = \sum_{i=1}^{\infty} P(A_i\backslash A_{i-1}) = \lim_{n \to \infty} \sum_{i=1}^{n} P(A_i\backslash A_{i-1})$$

Notice the result in exercise (d), where we obtained $P(B) = P(B\backslash A) + P(A)$ for $A \subseteq B$. Therefore

$$P(A_i) = P(A_i\backslash A_{i-1}) + P(A_{i-1})$$

$$P(A_i\backslash A_{i-1}) = P(A_i) - P(A_{i-1})$$

$$P(A) = \lim_{n \to \infty} \sum_{i=1}^{n} P(A_i) - P(A_{i-1}) = \lim_{n \to \infty} \left(\sum_{i=1}^{n} P(A_i) - \sum_{i=1}^{n-1} P(A_i)\right) = \lim_{n \to \infty} P(A_n)$$

where we used $P(A_0) = P(\varnothing) = 0$. In conclusion,

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = P(A) = \lim_{i \to \infty} P(A_i)$$

5.2 Linear Algebra

Question 8 $\mathbf{x}^\top \mathbf{y} = 1 \times 0 + (-2) \times 4 + 5 \times (-3) + (-1) \times 7 = 0 + (-8) + (-15) + (-7) = -30.$

Question 9 $\mathbf{y} = (24, -14, -12)^\top$, $||\mathbf{x}||_2 = \sqrt{23}$, $||\mathbf{y}||_2 = \sqrt{916}$.
Note that by definition the ℓ_2 norm of a vector is $||\mathbf{x}||_2 = \sqrt{\mathbf{x}^\top \mathbf{x}}$.

Question 10 1, 2, 3.

A set of vectors $\{\mathbf{b}_1, ..., \mathbf{b}_K\}$ with $\mathbf{b}_k \in \mathbb{R}^d$ can form a basis of $\mathbb{R}^d$ iff $K \geq d$ and there exists a subset of d vectors within the set, such that they are orthogonal to each other.

Question 11 2, 5.

A point $\mathbf{x} \in \mathbb{R}^d$ is in *span*$(\{\mathbf{b}_1, ..., \mathbf{b}_K\})$ with $\mathbf{b}_k \in \mathbb{R}^d$ iff we can find $a_1, ..., a_K \in \mathbb{R}$ such that $\mathbf{x} = \sum_{k=1}^K a_k \mathbf{b}_k$.

Question 12 The rotation matrix is

$$\begin{pmatrix} \cos\frac{\pi}{4} & -\sin\frac{\pi}{4} \\ \sin\frac{\pi}{4} & \cos\frac{\pi}{4} \end{pmatrix}.$$

Question 13 a) The matrix A and vector $\mathbf{b}$ are

$$A = \begin{pmatrix} 1 & 2 & 0 \\ 3 & 2 & 4 \\ -2 & 1 & -2 \end{pmatrix}, \quad \mathbf{b} = (2, 5, 1)^\top.$$

b) The inverse of A is

$$A^{-1} = \begin{pmatrix} 2/3 & -1/3 & -2/3 \\ 1/6 & 1/6 & 1/3 \\ 7/12 & 5/12 & 1/3 \end{pmatrix}.$$

Therefore we have $\mathbf{x} = A^{-1}\mathbf{b} = (-1, 3/2, 43/12)^\top$.

c) $\text{rank}(A) = 3$: as A is invertible, it must have full rank.

References

Marc Peter Deisenroth, A. Aldo Faisal, and Cheng Soon Ong (2020). *Mathematics for machine learning.* Cambridge University Press.