

Machine Learning: Algorithms and Applications

Philip O. Ogunbona

Advanced Multimedia Research Lab
University of Wollongong

Math Review - Elementary Probability Theory
Autumn 2024

Outline of Topics

- 1 Context
- 2 Events and probabilities
- 3 Conditional probability
- 4 Independent and mutually exclusive events
- 5 Joint and marginal probability distribution
- 6 Random variables
- 7 Normal distribution
- 8 Summary of key points

Pattern recognition and probability

- Machine learning algorithms help us to analyse data and extract or recognize inherent patterns.
- Real life data displays uncertainty and variability.
- Probability theory allows us to model the uncertainty and variability in data.
- The models are powerful and provide the basis for statistical/probabilistic reasoning.

Introduction to Probability

- Probability is the study of **non-deterministic** events.
- We represent the possible outcomes of an experiment as a set S (also called *sample space*).
- For example if,

$$S = \{f_1, f_2, f_3, \dots, f_6\},$$

is the set of outcomes of rolling a die. The probability of each outcome is,

$$p(f_i) = 1/6, \quad i = 1, 2, \dots, 6$$

- **Events** are a subsets of sample space. For example the sets:

$$\{even\} = \{f_2, f_4, f_6\} \text{ and } \{odd\} = \{f_1, f_3, f_5\}$$

are events in the rolling of the die.

Introduction to Probability

- Two events, A and B , are **mutually exclusive** if $A \cap B = \emptyset$.
- Suppose we have a **finite sample space** S and a set of events, $\{A, B, \dots\}$. A **probability function**, P gives a real value for each event such that:
 - for every event A we have $0 \leq P(A) \leq 1$.
 - $P(S) = 1$.
 - If A and B are mutually exclusive, then $P(A \cup B) = P(A) + P(B)$.
- If each sample point in the space has the same probability, we refer to this as **equiprobable** or **uniform space**.
- We can use **the number of points in A** to find its probability.

Introduction to Probability

Example 1:

Suppose we have a loaded die with

$$p(f_i) = 1/4, \quad \text{for } i = 1, 2, 3$$

$$p(f_i) = 1/12, \quad \text{for } i = 4, 5, 6$$

Then,

$$p(\text{even}) = \frac{1}{4} + 2 \times \frac{1}{12} = \frac{5}{12}$$

It can be shown that for two events A and B

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

Introduction to Probability

Example 2:

Consider the events:

A= even; B= multiple of 3
such that,

$$P(A) = \frac{1}{2}, \quad P(B) = \frac{1}{3}$$

$$P(A \cap B) = P(\text{even and multiple of 3}) = \frac{1}{6}$$

and

$$P(A \cup B) = \frac{1}{3} + \frac{1}{2} - \frac{1}{6} = \frac{2}{3}$$

Introduction to Probability

Example 3:

Consider the toss of two dice.

Each point of the sample space can be represented by a two tuple,

$$(i, j), \quad i = 1, 2, \dots, 6, \quad j = 1, 2, \dots, 6$$

For an un-biased die every outcome has equal probability, $p(i, j) = \frac{1}{36}$ of occurrence.

For illustrative purposes, let us define some events. (See next slide)

Introduction to Probability

We define 12 different events v_1, \dots, v_{12} where v_u is the event that $i + j = u$. So, we have:

$$v_1 \Rightarrow p(v_1) = 0$$

$$v_2 (1, 1) \Rightarrow p(v_2) = 1/36$$

$$v_3 (1, 2), (2, 1) \Rightarrow p(v_3) = 1/18$$

$$v_4 (1, 3), (2, 2) (3, 1) \Rightarrow p(v_4) = 1/12$$

$$v_5 (1, 4), (2, 3) (3, 2) (4, 1) \Rightarrow p(v_5) = 1/9$$

$$v_6 (1, 5), (2, 4) (3, 3) (4, 2) (5, 1) \Rightarrow p(v_6) = 5/36$$

$$v_7 (1, 6), (2, 5) (3, 4) (4, 3) (5, 2) (6, 1) \Rightarrow p(v_7) = 1/6$$

$$v_8 (2, 6) (3, 5) (4, 4) (5, 3) (6, 2) \Rightarrow p(v_8) = 5/36$$

$$v_9 (3, 6) (4, 5) (5, 4) (6, 3) \Rightarrow p(v_9) = 1/9$$

$$v_{10} (4, 6) (5, 5) (6, 4) \Rightarrow p(v_{10}) = 1/12$$

$$v_{11} (5, 6) (6, 5) \Rightarrow p(v_{11}) = 1/18$$

$$v_{12} (6, 6) \Rightarrow p(v_{12}) = 1/36$$

Conditional Probability

- Events may be dependent; that is, occurrence of one influences the probability of the other.
- The probability of event A given that event E has occurred is given by

$$P(A|E) = \frac{P(A \cap E)}{P(E)}$$

Example 4:

Consider two events described as follows:

$$\begin{aligned} A &= \{ \text{points with } i + j = 2u \} \\ &= \{ v_2, v_4, v_6, v_8, v_{10}, v_{12} \} \\ E &= \{ \text{points with } i = j \} \\ &= \{ (1, 1), (2, 2), (3, 3), (4, 4), (5, 5), (6, 6) \} \end{aligned}$$

Conditional Probability

Example 4 cont.

Then,

- $P(A) = 18/36 = 1/2$
- $P(E) = 6/36 = 1/6$
- E is a subset of A:
 - $a \in E \Rightarrow a \in A.$
- We can then infer that $P(A \cap E) = P(E) = 6/36 = 1/6$
So $P(E|A) = P(A \cap E)/P(A) = (1/6)/(1/2) = 1/3$

Independent Events

- Two events E and A are **independent** if

$$P(A|E) = P(A)$$

That is, occurrence of E does not influence probability of occurrence of A .

- Equivalently we can write,

$$P(A \cap E) = P(A|E)P(E) = P(A)P(E)$$

Example 5:

A fair coin is tossed three times. Consider two events:

$A = \{\text{first toss is heads}\}$

$B = \{\text{second toss is heads}\}$

We have

$$P(A) = P(\{hhh, hht, hth, htt\}) = 1/2$$

$$P(B) = P(\{hhh, hht, thh, tht\}) = 1/2$$

$$P(A \cap B) = P(\{hht, hhh\}) = 1/4$$

So $P(A \cap B) = P(A) \times P(B)$ since the two events are independent.

Mutually Exclusive Events

- If events E and A are **mutually exclusive**, occurrence of one excludes the occurrence of the other and so,

$$P(A|E) = 0$$

- Note that this means that mutually exclusive events are *not* independent. Recall that independence requires that $P(A \cap E) = P(A|E)P(E) = P(A)P(E)$.
- If $A \subset E$ then

$$P(A \text{ and } E) = P(A)$$

Mutually Exclusive Events

- S consists of the outcomes of rolling a die.

$$P(f_i) = \frac{1}{6} \quad i = 1, 2, \dots, 6$$

- Let $E = \text{even}$ and $A = \{f_2\}$.

Then

$$P(f_2|E) = \frac{P(f_2 \text{ and } E)}{P(E)} = \frac{1/6}{1/2} = \frac{1}{3}$$

- Suppose we are interested in the event $A = \{ \text{even and divisible by 3} \}$.

Then

$$A \subset E$$

and

$$P(A \text{ and } E) = P(A) = 1/6.$$

Joint Probability Distribution

We may have **joint probability** distribution: In this case the sample space consists of pairs of outcomes of two experiments. We illustrate by way of example.

Example 6:

Joint probability of (temperature, humidity) in summer:

	Low	High
30°C	1/6	1/12
40°C	5/18	1/9
45°C	7/36	2/12

Marginal Distribution

- Given a joint distribution for (X, Y) , one can derive distribution of X and Y .
- Let $P(x, y)$ denote probability of the sample point (x, y) .

$$P(x) = \sum_{\text{all values of } y} P(x, y)$$

$$P(y) = \sum_{\text{all values of } x} P(x, y)$$

- $P(x)$ and $P(y)$ are *marginal distributions*.

Example 7:

In the temperature and humidity example (6) we have marginal distributions as follows:

$$P(\text{Temperature}) = \left(\frac{9}{36}, \frac{14}{36}, \frac{13}{36} \right)$$

$$P(\text{Humidity}) = \left(\frac{23}{36}, \frac{13}{36} \right)$$

Conditional independence

- Recall the joint probability of two events A and B ,
 $P(A, B) = P(A)P(B|A) = P(B)P(A|B)$
- If events A and B are independent we write:

$$P(A, B) = P(A)P(B)$$

- If we have three events A, B, C , the joint probability is written as:

$$P(A, B, C) = P(A)P(B|A)P(C|A, B)$$

- Of course if the three events are independent we write:

$$P(A, B, C) = P(A)P(B)P(C)$$

- Can we have conditional independence? What does this mean?
How useful is it?

Conditional independence (adapted from QMC London slides)

Example of independence

Suppose Jane and John each toss separate coins. Let A represent the variable “Jane’s toss outcome”, and B represent the variable “John’s toss outcome”. Both A and B have two possible values (Heads and Tails). Without any controversy we can assume that A and B are independent. Evidence about B will not change our belief in A .

Conditional independence (adapted from QMC London slides)

Example of conditional independence

Now suppose both Jane and John toss the same coin. Again let A represent the variable “Jane’s toss outcome”, and B represent the variable “John’s toss outcome”. Assume also that there is a possibility that the coin is biased towards heads but we do not know this for certain. In this case A and B are not independent. For example, observing that A is Heads causes us to increase our belief in B being Heads

Conditional independence (adapted from QMC London slides)

In the conditional independence example the variables A and B are both dependent on a separate variable C , “the coin is biased towards Heads” (which has the value True or False). Although A and B are not independent, it turns out that once we know for certain the value of C then any evidence about A cannot change our belief about B (and in general vice versa). Specifically:

$$P(B|C) = P(B|A, C)$$

$$P(A|C) = P(A|B, C)$$

$$P(A, B|C) = P(A|C)P(B|C)$$

Random Variable

- A **random variable** assigns **values** to events in a probability space.

Example 8:

X is a random variable that assigns 0 or 1 to the outcome of the an experiment such as in Example 3 (this is discrete case).

If the sum is divisible by 4, then we assign $X = 0$ otherwise it is 1.

We have

$$\begin{aligned} p(X = 0) &= \sum p(\text{sum divisible by } 4) \\ &= 1/12 + 5/36 + 1/36 = 9/36 = 1/4 \end{aligned}$$

$$p(X = 1) = 1 - 1/4 = 3/4$$

Example 8 contd.:

- Let a random variable X take values x_1, x_2, \dots, x_N .
- **Expected value** of X is

$$E(X) = \sum x_n p(x_n)$$

Expected value can be regarded as the average value of the random variable.

In example 8

$$E(X) = 0 \times 1/4 + 1 \times 3/4 = 3/4$$

Second moment and Variance

- The **second moment** of a random variable X that takes on values x_i is

$$E[X^2] = \sum_i x_i^2 P(x_i) \quad (1)$$

- If we denote the expected value of X as $\mu = E[X]$, the **variance** of X is,

$$\text{Var}\{X\} = \sigma^2 = E[(X - \mu)^2] = \sum_i (x_i - \mu)^2 P(x_i) \quad (2)$$

σ is the standard deviation of the X .

Functions of two random variables

- If we have a function of two random variables $f(x, y)$ that are jointly distributed as $P(x, y)$. The expected value of the function is

$$E[f(x, y)] = \sum_{R_X} \sum_{R_Y} f(x, y)P(x, y)$$

R_X and R_Y are the ranges of X and Y respectively.

- The expectation operator is a linear function since,

$$E[\alpha_1 f_1(x, y) + \alpha_2 f_2(x, y)] = \alpha_1 E[f_1(x, y)] + \alpha_2 E[f_2(x, y)]$$

for any two functions $f_1(x, y)$ and $f_2(x, y)$; and scalars α_1 and α_2 .

Functions of two random variables

- The expected value of X and Y can be written as,

$$\mu_x = E[X] = \sum_{x \in R_X} \sum_{y \in R_Y} xP(x, y)$$

$$\mu_y = E[Y] = \sum_{x \in R_X} \sum_{y \in R_Y} yP(x, y)$$

The variances are,

$$\sigma_x^2 = E[(X - \mu_x)^2] = \sum_{x \in R_X} \sum_{y \in R_Y} (x - \mu_x)^2 P(x, y)$$

and

$$\sigma_y^2 = E[(Y - \mu_y)^2] = \sum_{x \in R_X} \sum_{y \in R_Y} (y - \mu_y)^2 P(x, y)$$

Functions of two random variables

- The **covariance** of X and Y is defined as,

$$\sigma_{xy} = E[(X - \mu_x)(Y - \mu_y)] = \sum_{x \in R_X} \sum_{y \in R_Y} (x - \mu_x)(y - \mu_y)P(x, y)$$

- Covariance is a measure of the degree of statistical dependence between X and Y . If X and Y are statistically independent, $\sigma_{xy} = 0$.
- If $\sigma_{xy} = 0$, then the random variables are said to be **uncorrelated**.
- It does not follow that uncorrelated variables must be statistically independent.

Functions of two random variables

- Note that uncorrelated variables are statistically independent if they have multivariate normal distribution.
- If we have the relation, $Y = \alpha X$, which denotes that Y depends strongly on X , then $\sigma_{xy} = \alpha\sigma_x$.
- Covariance is **positive** if X and Y increase or decrease together and **negative** if Y decreases when X increases.
- The **correlation coefficient** defined as

$$\rho = \frac{\sigma_{xy}}{\sigma_x\sigma_y}$$

is normalized covariance and takes values between -1 and +1

Vector random variables

- If we have more than two variables it is convenient to represent them in a vector notation.
- \mathbf{x} denotes a random vector with components x_1, x_2, \dots, x_d (d-dimensional vector).
- The joint distribution can be obtained and will in general be complicated. If the variables are statistically independent, the joint probability mass function, $P(\mathbf{x})$ is the product,

$$\begin{aligned} P(\mathbf{x}) &= P_{x_1}(x_1)P_{x_2}(x_2) \cdots P_{x_d}(x_d) \\ &= \prod_{i=1}^d P_{x_i}(x_i) \end{aligned}$$

The marginal distributions, $P_{x_i}(x_i)$ are obtained by summing the joint distribution over the other variables.

Expectation, Mean Vectors and Covariance Matrices

- The expected value of a vector is defined as the vector whose components are the expected values of the original components.
- The d -dimensional mean vector μ is defined by,

$$\mu = E[\mathbf{x}] = \begin{bmatrix} E[x_1] \\ E[x_2] \\ \vdots \\ E[x_d] \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_d \end{bmatrix} = \sum_x \mathbf{x}P(\mathbf{x})$$

Expectation, Mean Vectors and Covariance Matrices

- The **covariance matrix** denoted by Σ , is defined as

$$\sigma_{ij} = \sigma_{ji} = E[(x_i - \mu_i)(x_j - \mu_j)] \quad i, j = 1, \dots, d$$

$$\begin{aligned} \Sigma &= \begin{bmatrix} E[(x_1 - \mu_1)(x_1 - \mu_1)] & \dots & E[(x_1 - \mu_1)(x_d - \mu_d)] \\ E[(x_2 - \mu_2)(x_1 - \mu_1)] & \dots & E[(x_2 - \mu_2)(x_d - \mu_d)] \\ \vdots & \ddots & \vdots \\ E[(x_d - \mu_d)(x_1 - \mu_1)] & \dots & E[(x_d - \mu_d)(x_d - \mu_d)] \end{bmatrix} \\ &= \begin{bmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1d} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{d1} & \sigma_{d2} & \dots & \sigma_{dd} \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1d} \\ \sigma_{21} & \sigma_2^2 & \dots & \sigma_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{d1} & \sigma_{d2} & \dots & \sigma_d^2 \end{bmatrix} \end{aligned}$$

Expectation, Mean Vectors and Covariance Matrices

- The covariance matrix, Σ , can be written as the vector outer product,

$$\Sigma = E[(\mathbf{x} - \mu)(\mathbf{x} - \mu)^t]$$

- The covariance matrix is symmetric and the diagonal elements are the variances of the individual elements of \mathbf{x} . They are always positive.
- The off-diagonal entries of Σ are the covariances and they are positive or negative.

Expectation, Mean Vectors and Covariance Matrices

- If the variables are statistically independent, the covariances are zero and the covariance matrix is diagonal,

$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_d^2 \end{bmatrix}$$

- For any d -dimensional vector \mathbf{w} , if the quadratic form

$$\mathbf{w}^t \Sigma \mathbf{w} > 0$$

Σ is positive definite. It is positive semi-definite if,

$$\mathbf{w}^t \Sigma \mathbf{w} \geq 0$$

Continuous Random Variables

- If the random variable can take values in the continuum we refer to it as a continuous random variable and describe it in terms of its probability density function.
- Rather than the probability mass function $P(X)$ we define the probability density function $p(x)$ with the property,

$$Pr[x \in (a, b)] = \int_a^b p(x) dx.$$

This is the probability that the continuous random variable, x , has value lying in the interval (a, b)

- The following is true:

$$p(x) \geq 0 \quad \text{and} \quad \int_{-\infty}^{\infty} p(x) = 1$$

Normal Distributions

- The normal or Gaussian probability density function is defined (for a single random variable) by,

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp^{-1/2((x-\mu)^2/\sigma^2)}$$

- The following expected values can be computed,

$$E[1] = \int_{-\infty}^{\infty} p(x) dx = 1$$

$$E[x] = \int_{-\infty}^{\infty} xp(x) dx = \mu$$

$$E[(x - \mu)^2] = \int_{-\infty}^{\infty} (x - \mu)^2 p(x) dx = \sigma^2$$

Bayes Rule

- We can combine the conditional probability with the product and sum rules to yield **Bayes** rule (also called **Bayes** theorem)
- Let us consider two random variable X and Y :

$$\begin{aligned} p(X = x|Y = y) &= \frac{p(X = x, Y = y)}{p(Y = y)} \\ &= \frac{p(X = x)p(Y = y|X = x)}{\sum_{x'} p(X = x')p(Y = y|X = x')} \end{aligned}$$

- This rule is used in classification task.

Example

Medical diagnosis

- Suppose a person went to have a medical test for some cancer. We are interested in the probability of the person having cancer if the test is positive. It is known that the **sensitivity** of the test is 80% (i.e if the person has cancer, the test will be positive with probability of 0.8). We can write this as $p(x = 1|y = 1) = 0.8$; x is the random variable “test is positive” and y is the random variable “has cancer”.
- Assuming that this cancer is very rare and has a low probability of occurrence: $p(y = 1) = 0.004$ and equivalently $p(y = 0) = 0.996$.
- Let us further assume that the probability of **false positive**, $p(x = 1|y = 0) = 0.1$
- The actual probability of the person having the cancer under the given conditions can be computed by using **Bayes** rule:

$$\begin{aligned} p(y = 1|x = 1) &= \frac{p(x = 1|y = 1)p(y = 1)}{p(x = 1|y = 1)p(y = 1) + p(x = 1|y = 0)p(y = 0)} \\ &= \frac{0.8 \times 0.004}{0.8 \times 0.004 + 0.1 \times 0.996} = 0.031 \end{aligned}$$

- This means, testing positive implies there is only 3% chance of having the cancer.

Summary of key points and relations - I

- 1 Sample space, (S) - the collection of possible outcomes for an experiment
- 2 Event - A subset of the sample space
- 3 Mutually exclusive events - $A \cap B = \emptyset$
- 4 Probability of event - real number associated with the event that represents relative frequency of occurrence of that event
- 5 $P(S) = 1$
- 6 $P(A) \geq 0$
- 7 $P(A \cup B) = P(A) + P(B)$ if $A \cap B = \emptyset$
- 8 $P(B_1 \cup B_2 \cup \dots \cup B_n \dots) = \sum_{i=1}^{\infty} P(B_i)$ if $B_i \cap B_j = \emptyset$ for all $i \neq j$
 - $P(\emptyset) = 0$
 - $P(\bar{A}) = 1 - P(A)$
 - $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- 9 Equally likely rule : $P(A) = \frac{\text{number of elements in } A}{\text{number of elements in } S}$
- 10 Conditional probability: $P(A|B) = \frac{P(A \cap B)}{P(B)}$
- 11 $P(A \cap B \cap C) = P(A)P(B|A)P(C|A \cap B)$

Summary of key points and relations - II

- 12 Partition: $E_1 \cup E_2 \cup \dots \cup E_n = S$ and $E_i \cap E_j = \emptyset$ if $i \neq j$
- 13 Bayes Theorem: $P(E_i|A) = \frac{P(E_i)P(A|E_i)}{\sum_{j=1}^n P(E_j)P(A|E_j)}$ if E_1, E_2, \dots, E_n is a partition of S .
- 14 Conditional independence: $P(A|C) = P(A|B, C)$;
 $P(A, B|C) = P(A|C) \times P(B|C)$; A and B are conditionally independent given C .
A typical situation where conditional independence holds between events, is when the events share a common cause.
- 15 Chain rule with conditional independence: Consider the usual chain rule

$$P(E_i, E_2, \dots, E_n|A = a) = P(E_1|A = a) \times P(E_2, E_1|A = a) \\ \times P(E_n, E_{n-1}, \dots, E_1|A = a)$$

Under the assumption of **conditional independence** (i.e. the E_i have a common cause in the event $A = a$;

$$P(E_i, E_2, \dots, E_n|A = a) = P(E_1|A = a) \times P(E_2|A = a) \times \dots \times P(E_n|A = a)$$