Linear Algebra for Statistical Learning and Data Mining

(HTIN5005)

Contents

1 Vectors

1.1 What is a vector?

A **vector** is an ordered finite list of numbers. We typically write vectors as vertical arrays, surrounded by square or curved brackets, as in

$$\begin{bmatrix} -1 \\ 0 \\ 2.5 \\ -7.2 \end{bmatrix} \text{ or } \begin{pmatrix} -1 \\ 0 \\ 2.5 \\ -7.2 \end{pmatrix}$$

$$\mathbf{a} = \begin{bmatrix} 5 \\ -2 \\ -3 \end{bmatrix}, \quad \mathbf{a} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

We also write vectors as numbers separated by commas.

$$\mathbf{a} = (5, -2, -3), \quad \mathbf{a} = (1, 1).$$

The **elements** or **entries** of a vector are the values in the array.

The **size** (or **dimension**) of a vector is the number of elements it contains.

A vector of size n is called an $n-$vector.

$$\mathbf{a} = \begin{bmatrix} a_1 \\ \vdots \\ a_i \\ \vdots \\ a_n \end{bmatrix}$$

$$\mathbf{a} = (a_1, \cdots, a_i, \cdots, a_n)$$

A vector with n entries, each belonging to $\mathbb{R}$ (the set of real numbers), is called a n-vector over $\mathbb{R}$. We denote the set of n-vectors over $\mathbb{R}$ as $\mathbb{R}^n$.

$$\boldsymbol{a} = (a_1, a_2, a_3) \in (\mathbb{R}, \mathbb{R}, \mathbb{R}) \equiv \mathbb{R}^3$$

We can also define a vector as a function from a finite set D to $\mathbb{R}$. For example, a function from $D = \{0, 1, 2, \ldots, d-1\}$ to $\mathbb{R}$. The vector

$$\mathbf{a} = (6, -4, -3.7)$$

is the function

$$\begin{aligned} 0 &\longmapsto 6 \\ 1 &\longmapsto -4 \\ 2 &\longmapsto -3.7, \end{aligned}$$

where $\longmapsto$ reads "maps to".

This last definition is useful as it matches how we work with vectors in Python. For example, if we store the above vector as a Python list called `a`, the command `a[0]` returns 6, the first element of the vector. More formally, we say that the above definition lends itself to representation in a data structure (a format for organising and storing data). Python objects such as lists, dictionaries, and `NumPy` arrays are data structures that can represent vectors.

1.2 Special vectors

Zero vector. A zero vector has all elements equal to zero $\mathbf{0} = (0, 0, ..., 0)$. $\mathbf{0}_n$ indicates a zero vector with dimension n.

Unit vector. A unit vector has all elements equal zeros, except one element which is equal to one. $\mathbf{e}_i = [0, ..., 0, 1, 0, ..., 0]^T$ (all zeros except for 1 at i-th position).

Ones vector. A ones vector has elements equal to one, $\mathbf{1} = [1, 1, ..., 1]^T$. $\mathbf{1}_n$ indicates a ones vector with dimension n. We also use the notation $\boldsymbol{\iota}$ for this type of vector

Sparse vector. A vector is said to be sparse if many of its elements are equal to zero.

1.3 Vector operations

Vector equality. $\mathbf{a} = \mathbf{b} \Longleftrightarrow a_i = b_i$ for all $i = 1, 2, ..., n$.

Scalar-vector multiplication. Let α denote a scalar. The vector $\alpha\,\mathbf{a}$ is the vector with elements $\{\alpha\, a_i\}$. For example, let $\mathbf{a} = (5, -2, -3)$, then

$$0.5\,\mathbf{a} = (0.5 \times 5, 0.5 \times -2, 0.5 \times -3) = (2.5, -1, -1.5)$$

Addition. let $\mathbf{a}$ and $\mathbf{b}$ be two vectors with the same size n. The sum $\mathbf{c} = \mathbf{a} + \mathbf{b}$ is the vector with elements $c_i = a_i + b_i$.

Let $\mathbf{a} = (5, -2, -3)$ and $\mathbf{b} = (-1, 2, 4)$. Then,

$$\mathbf{c} = \mathbf{a} + \mathbf{b} = (5, -2, -3) \; + \; (-1, 2, 4) = (5 - 1, -2 + 2, -3 + 4) = (4, 0, 1).$$

Linear combination. Let $\mathbf{a}$ and $\mathbf{b}$ are $n-$vectors and β_1 and β_2 are scalars, the $n-$vector

$$\beta_1\mathbf{a} + \beta_2\mathbf{b}$$

is called a linear combination of $\boldsymbol{a}$ and $\boldsymbol{b}$. The scalars β_1 and β_2 are the **coefficients** of the linear combination.

Let $\mathbf{a} = (5, -2, -3)$, $\mathbf{b} = (-1, 2, 4)$, $\beta_1 = 2$, and $\beta_2 = 3$.

$$2\mathbf{a} + 3\mathbf{b} = (2 \times 5, 2 \times -2, 2 \times -3) + (3 \times -1, 3 \times 2, 3 \times 4) = (7, 2, 6)$$

1.4 Inner Product

We define the dot or **inner product** of two n-dimensional vectors $\mathbf{a}$ and $\mathbf{b}$ as

$$\boldsymbol{a}^T\boldsymbol{b} = a_1b_1 + a_2b_2 + \ldots + a_nb_n = \sum_{i=1}^{n} a_ib_i$$

Example: $\mathbf{a} = (2, -1, 3)$ and $\mathbf{b} = (5, -2, -3)$, then

$$\boldsymbol{a}^T\boldsymbol{b} = 2 \times 5 + (-1) \times (-2) + 3 \times (-3) = 3$$

Some authors use the notation $\langle \mathbf{a}, \mathbf{b} \rangle$ for inner products.

Properties. The following are useful properties of inner products that follow easily from the definition.

$$(\alpha\mathbf{a})^T\mathbf{b} = \alpha(\mathbf{a}^T\mathbf{b})$$

$$\mathbf{a}^T\mathbf{b} = \mathbf{b}^T\mathbf{a}$$

$$\mathbf{a}^T(\mathbf{b} + \mathbf{c}) = \mathbf{a}^T\mathbf{b} + \mathbf{a}^T\mathbf{c}$$

.

Examples.

Sum. $\boldsymbol{\iota}^T\boldsymbol{a} = a_1 + a_2 + \ldots + a_n$ is the sum the elements of $\boldsymbol{a}$.

Average. $(1/n)(\boldsymbol{\iota}^T\boldsymbol{a})$ is the average of the elements of $\boldsymbol{a}$.

Sum of squares. $\boldsymbol{a}^T\boldsymbol{a} = a_1^2 + \ldots + a_n^2$ is the sum of squares of the elements of $\boldsymbol{a}$.

$$n = 4, \quad \mathbf{x} = \begin{bmatrix} 3 \\ 4 \\ 2 \\ 7 \end{bmatrix}, \quad \mathbf{1} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix},$$

$$\Rightarrow \frac{1}{4}\boldsymbol{\iota}^T\mathbf{x} = \frac{1}{4}(1 \times 3 + 1 \times 4 + 1 \times 3 + 1 \times 7) = 4.$$

1.5 Linear functions

The notation $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ means that f is a function that maps an $n-$vector to a real number. If $\boldsymbol{x}$ is an $n-$vector, then $f(\boldsymbol{x})$ (a scalar) is the value of the function at $\boldsymbol{x}$. In this setting, we refer to $\boldsymbol{x}$ as the **argument** of the function.

Let $\boldsymbol{x}$ and $\boldsymbol{y}$ be $n-$vectors and α and β be scalars. A **linear function** is a function that satisfies the property

$$f(\alpha\boldsymbol{x} + \beta\boldsymbol{y}) = \alpha f(\boldsymbol{x}) + \beta f(\boldsymbol{y})$$

We can always represent a linear function as an inner product. Let $\boldsymbol{a}$ be an $n-$vector. The we can write any linear function using the form

$$f(\boldsymbol{x}) = \boldsymbol{a}^T\boldsymbol{x} = a_1x_1 + a_2x_2 + \ldots a_nx_n,$$

where $\boldsymbol{x}$ is an $n-$vector. Here, $\boldsymbol{a}$ is fixed, and the argument $\boldsymbol{x}$ can be any $n-$vector.

For example, in a linear regression model $f(\boldsymbol{x})$ is the regression function, $\boldsymbol{x}$ are the predictor values, and $\boldsymbol{a}$ are the model parameters.

An **affine function** is a linear function plus a constant, that is

$$f(\boldsymbol{x}) = \boldsymbol{a}^T\boldsymbol{x} + b,$$

for a scalar b.

1.6 Norm

The **Euclidean norm** or ℓ_2-norm of a vector is

$$\|\mathbf{a}\|_2 = \sqrt{\mathbf{a}^T\mathbf{a}} = \left(\sum_{i=1}^{n} a_i^2\right)^{1/2}.$$

This is the distance from the origin to the point $\mathbf{a}$ or the **length** of the vector.

The **normalized vector** $\frac{\mathbf{a}}{\|\mathbf{a}\|_2}$ has unit norm.

Example.

Let $\mathbf{x}$ be a vector with sample average zero. Then $\|\mathbf{x}\|^2$ is the sum of squares of $\boldsymbol{x}$ and $s_x = \|\mathbf{x}\|^2/n$ is the sample variance.

General definition. A norm $\|\cdot\|$ is a function that satisfies the following properties:

1. $\|\boldsymbol{a}\| \geq 0$ (non-negativity).
2. $\|\boldsymbol{a}\| = 0$ only if $\boldsymbol{x} = \mathbf{0}$ (definiteness).
3. $\|\alpha\boldsymbol{a}\| = |\alpha| \times \|\boldsymbol{a}\|$ (homogeneity).
4. $\|\boldsymbol{a} + \boldsymbol{b}\| \leq \|\boldsymbol{a}\| + \|\boldsymbol{b}\|$ (triangle inequality).

The ℓ_1 norm of a vector is

$$\|\mathbf{a}\|_1 = |a_1| + |a_2| + \ldots + |a_n| = \sum_{i=1}^{n} |a_i|.$$

The **Chebyshev** or ℓ_∞ norm is given by

$$\|\mathbf{a}\|_\infty = \max\{|a_1|, |a_2|, \ldots, |a_n|\}.$$

The **Minkowski norm** of order p is

$$\|\mathbf{a}\|_p = \left(\sum_{i=1}^{n} |a_i|^p\right)^{1/p},$$

for $p \geq 1$. This is a generalisation of the previous norms. We include this here because the `scikit-learn` package in Python sometimes refers to the Minkowski norm by default, even if $p = 2$.

1.7 Distance

The **Euclidean distance** between two vectors $\mathbf{x}$ and $\mathbf{y}$ is the norm of the difference vector $\mathbf{x} - \mathbf{y}$:

$$\mathbf{dist}(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|_2 = \sqrt{(\boldsymbol{x} - \boldsymbol{y})^T(\boldsymbol{x} - \boldsymbol{y})} = \left(\sum_{i=1}^{n}(x_i - y_i)^2\right)^{1/2}$$

Every norm $\|\cdot\|$ induces a distance metric $\|\boldsymbol{x} - \boldsymbol{y}\|$.

1.8 Orthogonal vectors

Two vectors are **orthogonal**, written $\mathbf{a} \perp \mathbf{b}$, if and only if their inner product is zero, $\mathbf{a}^T\mathbf{b} = 0$.

Example:

$$\mathbf{a} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}, \quad \|\mathbf{a}\| = \|\mathbf{b}\| = \sqrt{2}, \ \mathbf{a}^T\mathbf{b} = 0.$$

$$\mathbf{c} = \begin{bmatrix} 1 \\ -3 \end{bmatrix}, \quad \mathbf{d} = \begin{bmatrix} .6 \\ .2 \end{bmatrix},$$

$$\|\mathbf{c}\| = \sqrt{10} \approx 3.16, \|\mathbf{d}\| = \sqrt{.4} \approx 0.63, \quad \mathbf{c}^T\mathbf{d} = 0, \text{ and } \mathbf{a}^T\mathbf{c} = -2.$$

2 Matrices

2.1 What is a matrix?

A matrix is a rectangular two-dimensional array of numbers such as

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & -2.3 & 0.1 \\ 1.3 & 4 & -0.1 & 7 \\ 4.1 & -1 & 0 & 1.7 \end{bmatrix}$$

The **size** (or **dimensions**) of a matrix are the number of rows and columns. The matrix above has 3 rows and 4 columns, so the size is 3×4 (it reads 3-by-4).

We represent an $(m \times n)$ matrix as

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1j} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2j} & \cdots & a_{2n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{i1} & a_{i2} & \cdots & a_{ij} & \cdots & a_{in} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mj} & \cdots & a_{mn}, \end{bmatrix}$$

with $\mathbf{A} \in \mathbb{R}^{m \times n}$.

We also represent a matrix as $\mathbf{A} = \{a_{ij}\}$. In a design matrix in regression analysis, the index $i = 1, 2, ..., m$ refers to the statistical units, and the index $j = 1, 2, ..., n$ to the variables or attributes.

2.2 Row and column vectors

A column vector is a $m \times 1$ matrix. We do not distinguish between vectors and column vectors.

$$\mathbf{a} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{bmatrix}$$

In the same way, a row vector is a $1 \times m$ matrix.

$$\mathbf{a} = \begin{bmatrix} a_1 & a_2 & \dots & a_m \end{bmatrix}$$

The **transpose** of a column vector is the corresponding row vector and vice-versa.

$$\begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{bmatrix}^T = \begin{bmatrix} a_1 & a_2 & \dots & a_m \end{bmatrix}$$

$$\begin{bmatrix} a_1 & a_2 & \dots & a_m \end{bmatrix}^T = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{bmatrix}$$

We can represent a matrix $\mathbf{X}$ as a partitioned matrix whose generic block is the $1 \times n$ row vector $\mathbf{x}_i^T = [x_{i1}, x_{i2}, ..., x_{ij}, ..., x_{in}]$, which contains the profile of the i-th row unit,

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_i^T \\ \vdots \\ \mathbf{x}_m^T \end{bmatrix}$$

Alternatively, we can partition as

$$\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_j, ..., \mathbf{x}_n],$$

where $\mathbf{x}_j$ is the $m \times 1$ column vector referring to the j-th variable or attribute.

2.3 Transpose

The **transpose** of an $m \times n$ matrix $\boldsymbol{A}$ yields an $n \times m$ matrix that interchanges the rows and columns of $\boldsymbol{A}$.

$$\mathbf{A}^T = \begin{bmatrix} a_{11} & a_{21} & \cdots & a_{i1} & \cdots & a_{m1} \\ a_{12} & a_{22} & \cdots & a_{i2} & \cdots & a_{m2} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{1j} & a_{2j} & \cdots & a_{ij} & \cdots & a_{mj} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{1n} & a_{2n} & \cdots & a_{in} & \cdots & a_{mn} \end{bmatrix}$$

The transpose has the property that $(\mathbf{A}^T)^T = \mathbf{A}$

Example.

$$\mathbf{A} = \begin{bmatrix} 2 & 3 & 4 \\ 1 & 2 & -1 \\ 3 & -2 & 5 \\ -2 & 4 & 1 \end{bmatrix}, \qquad \mathbf{A}^T = \begin{bmatrix} 2 & 1 & 3 & -2 \\ 3 & 2 & -2 & 4 \\ 4 & -1 & 5 & 1 \end{bmatrix}$$

2.4 Addition and scalar multiplication

Scalar Multiplication. $\alpha\mathbf{a} = \{\alpha a_{ij}\}$.

Matrix addition. If $\boldsymbol{A}$ is an $m \times n$ matrix and $\boldsymbol{B}$ an $m \times n$ matrix, then

$$\boldsymbol{A} + \boldsymbol{B} = \{a_{ij} + b_{ij}\}.$$

This can only be performed if $\boldsymbol{A}$ and $\boldsymbol{B}$ have the exact same dimensions.

Note that $(\mathbf{A}+\mathbf{B})^T = \mathbf{A}^T + \mathbf{B}^T$.

Example.

$$\begin{bmatrix} 2 & 3 \\ 3 & -2 \end{bmatrix} + \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} = \begin{bmatrix} 2+1 & 3+2 \\ 3+3 & -2+4 \end{bmatrix} = \begin{bmatrix} 3 & 5 \\ 6 & 2 \end{bmatrix}$$

2.5 Matrix-vector multiplication

The product of an $m \times n$ matrix $\boldsymbol{A}$ with an n-vector $\boldsymbol{b}$ is an m-vector $\boldsymbol{c}$ with element i equal to the inner product of the row i of $\boldsymbol{A}$ with $\boldsymbol{b}$.

$$c_i = \boldsymbol{a}_i^T \boldsymbol{b} = \sum_{j=1}^{n} a_{ij} b_j,$$

where $\boldsymbol{a}_i^T$ denotes i-th row of $\boldsymbol{A}$.

Example.

$$\begin{bmatrix} 1 & 4 \\ 7 & -3 \\ 2 & -5 \end{bmatrix} \times \begin{bmatrix} 2 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \times 2 + 4 \times 1 \\ 7 \times 2 - 3 \times 1 \\ 2 \times 2 - 5 \times 1 \end{bmatrix} = \begin{bmatrix} 6 \\ 11 \\ -1 \end{bmatrix}$$

2.6 Matrix-matrix multiplication

The product of an $m \times p$ matrix $\boldsymbol{A}$ with an $p \times n$ matrix $\boldsymbol{B}$ is an $m \times n$ matrix $\boldsymbol{C}$ with element ij equal to the inner product of the row i of $\boldsymbol{A}$ with column j of $\boldsymbol{B}$

$$c_{ij} = \boldsymbol{a}_i^T \boldsymbol{b}_j = \sum_{k=1}^{p} a_{ik} b_{kj}$$

where $\boldsymbol{a}_i^T$ denotes i-th row of $\boldsymbol{A}$ and $\boldsymbol{b}_j$ denotes j-th column of $\boldsymbol{B}$.

The matrix partitions that we use in the multiplication are

$$\mathbf{A} = \begin{bmatrix} \mathbf{a}_1^T \\ \vdots \\ \mathbf{a}_i^T \\ \vdots \\ \mathbf{a}_m^T \end{bmatrix}, \quad \boldsymbol{B} = [\mathbf{b}_1, ..., \mathbf{b}_j, ..., \mathbf{b}_n]$$

The multiplication ($\boldsymbol{AB}$) is only defined when the column dimension of $\boldsymbol{A}$ ($m \times p$ matrix) equals the row dimension$\boldsymbol{B}$ ($p \times n$ matrix).

Example.

$$\begin{bmatrix} 2 & 3 \\ 3 & -2 \end{bmatrix} \times \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} = \begin{bmatrix} 2 \times 1 + 3 \times 3 & 2 \times 2 + 3 \times 4 \\ 3 \times 1 - 2 \times 3 & 3 \times 2 - 2 \times 4 \end{bmatrix} = \begin{bmatrix} 11 & 16 \\ -3 & -2 \end{bmatrix}$$

Properties of matrix multiplication.

$$(\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T$$

$$(\mathbf{AB})\mathbf{C} = \mathbf{A}(\mathbf{BC})$$

$$\mathbf{A}(\mathbf{B} + \mathbf{C}) = \mathbf{AB} + \mathbf{AC}$$

$$(\mathbf{A} + \mathbf{B})\mathbf{C} = \mathbf{AC} + \mathbf{BC}$$

Unlike in scalar multiplication, the order of multiplication matters for matrices: in general, $\mathbf{AB} \neq \mathbf{BA}$. Moreover, remember that if $m \neq n$, $\mathbf{BA}$ is not even defined.

Example.

$$\mathbf{A} = \begin{bmatrix} 1 & 0 \\ 5 & -1 \\ 3 & 2 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 2 & -1 \\ 3 & 6 \end{bmatrix}, \quad \boldsymbol{\iota}_3 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

$$\mathbf{C} = \mathbf{AB} = \begin{bmatrix} 2 & -1 \\ 7 & -11 \\ 12 & 9 \end{bmatrix}, \quad \boldsymbol{\iota}^T\mathbf{C} = \begin{bmatrix} 21 & -3 \end{bmatrix}.$$

BA is not defined.

Vector outer product. If $\mathbf{a}$ is an m-vector and $\mathbf{b}$ is an n- vector, the outer product $\mathbf{ab}^T$ is the $m \times n$ matrix

$$\mathbf{ab}^T = \begin{bmatrix} a_1b_1 & a_1b_2 & \dots & a_1b_n \\ a_2b_1 & a_2b_2 & \dots & a_2b_n \\ \vdots & \vdots & & \vdots \\ a_mb_1 & a_mb_2 & \dots & a_mb_n \end{bmatrix}$$

2.7 Square matrices

A **square matrix** has the same number of rows and columns $m = n$.

Symmetric matrix. A square matrix $\mathbf{A}$ is symmetric if $\mathbf{A}^T = \mathbf{A}$.

Quadratic form. Let $\mathbf{A}$ be an n dimensional square matrix and $\mathbf{x}$ an $n \times 1$ vector. The scalar $\mathbf{x}^T\mathbf{Ax}$ is called a quadratic form.

A symmetric matrix $\mathbf{A}$ with the property that $\mathbf{x}^T\mathbf{Ax} > 0$ for any vector $\mathbf{x}$ is said to be **positive definite**.

2.8 Identity and diagonal matrices

The diagonal elements of a matrix are the elements a_{ij} such that $i = j$ (same row and column index).

An **identity matrix** of order n is a matrix with all diagonal elements equal to one ($a_{ii} = 1$ for $i = 1, \ldots, n$), and all non-diagonal elements equal to zero, that is

$$\mathbf{I}_n = \begin{bmatrix} 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \ddots & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & \ddots & 1 & 0 \\ 0 & 0 & \cdots & 0 & 1 \end{bmatrix} = \text{diag}(1, ..., 1)$$

For example,

$$\mathbf{I}_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \qquad \mathbf{I}_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

Properties.

Let $\boldsymbol{A}$ be an $m \times n$ matrix.

$$\boldsymbol{I}_n^2 = \boldsymbol{I}_n$$

$$\boldsymbol{I}_m \boldsymbol{A} = \boldsymbol{A}$$

$$\boldsymbol{A} \boldsymbol{I}_n = \boldsymbol{A}$$

Diagonal matrix. A diagonal matrix is a square matrix with zeros in all the non-

diagonal positions.

$$\mathbf{D} = \begin{bmatrix} d_1 & 0 & \cdots & 0 & 0 \\ 0 & d_2 & \ddots & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & \ddots & d_{n-1} & 0 \\ 0 & 0 & \cdots & 0 & d_n \end{bmatrix} = \text{diag}(d_1, ..., d_n)$$

Let $\boldsymbol{D}$ be an $n \times n$ diagonal matrix and $\boldsymbol{A}$ an $n \times p$ matrix. The operation $\boldsymbol{DA}$ multiplies each row i of $\boldsymbol{A}$ by the diagonal element d_i of $\boldsymbol{D}$.

2.9 Systems of Linear Equations

Consider a system of m linear equations with n variables.

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n &= b_2 \\ &\vdots \\ a_{m1}x_1 + a_{m2}x_2 + \ldots + a_{mn}x_n &= b_2 \end{aligned}$$

This system has a compact representation in matrix notation

$$\boldsymbol{Ax} = \boldsymbol{b},$$

where

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{21} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \ldots & a_{mn} \end{bmatrix}, \ \boldsymbol{x} = \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_n \end{bmatrix}, \ \boldsymbol{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}.$$

The study of linear systems is a fundamental part of linear algebra, which allows us to

determine whether a system has an unique solution, infinitely many solutions, or no solutions, and to obtain a solution if one exists.

Example.

We can write the system

$$\begin{cases} 2x_1 + 2x_2 + x_3 &= 9 \\ 2x_1 - x_2 + 2x_3 &= 6 \\ x_1 - x_2 + 2x_3 &= 5 \end{cases}$$

as

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b},$$

where

$$\boldsymbol{A} = \begin{bmatrix} 2 & 2 & 1 \\ 2 & -1 & 2 \\ 1 & -1 & 2 \end{bmatrix}, \ \boldsymbol{b} = \begin{bmatrix} 9 \\ 6 \\ 5 \end{bmatrix}.$$

The unique solution is $\boldsymbol{x} = (1, 2, 3)$.

2.10 Matrix inverse

An $n \times n$ matrix $\boldsymbol{A}$ is **invertible** is there exists a matrix $\boldsymbol{B}$ such that

$$\boldsymbol{A}\boldsymbol{B} = \boldsymbol{I}_n$$

If that is the case then we call $\boldsymbol{B}$ the **inverse** of $\boldsymbol{A}$, and use the notation $\boldsymbol{A}^{-1}$.

There are several methods for calculating a matrix inverse, but we will leave the details in the background. It is often the case in practice that the we do not actually need to explictly compute the matrix inverse to evaluate expressions in which it appears (for example in the formula for the OLS).

Properties.

$$(\boldsymbol{A}^{-1})^{-1} = \boldsymbol{A}$$

$$(\alpha\boldsymbol{A})^{-1} = (1/\alpha)\boldsymbol{A}^{-1}$$

$$(\boldsymbol{A}^{T})^{-1} = (\boldsymbol{A}^{-1})^{T}$$

$$(\boldsymbol{AB})^{-1} = \boldsymbol{B}^{-1}\boldsymbol{A}^{-1}$$

3 Example: Linear Regression and Least squares

3.1 Linear regression model

In this unit, we use the linear regression model as an algorithm to predict a continuous response variable Y as a function of predictor variables $X_1, X_2, \ldots, X_p$. The model is based on the linear predictive function

$$f(\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p,$$

where $\boldsymbol{x} = (x_1, \ldots, x_p)$ is vector of observed values of the predictors.

The parameter vector is

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}.$$

3.2 Least squares

We write a dataset as $=\{(y_i, x_{i1}, \ldots, x_{ip})\}_{i=1}^{n}$, where y_i denotes the response value for observation i and x_{ij} denotes the value of predictor j for observation i.

In the **ordinary least squares** (OLS) method, we obtain the coefficients by solving

the minimisation problem

$$\underset{\boldsymbol{\beta}}{\text{minimise}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\}$$

Define

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix},$$

where $\boldsymbol{X}$ is called the **design matrix**. Furthermore, let $J(\boldsymbol{\beta})$ denote the objective function. We can write it more compactly as

$$\begin{aligned} J(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \\ &= (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \\ &= ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2, \end{aligned}$$

which is the type of notation that you may find in certain material.

We can show that least squares estimate $\widehat{\boldsymbol{\beta}}$ satisfies the system of linear equations

$$\boldsymbol{X}^T \boldsymbol{X} \widehat{\boldsymbol{\beta}} = \boldsymbol{X}^T \boldsymbol{y}$$

, which has the form of Section 2.9 since $\boldsymbol{X}^T\boldsymbol{X}$ is a $(p+1) \times (p+1)$ matrix and $\boldsymbol{X}^T\boldsymbol{y}$ is a $(p+1)$-vector.

If $(\boldsymbol{X}^T\boldsymbol{X})$ is invertible, left multiplication with $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ gives the unique solution

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y},$$

the widely known formula for the OLS estimator.

References

Boyd, S. and L. Vandenberghe (2016). Vectors, matrices, and least squares. *Available: stanford.edu/class/ee103/mma.pdf*.

Klein, P. N. (2013). *Coding the matrix: Linear algebra through applications to computer science*. Newtonian Press.

Mathematics for Statistical Learning and Data Mining

(HTIN5005)

Contents

1 Exponentials and Logarithms

1.1 Exponential function

The mathematical constant $e = 2.71828\ldots$ is useful in several areas of application. We define it as

$$e = \lim_{n\to\infty} \left(1 + \frac{1}{n}\right)^n.$$

That is, if we compute the sequence $(1 + 1/2)^2 = 9/4 = 2.25$, $(1 + 1/3)^3 = 64/27 = 2.3704$, $(1 + 1/4)^4 = 625/256 = 2.4414$, ..., $(1 + 1/10)^{10} = 2.5937$, ..., $(1 + 1/20)^{20} = 2.6533$, and so on, the value approaches the constant e.

We can also define e as a sum

$$e = 1 + 1 + \frac{1}{2} + \frac{1}{2 \times 3} + \frac{1}{2 \times 3 \times 4} + \ldots$$

This produces much quicker convergence; if we carry on and include all the terms up to $1/(2 \times 3 \times 4 \times 5 \times 6)$ we get 2.7181.

The (natural) **exponential function** is e^x. We often write it as $\exp(x)$. In this case we say that e is the **base** for exponentiation.

1.2 Natural logarithm

The natural logarithm is the inverse of the exponential function. That is, the natural logarithm of a number x is the power to which e must be raised to equal x. For example, $\log(2)$ is the number x such that $e^x = 2$. Since the result is always positive no matter what power e is raised to, it is easy to see that $\log(x)$ is only defined for positive values of x.

We say the the natural logarithm is the logarithm to base e. Unless explicitly stated otherwise, we refer to the natural logarithm when writing simply "logarithm".

1.3 Properties of exponentials and logarithms

Let x and y be scalars.

$$\exp(0) = 1$$

$$\exp(-x) = 1/\exp(x)$$

$$\exp(x + y) = \exp(x)\exp(y)$$

$$\exp(x - y) = \exp(x)/\exp(y)$$

$$\exp(\log(x)) = x$$

$$\log(1) = 0$$

$$\log(xy) = \log(x) + \log(y)$$

$$\log(x/y) = \log(x) - \log(y)$$

$$\log(x^y) = y\log(x)$$

$$\log(1/x) = -\log(x)$$

1.4 Approximation to percentage changes

Suppose that you make a deposit in a financial instrument that pays your initial investment plus 5% interest after one year. Let $r = 0.05$ be the interest rate. The payout after one year is then

$$P = 100 \times (1 + r) = 100 \times 1.05 = 105.$$

Now, suppose instead that the investment pays 2.5% compound interest every six months. Then,

$$P = 100 \times (1 + r/2)^2 = 100 \times 1.0506 = 105.06,$$

or 1.25% every three months,

$$P = 100 \times (1 + r/4)^4 = 100 \times 1.0510 = 105.10$$

Taking the limit,

$$P = 100 \times \exp(r) = 105.13,$$

where

$$\exp(r) = \lim_{n \to \infty} \left(1 + \frac{r}{n}\right)^n.$$

This illustrates that

$$\exp(r) \approx 1 + r$$

for r not large. The smaller r, the more accurate the approximation. For example, try computing exp(.01), exp(.1), and exp(.2).

2 Functions of one variable

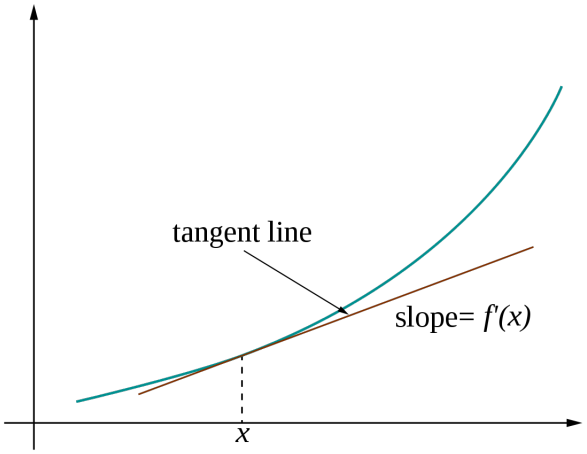
A **function** $f : \mathcal{X} \to \mathcal{Y}$ is a relation that associates each element x of a set $\mathcal{X}$, called the **domain** of the function, to a single element y of another set $\mathcal{Y}$, the **codomain** of the function. For example, if the function is called f, we may write $y = f(x)$ (read f of x). The element x is the **argument** or **input**, and y is the value or **output**.

2.1 Derivatives

The **derivative** f' of a function $f(x)$ is the slope of the graph of f at the point x. We can think of it as a function that measures the how fast the value of f (the output) increases (or decreases) as we change the argument (input) x. We denote the derivative as

$$f'(x) \text{ or } \frac{df}{dx}.$$

The following figure illustrates the concept:



Formally, we define the derivative as the limit

$$\frac{df}{dx} = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}.$$

This is equivalent to looking at the slope of f between the two points x and $x + h$ and letting h tend to zero. When the limit exists, we say that f is differentiable at x. However, if the graph of f is not smooth at a point x (for example f has a jump or a corner there) then the derivative of f will not be defined at that point.

The process of finding a derivative is called **differentiation**.

2.2 Derivatives of basic functions

Constant. If $f(x) = c$, then $f'(x) = 0$.

Line. If $f(x) = cx$, then $f'(x) = c$.

Exponential. If $f(x) = e^x$, then $f'(x) = e^x$.

Logarithm. If $f(x) = \log(x)$, then $f'(x) = 1/x$.

Power. If $f(x) = x^p$, then $f'(x) = px^{p-1}$.

2.3 Derivatives of combinations of functions

Constant factor. If $h(x) = cf(x)$, then $h'(x) = cf'(x)$.

Sum rule. If $h(x) = f(x) + g(x)$, then $h'(x) = f'(x) + g'(x)$.

Difference rule. If $h(x) = f(x) - g(x)$, then $h'(x) = f'(x) - g'(x)$.

Linearity. If $h(x) = \alpha f(x) + \beta g(x)$, then $h'(x) = \alpha f'(x) + \beta g'(x)$.

Product rule. If $h(x) = f(x)g(x)$, then $h'(x) = f'(x)g(x) + g'(x)f(x)$.

Quotient rule. If $h(x) = f(x)/g(x)$, then $h'(x) = (f'(x)g(x) - g'(x)f(x))/(g(x)^2)$.

Reciprocal rule. If $h(x) = 1/f(x)$, then $h'(x) = -f'(x)/(f(x)^2)$. Special case: if $h(x) = 1/x$, then $h'(x) = -1/x^2$.

Chain rule. If $h(x) = f(g(x))$, then $h'(x) = f'(g(x))g'(x)$.

2.4 Higher order derivatives

The **second derivative** of a function f at a point x is the derivative of f' evaluated at x. We use the notation $f''(x)$ for the second derivative. Likewise, we can define higher order derivatives in the same way.

3 Functions of several variables

Denote the set of all p dimensional vectors by $\mathbb{R}^p$. Given two sets $\mathcal{X} \subseteq \mathbb{R}^p$ and $\mathcal{Y} \subseteq \mathbb{R}$, a function $f : \mathcal{X} \to \mathcal{Y}$ assigns each point $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)$ in the set $\mathcal{X}$ to a unique

$y \in \mathcal{Y}$ denoted by

$$y = f(x_1, x_2, \ldots, x_p).$$

In this case $x_1, x_2, \ldots, x_p$ are the input variables and y is the output variable. We also write $y = f(\boldsymbol{x})$.

3.1 Partial derivatives

The **partial derivative** of a function of multiple variables measures how fast the value of f (the output) increases (or decreases) as we change one of the inputs while holding the others constant. In other words, it is the derivative of the function with respect to one of the variables, while keeping the remaining constant.

For example, the partial derivative of the function with respect to the first input is

$$\frac{\partial f}{\partial x_1} = \lim_{h \to 0} \frac{f(x_1 + h, x_2, \ldots, x_p) - f(x_1, x_2, \ldots, x_p)}{h},$$

with the respect to the second is

$$\frac{\partial f}{\partial x_2} = \lim_{h \to 0} \frac{f(x_1, x_2 + h, \ldots, x_p) - f(x_1, x_2, \ldots, x_p)}{h},$$

and so on.

3.2 Gradient

The **gradient** of a function of several variables is the vector of partial derivatives, written as

$$\bigtriangledown f = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_p} \end{pmatrix}$$

or

$$\frac{df}{d\boldsymbol{x}} = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \ldots, \frac{\partial f}{\partial x_n} \right).$$

3.3 Hessian

The **Hessian** of a function of many variables is the square matrix of second order partial derivatives,

$$\mathbf{H}(f(\boldsymbol{x})) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \, \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \, \partial x_p} \\ \frac{\partial^2 f}{\partial x_2 \, \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \, \partial x_p} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_p \, \partial x_1} & \frac{\partial^2 f}{\partial x_p \, \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_p^2} \end{bmatrix}.$$

Probability for Statistical Learning and Data Mining (HTIN5005)

Contents

1 Probability

Probability is the mathematical language for quantifying uncertainty.

1.1 Sample spaces and events

A **random experiment** is a process with uncertain outcome. The **sample space** Ω is the set of all possible outcomes of a random experiment. Points ω in Ω are called sample **outcomes**, **realisations**, or **elements**. Subsets of Ω are called **events**.

Example. If we toss a coin twice, then $\Omega = \{\mathrm{HH}, \mathrm{HT}, \mathrm{TH}, \mathrm{TT}\}$. The event that the first toss is heads is $A = \{\mathrm{HH}, \mathrm{HT}\}$.

Example. Let the outcome ω be value of a stock in one year. Then $\Omega = (0, +\infty)$.

Example. If we toss a coin forever, the sample space is the infinite set

$$\Omega = \{\omega = (\omega_1, \omega_2, \omega_3, \dots,) : \omega_i \in \{\mathrm{H}, \mathrm{T}\}\}$$

1.2 Complement, union, and intersection of events

Given an event A, let $A^c = \{\omega \in \Omega : \omega \notin A\}$ denote the **complement** of A. We can read it as "not A". The complement of Ω is the empty set $\emptyset$.

We define the union of events A and B as

$$A \cup B = \{\omega \in \Omega : \omega \in A \text{ or } \omega \in B \text{ or } \omega \in \text{ both}\},$$

which we can think of as "A or B".

The intersection of A and B as

$$A \cap B = \{\omega \in \Omega : \omega \in A \text{ and } \omega \in B\},$$

read "A and B'. We sometimes write $A \cap B$ as AB or (A, B).

We say that A and B are **disjoint** or **mutually exclusive** if $A \cap B = \emptyset$.

1.3 Probability

We assign a real number $\mathbb{P}(A)$ to each event A. To qualify as a probability, $\mathbb{P}$ must satisfy the following axioms.

Definition. A function $\mathbb{P}$ that assigns a real number $\mathbb{P}(A)$ to each event A is a **probability measure** (or just **probability**) if it satisfies:

1. $0 \leq \mathbb{P}(A) \leq 1$ for every A.
2. $\mathbb{P}(\Omega) = 1$.
3. If $A_1, A_2, ...$ are disjoint then

$$\mathbb{P}(\cup A_i) = \sum_i \mathbb{P}(A_i).$$

The definition of probability implies the following useful properties.

$\mathbb{P}(\emptyset) = 0$

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

If $A \subset B$, then $\mathbb{P}(A) \leq \mathbb{P}(B)$

Lemma. For any events A and B,

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

Proof.

$$\begin{aligned}
\mathbb{P}(A \cup B) &= \mathbb{P}\left((AB^c) \cup (AB) \cup (A^cB)\right) \\
&= \mathbb{P}(AB^c) + \mathbb{P}(AB) + \mathbb{P}(A^cB) \\
&= \mathbb{P}(AB^c) + \mathbb{P}(AB) + \mathbb{P}(A^cB) + \mathbb{P}(AB) - P(AB) \\
&= \mathbb{P}\left((AB^c) \cup (AB)\right) + \mathbb{P}\left((A^cB) \cup (AB)\right) - \mathbb{P}(AB) \\
&= \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(AB)
\end{aligned}$$

□

1.4 Independent events

Two events A and B are **independent** if

$$P(A \cap B) = P(A)P(B).$$

A set of events $\{A_1, A_2, \ldots\}$ is independent if

$$\mathbb{P}(\cap_i A_i) = \prod_i \mathbb{P}(A_i)$$

We sometimes assume independence, and sometimes derive it.

1.5 Conditional probability

If $\mathbb{P}(B) > 0$ then the **conditional probability** of A given B is

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.$$

For a given B, $\mathbb{P}(\cdot|B)$ satisfies the definition of probability.

Lemma. For any pair of events, $\mathbb{P}(A \cap B) = \mathbb{P}(A|B)\mathbb{P}(B) = \mathbb{P}(B|A)\mathbb{P}(A)$.

Lemma. If A and B are independent then $\mathbb{P}(A|B) = \mathbb{P}(A)$.

In general, $\mathbb{P}(A|B) \neq \mathbb{P}(B|A)$.

1.6 Bayes' theorem

Note: we will not use Bayes' theorem this semester, but this is important for your own knowledge.

A **partition** of Ω is a collection of disjoint sets $A_1, A_2, \ldots, A_k$ such that $\bigcup_{i=1}^{k} A_i = \Omega$.

Theorem (The Law of Total Probability). Let $A_1, A_2, \ldots, A_k$ be a partition of Ω. Then, for any event B,

$$\mathbb{P}(B) = \sum_{i=1}^{k} \mathbb{P}(B|A_i)\mathbb{P}(A_i).$$

Theorem (Bayes' Theorem). Let $A_1, \ldots, A_k$ be a partition of Ω such that $\mathbb{P}(A_i) > 0$ for each i. If $\mathbb{P}(B) > 0$ then, for each $i = 1, \ldots, k$,

$$\mathbb{P}(A_i|B) = \frac{\mathbb{P}(B|A_i)\mathbb{P}(A_i)}{\sum_j \mathbb{P}(B|A_j)\mathbb{P}(A_j)}.$$

Bayes' theorem follows from definition of conditional probability and the law of total probability.

Application (Bayesian updating). Let $A_1, \ldots, A_k$ are competing hypotheses (widely defined), one of which is true. Each hypothesis has **prior probability** (initial plausibility) $\mathbb{P}(A_i)$ for $i = 1, \ldots, k$. We then receive some new information B, which makes us update the probability that each hypothesis is the correct one. In doing so, we take into account both the priors and the probability that event B would occur if each hypothesis was true.

Example. This example is from Statistical Rethinking. Consider a blood test for detecting vampires. Suppose that the test has **sensitivity** 95%,

$$\mathbb{P}(\text{positive}|\text{vampire}) = 0.95,$$

and **specificity** of 99%,

$$\mathbb{P}(\text{negative}|\text{normal}) = 0.99.$$

What is the probability that the suspect is a vampire, given that test returns positive?

If we think of this problem as Bayesian updating, the hypotheses are normal and vampire, and the new information is the result of the test. A common but mistaken intuition is that the probability that the suspect is a vampire in case of a positive result is 95%, since this the accuracy of the test for vampires. However, this ignores the prior probabilities (base rates). This logical error is known as the **base rate fallacy**.

To get the correct answer, we need to use Bayes' theorem,

$$\mathbb{P}(\text{vampire}|\text{positive}) = \frac{\mathbb{P}(\text{positive}|\,\text{vampire})\mathbb{P}(\text{vampire})}{\mathbb{P}(\text{positive})}$$

where

$$\begin{aligned}\mathbb{P}(\text{positive}) =& \mathbb{P}(\text{positive}|\,\text{vampire})\mathbb{P}(\text{vampire}) \\ &+\mathbb{P}(\text{positive}|\,\text{normal})\mathbb{P}(\text{normal}).\end{aligned}$$

Suppose that vampires are only 0.1% of the population, i.e. $\mathbb{P}(\text{vampire}) = 0.001$.

$$\mathbb{P}(\text{vampire}|\text{positive}) = \frac{0.95 \times 0.001}{0.95 \times 0.001 + 0.01 \times 0.999} = 8.7\% \qquad \square$$

2 Random variables

2.1 Definition of a random variable

Statistics and machine learning are concerned with data. The concept of a random variable links sample spaces and events to data.

Definition. A **random variable** is a mapping $X : \Omega \to \mathbb{R}$ that assigns a real number $X(\omega)$ to each outcome.

Example. Toss a fair coin twice and let X be the number of heads. The sample space is $\Omega = \{\text{TT}, \text{HT}, \text{TH}, \text{HH}\}$. Then X is a function that assigns each outcome $\omega \in \Omega$ to a number as follows:

ω	$\mathbb{P}(\omega)$	$X(\omega)$
TT	1/4	0
HT	1/4	1
TH	1/4	1
HH	1/4	2

Once we start working directly with random variables, we tend not to mention the sample space anymore. It's good to keep in mind that the sample space is always there in the background.

2.2 Distribution functions and probability functions

Definition. The **cumulative distribution function**, or CDF, is the function $F_X : \mathbb{R} \to [0, 1]$ defined by $F_X(x) = \mathbb{P}(X \leq x)$.

The CDF effectively contains all the information about a random variable. We sometimes write it as F instead of F_X.

X is **discrete** if it takes countably many values $\{x_1, x_2, \ldots\}$. We define the **probability function** or **probability mass function** for X as $f_X(x) = \mathbb{P}(X = x)$.

Example. Toss a fair coin twice and let X be the number of heads. $\mathbb{P}(X = 0) = \mathbb{P}(X = 2) = 1/4$ and $\mathbb{P}(X = 1) = 1/2$. The distribution is

$$F_X(x) = \begin{cases} 0 & x < 0 \\ 1/4 & 0 \leq x < 1 \\ 3/4 & 1 \leq x < 2 \\ 1 & x \geq 2 \end{cases}$$

The probability mass function is

$$f_X(x) = \begin{cases} 1/4 & x = 0 \\ 1/2 & x = 1 \\ 1/4 & x = 2 \\ 0 & \text{otherwise.} \end{cases}$$

A random variable X is **continuous** if there exists a function f_X such that $f_X(x) \geq 0$ for all x, $\int_{-\infty}^{\infty} f_X(x)dx = 1$, and for every $a \leq b$,

$$\mathbb{P}(a < X < b) = \int_a^b f_X(x)dx.$$

The function f_X is called a **probability density function** (PDF). We have that

$$F_X(x) = \int_{\infty}^{x} f_X(t)dt,$$

and $f_X(x) = F'_X(x)$ at all point at which F_X is differentiable.

Let X be a random variable with CDF F. If F is strictly increasing and continuous, the **inverse CDF** or **quantile** function $F^{-1}(q)$ is the unique real number x such that $F(x) = q$. $F^{-1}(1/4)$ is the first quartile, $F^{-1}(1/2)$ is the median, $F^{-1}(3/4)$ is the third quartile. You will recall using the inverse CDF to get critical values in introductory units.

2.3 Bivariate distributions

Given a pair of discrete random variables X and Y, define the **joint mass function** by $f(x, y) = \mathbb{P}(X = x \text{ and } Y = y)$. We write $\mathbb{P}(X = x \text{ and } Y = y)$ as $\mathbb{P}(X = x, Y = y)$ and f as $f_{X,Y}$.

Definition. In the continuous case, a function $f(x, y)$ is a joint PDF for the random variables (X, Y) if

1. $f(x,y) \geq 0$ for all (x,y),

2. $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dxdy = 1$, and

3. for any set $X \subset \mathbb{R} \times \mathbb{R}$, $\mathbb{P}((X,Y) \in A) = \int\int_A f(x,y)dxdy$.

In the discrete or continuous case we define the **joint CDF** as $F_{X,Y}(x,y) = \mathbb{P}(X \leq x, Y \leq y)$.

2.4 Marginal distributions

Definition. If (X,Y) have joint distribution with mass functions $f_{X,Y}$, then the **marginal mass function** for X is defined by

$$f_X(x) = \mathbb{P}(X = x) = \sum_y \mathbb{P}(X = x, Y = y) = \sum_y f(x,y),$$

and in the same way the marginal for Y is

$$f_Y(y) = \mathbb{P}(Y = y) = \sum_x \mathbb{P}(X = x, Y = y) = \sum_x f(x,y).$$

Definition. For continuous random variables, the **marginal densities** are

$$f_X(x) = \int f(x,y)dy \text{ and } f_Y(y) = \int f(x,y)dx.$$

We denote the corresponding marginal distributions as F_X and F_Y.

2.5 Independent random variables

Two random variables X and Y are **independent** if for every A and B,

$$P(X \in A, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B).$$

In practice, we use the following result.

Theorem. Let X and Y have joint PDF $f_{X,Y}$. Then X and Y are independent if and only if $f_{X,Y}(x,y) = f_X(x)f_Y(y)$ for all values x and y.

This theorem also applies to discrete random variables, even though we formulated it for continuous variables.

2.6 Conditional distributions

Then **conditional probability mass function** is

$$f_{X|Y}(x|y) = \mathbb{P}(X = x, Y = y) = \frac{\mathbb{P}(X = x, Y = x)}{\mathbb{P}(Y = y)} = \frac{f_{X,Y}(x,y)}{f_Y(y)},$$

if $f_Y(y) > 0$.

For continuous random variables, the **conditional probability density function** is

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)},$$

if $f_Y(y) > 0$. Then,

$$P(a < X < b|Y = y) = \int_a^b f_{X|y}(x|y)dy.$$

2.7 Random vectors

Let $X = (X_1, \ldots, X_n)$, where $X_1, \ldots, X_n$ are random variables. We call X a **random vector**. Let $f(x_1, \ldots, x_n)$ denote the PDF. We define the marginals, conditionals, etc in the same way as in the bivariate case.

2.8 IID samples

We say that $X_1, \ldots, X_n$ are independent if $f(x_1, \ldots, x_n) = \prod_{i=1}^n f_{X_i}(x_i)$.

If $X_1, \ldots, X_n$ are independent and each have the same marginal distribution with CDF F, we say that $X_1, \ldots, X_n$ are **IID** (**independent and identically distributed**) and write $X_1, \ldots, X_n \sim F$.

3 Some important distributions

3.1 Discrete distributions

3.1.1 Discrete uniform

Let k be an integer. Suppose that X has probability function given by

$$f(x) = \begin{cases} 1/k & \text{for } x = 1, \ldots, k \\ 0 & \text{otherwise.} \end{cases}$$

We say that X has a uniform distribution on $\{1, \ldots, k\}$.

3.1.2 Bernoulli

Let X be a binary variable. Then $\mathbb{P}(X = 1) = p$ and $\mathbb{P}(X = 0) = 1 - p$ for some $p \in [0, 1]$. We say that X has the Bernoulli distribution and write $X \sim \text{Bernoulli}(p)$. The probability function is $f(x) = p^x(1-p)^{(1-x)}$ for $x \in \{0, 1\}$.

3.1.3 Binomial distribution

Let X be the sum of n independent Bernoulli random variables with parameter p. Then X follows the binomial distribution, which has probability function

$$f(x) = \binom{n}{k} p^x (1-p)^x,$$

$x = 0, 1, \ldots, n$. We write $X \sim \text{Binomial}(n, p)$.

3.1.4 Poisson distribution

Let X be a discrete random variable with probability function

$$f(x) = \mathbb{P}(X = x) = e^{-\lambda}\frac{\lambda^x}{x!}, \quad \lambda > 0,$$

for $x \geq 0$. Then X is a **Poisson** random variable with parameter λ. We write $X \sim \text{Poisson}(\lambda)$. We often use the Poisson to model rare events, such as the number of corporate defaults in a given amount of time.

3.2 Continuous distributions

3.2.1 Uniform distribution

X has a uniform distribution between a and b, written $X \sim \text{Uniform}(a, b)$, if

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{for } x \in [a, b] \\ 0 & \text{otherwise,} \end{cases}$$

where $b > a$. We have that

$$\mathbb{P}(u < X < v) = \frac{v - u}{b - a}$$

for $a \leq u < v \leq b$.

3.2.2 Normal (Gaussian) distribution

X has the normal (Gaussian) distribution with parameters μ and σ, denoted by $X \sim N(\mu, \sigma^2)$, if

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

The parameter μ is the mean (centre) of the distribution and σ is the standard deviation of the distribution, to be defined in the next section.

We say that X has the **standard normal distribution** when $\mu = 0$ and $\sigma = 1$,

traditionally denoted Z.

We often use the following properties:

(i) If $X \sim N(\mu, \sigma^2)$, then $Z = (X - \mu)/\sigma \sim N(0, 1)$.

(ii) If $Z \sim N(0, 1)$, then $X = \mu + \sigma Z \sim N(0, 1)$.

(iii) If $X_i \sim N(\mu_i, \sigma_i^2)$ for $i = 1, \ldots, n$ are independent, then

$$\sum_{i=1}^{n} X_i \sim N\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right).$$

3.2.3 Exponential distribution

X has the an exponential distribution with rate parameter λ, denoted $X \sim \text{Exp}(\lambda)$, if

$$f(x) = \lambda e^{-\lambda x}, \quad x > 0$$

where $\lambda > 0$. We often use the exponential distribution to model the time between events.

3.2.4 The t and Cauchy distributions

X has as a t distribution with ν degrees of freedom, written as $X \sim t_\nu$, if

$$f(x) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}.$$

The t distribution is similar to the Gaussian distribution, but with thicker tails. When $\nu \to \infty$ the t distribution tends to the Gaussian distribution.

The Cauchy distribution is a special case of the t distribution when $\nu = 1$. The density

has the simple form

$$f(x) = \frac{1}{\pi(1+x^2)}.$$

4 Expectation

4.1 Expectation

We define the **expected value** or **mean** of a random variable X as

$$\mathbb{E}(X) = \begin{cases} \sum_x x f(x) & \text{if } X \text{ is discrete} \\ \int_x f(x)dx & \text{if } X \text{ is continuous,} \end{cases}$$

provided that the sum or integral is well defined.

We often use the notation $\mathbb{E}(X) = \mathbb{E}X = \mu = \mu_X$.

We can think of $\mathbb{E}(X)$ as the average $\sum_i^n X_i/n$ for a very large number of IID draws $X_1, \ldots, X_n$.

4.2 Properties of expectations

If X and Y are random variables and a and b are constants,

$$\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}(Y).$$

More generally, if $X_1, \ldots, X_n$ are random variables and $a_1, \ldots, a_n$ are constants,

$$\mathbb{E}\left(\sum_{i=1}^n a_i X_i\right) = \sum_{i=1}^n a_i \mathbb{E}(X_i).$$

We call above property the linearity of expectations. Note that it does not require independence, unlike the result below.

Theorem. Let $X_1, \ldots, X_n$ be independent random variables. Then

$$\mathbb{E}\left(\prod_{i=1}^{n} X_i\right) = \prod_{i=1}^{n} \mathbb{E}(X_i).$$

4.3 Variance and covariance

Let X be a random variable with mean μ. We define the **variance** of X, denoted by $\mathbb{V}(X)$, σ^2, or σ_x^2 as

$$\sigma^2 = \mathbb{E}(X - \mu)^2.$$

If X and Y are random variables, then the covariance and correlation between X and Y measure the strength of linear relationship between the two variables.

Let X and Y be random variables with means μ_X and μ_Y and standard deviations σ_X and σ_Y. We define the **covariance** between X and Y as

$$\mathsf{Cov}(X, Y) = \mathbb{E}\left((X - \mu_X)(Y - \mu_Y)\right),$$

as the **correlation** as

$$\rho = \rho_{X,Y} = \rho(X, Y) = \frac{\mathsf{Cov}(X, Y)}{\sigma_X \sigma_Y}.$$

The correlation satisfies $-1 \leq \rho_{X,Y} \leq 1$.

4.4 Properties of variances and covariances

We frequently use the following properties.

$$\mathbb{V}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$$

$$\mathbb{V}(aX + b) = a^2 \mathbb{V}(X)$$

$$\mathsf{Cov}(X,Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$$

$$\mathsf{Cov}(X,Y) = \mathsf{Cov}(Y,X)$$

$$\mathsf{Cov}(X,X) = \mathbb{V}(X)$$

$$\mathbb{V}(X+Y) = \mathbb{V}(X) + \mathbb{V}(Y) + 2\mathsf{Cov}(X,Y)$$

$$\mathbb{V}(X-Y) = \mathbb{V}(X) + \mathbb{V}(Y) - 2\mathsf{Cov}(X,Y)$$

$$\mathbb{V}(aX+bY) = a^2\mathbb{V}(X) + b^2\mathbb{V}(Y) + 2ab\mathsf{Cov}(X,Y)$$

$$\mathsf{Cov}(aX+b,Y) = a\mathsf{Cov}(X,Y)$$

$$\mathsf{Cov}(aX,bY) = ab\mathsf{Cov}(X,Y)$$

$$\mathsf{Cov}(X,Y+Z) = \mathsf{Cov}(X,Y) + \mathsf{Cov}(X,Z)$$

If $X_1, \ldots, X_n$ are independent random variables and $a_1, \ldots, a_n$ are constants,

$$\mathbb{V}\left(\sum_{i=1}^n a_i X_i\right) = \sum_{i=1}^n a_i^2 \mathbb{V}(X_i).$$

More generally, if $X_1, \ldots, X_n$ are random variables and $a_1, \ldots, a_n$ are constants,

$$\mathbb{V}\left(\sum_{i=1}^n a_i X_i\right) = \sum_{i=1}^n a_i^2 \mathbb{V}(X_i) + 2\sum\sum_{i<j} a_i a_j \mathsf{Cov}(X_i, X_j).$$

4.5 Expectation and variance of important random variables

Distribution	Mean	Variance
Bernoulli(p)	p	$p(1-p)$
Binomial(n, p)	np	$np(1-np)$
Poisson(λ)	λ	λ
Uniform(a, b)	$(a+b)/2$	$(b-a)^2/12$
Normal(μ, σ^2)	μ	σ^2
Exponential(λ)	$1/\lambda$	$1/\lambda^2$
t_ν	0 if $\nu > 1$	$\nu/(\nu-2)$ if $\nu > 2$

4.6 Conditional expectation

The **conditional expectation** of X given Y is

$$\mathbb{E}(X|Y=y) = \begin{cases} \sum x f_{X|Y}(x|y) & \text{if } X \text{ is discrete} \\ \int x f_{X|Y}(x|y)dx & \text{if } X \text{ is continuous.} \end{cases}$$

The conditional expectation $\mathbb{E}(X|Y=y)$ is a function of y. When we write $\mathbb{E}(X|Y)$, we refer to a random variable denoted as such; we haven't yet observed Y.

Theorem (Law of Iterated Expectations). For random variables X and Y, we have that

$$\mathbb{E}(Y) = \mathbb{E}\left[\mathbb{E}(Y|X)\right] \text{ and } \mathbb{E}(X) = \mathbb{E}\left[\mathbb{E}(X|Y)\right].$$

The **conditional variance** is

$$\mathbb{V}(Y|X=x) = \int (y-\mu(x))^2 f(y|x)dy,$$

where $\mu(x) = \mathbb{E}(Y|X=x)$.

Theorem (Law of total variance). For random variables X and Y,

$$\mathbb{V}(Y) = \mathbb{E}\mathbb{V}(Y|X) + \mathbb{V}\mathbb{E}(Y|X).$$

5 Convergence of random variables

5.1 The Law of Large Numbers

Let $X_1, X_2, \ldots$ be a sequence of random variables and let c be a constant. We say that X_n **converges in probability** to X, written $X_n \xrightarrow{P} c$ if for every $\epsilon > 0$,

$$\mathbb{P}(|X_n - c| > \epsilon) \to 0$$

as $n \to \infty$. In words, we can pick any arbitrarily small $\epsilon > 0$, and the probability that the distance between X_n and c will be more than ϵ tends to zero when $n \to \infty$.

Theorem (The Weak Law of Large Numbers (WLLN)). Let $X_1, X_2, \ldots, X_n$ be an IID sample, where $\mathbb{E}(X_i) = \mu$ and $\mathbb{V}(X_i) = \sigma^2$ for $i = 1, \ldots, n$. Let the **sample mean** be $\overline{X}_n = n^{-1} \sum_{i=1}^{n} X_i$. Then $X_n \xrightarrow{P} \mu$.

The law of large numbers says that the distribution of $\overline{X}$ becomes more concentrated around μ as n gets larger.

5.2 The Central Limit Theorem

Note: we will not use the CLT in this unit, it's here just for completeness.

Let $X_1, X_2, \ldots$ be a sequence of random variables and let X be another random variable. Let F_n denote the CDF of X_n and F denote the CDF of X. We say that X_n **converges in distribution** to X, written $X_n \rightsquigarrow$, if

$$\lim_{n \to \infty} F_n(x) = F(x),$$

for every x.

Theorem (Central Limit Theorem (CLT)). Let $X_1, X_2, \ldots, X_n$ be an IID with mean μ and variance σ^2 $\overline{X}_n = n^{-1} \sum_{i=1}^n X_i$. Then

$$Z_n \equiv \frac{\overline{X}_n - \mu}{\sqrt{\mathbb{V}(\overline{X}_n)}} = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \rightsquigarrow Z,$$

where $Z \sim N(0,1)$.

Interpretation: when n is large, we can approximate probability statements about $\overline{X}_n$ using a normal distribution. We often use the notation $Z_n \approx N(0,1)$ to indicate that the distribution of Z_n is converging to a normal.

Theorem. Assume the same conditions as the Central Limit Theorem, and let

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \overline{X}_n)^2$$

be the sample standard deviation. Then

$$\frac{\sqrt{n}(\overline{X}_n - \mu)}{S_n} \rightsquigarrow N(0,1).$$

.

References

Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. The MIT Press.

Wasserman, L. (2013). *All of statistics: a concise course in statistical inference*. Springer Science & Business Media.