School of Electronic Engineering and Computer Science
Queen Mary University of London

CBU5201 Machine Learning
A bit of notation and basic maths

Credit to Dr Jesús Requena Carrión

26 Sept 2022

Agenda

The dataset

Datasets are collections of **items** described by a set of **attributes**.

Animal (ID)	Body mass [g]	Heart rate [bpm]
Wild mouse	22	480
Rabbit	$2.5{\times}10^3$	250
Humpback whale	$30{\times}10^6$	30
...	...	...

Types of attributes

The basic types of attributes are:

- **Continuous** (real numbers: temperature, voltage)
  - Equality, ordering and distance are defined
- **Categorical** (discrete, nominal: name, nationality)
  - Equality is defined, ordering and distance are not
- **Ordinal** (categories with ordering: low/medium/high)
  - Ordering and equality are defined, distance is not

The dataset as a table

Datasets can be represented as tables, where **rows correspond to items** and **columns to attributes** .

The first 5 instances of a dataset recording the age and salary of a group of people are shown below in a table form:

	Age	Salary
S_1	18	12000
S_2	37	68000
S_3	66	80000
S_4	25	45000
S_5	26	30000
...	...	...

The values S_1, S_2... are not attributes, but identifiers.

The dataset as a matrix

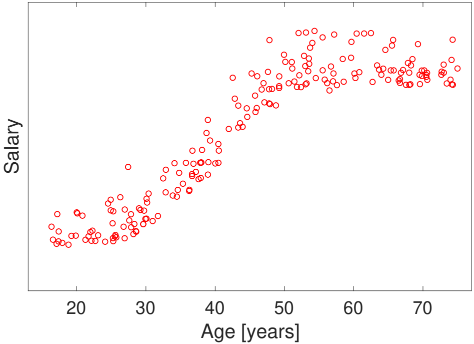
Datasets consisting of **numerical attributes** can also be represented in **matrix** form, for instance:

$$S = \begin{pmatrix} 18 & 12000 \\ 37 & 68000 \\ 66 & 80000 \\ 25 & 45000 \\ 26 & 30000 \\ \vdots & \vdots \end{pmatrix}$$

This is a useful and compact notation that will make it easier for us to formulate problems and represent computations.

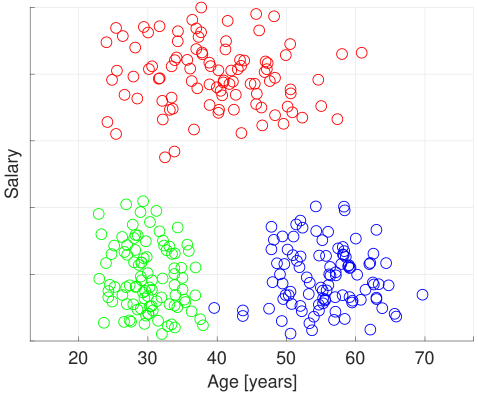
The dataset as a point cloud

Datasets can be represented **sets of points** in a space (known as the **feature** or **attribute space**), where each axis corresponds to one attribute. The values of the attributes are used as coordinates.



The dataset as a point cloud

Categorical values can be represented by **different symbols** in a point cloud, instead of a value in an axis.



From basket to table

In market basket analysis, raw items consist of a list of products, for instance:

$$
\begin{aligned}
S_1 &= \{\text{The Beatles}, \text{The Who}, \text{Cream}\} \\
S_2 &= \{\text{Muse}, \text{Franz Ferdinand}\} \\
S_3 &= \{\text{The Who}, \text{Franz Ferdinand}\} \\
\ldots &
\end{aligned}
$$

Such datasets can be represented as a table, where each product corresponds to a binary attribute:

	The Beatles	The Who	Cream	Muse	Franz Ferdinand
S_1	1	1	1	0	0
S_2	0	0	0	1	1
S_3	0	1	0	0	1
$\ldots$	$\ldots$	$\ldots$			

Datasets with grid support

Digital signals and images are collections of values on a temporal and spatial grid. Each value can be treated as separate attribute.

S_1

S_2

S_3

A collection of images consisting of 28×28 pixels can be represented as a table with 784 columns, where each column represents one pixel.

	x_1	x_2	x_3	$\dots$	x_{784}
S_1	0	0.1	0	$\dots$	0
S_2	0	0.1	0	$\dots$	0
S_3	0	0.7	0.6	$\dots$	0.1
$\dots$	$\dots$	$\dots$			

It is however more useful to represent each individual image as a 28×28 array, where each entry corresponds to a pixel value.

Agenda

Scalars

A **scalar** is value consisting of one numerical quantity. Scalars are usually represented by a letter of the alphabet in italics, for instance a.

When dealing with several scalars, we can use **different symbols** (a, b, c), or use **one symbol** with a **subscript** (also known as **index**) whose value identifies each scalar (a_1, a_2, a_3).

Using subscript notation, large collections of scalars can be represented, for instance $a_0, a_1, \ldots, a_{99}$ or equivalently a_i, where $0 \leq i \leq 99$.

Sum and product notation

The sum and product notations provide a compact way of expressing operations involving many scalars.

The **sum notation** uses the symbol Σ (meaning "sum"):

$$\sum_{i=1}^{N} a_i = a_1 + a_2 + a_3 + \cdots + a_N$$

It reads as *the sum of all the scalars* a_i *starting from* $i = 1$ *through* $i = N$. If the subscripts are known, we can write $\sum_i a_i$ or simply $\sum a_i$.

The **product notation** uses the symbol Π (meaning "product"):

$$\prod_{i=1}^{N} a_i = a_1 \times a_2 \times a_3 \times \cdots \times a_N$$

and sometimes will be written as $\prod_i a_i$ or $\prod a_i$.

Linear combination

Given N scalars a_i, a linear combination is the sum

$$\sum_{i=1}^{N} b_i a_i = b_1 \times a_1 + b_2 \times a_1 + b_3 \times a_3 + \cdots + b_N \times a_N$$

where the N scalars b_i are the **weights** of the linear combination.

Vectors

In some situations, we are interested in values that are represented by an **ordered arrangement of scalars**, for instance the coordinates of a point or a digital picture consisting of an array of pixels.

Vectors are **1D arrays of scalars arranged in order**. Typically vectors are represented in bold typeface and its elements are written in italic typeface with a subscript. By using brackets, a vector consisting of N elements is represented as a **column**:

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

Matrices

A matrix is a **2D array of scalars**. Matrices are usually represented as an uppercase variable in bold typeface and its elements are written in italic typeface with two subscripts instead of one, for instance:

$$\mathbf{A} = \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,M} \\ a_{2,1} & a_{2,2} & & a_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ a_{N,1} & a_{N,2} & \cdots & a_{N,M} \end{bmatrix}$$

Matrix $\mathbf{A}$ is said to have N rows (height) and M columns (width) or to be an $N \times M$ matrix, where $N \times M$ is known as the **shape** of the matrix.

Individual elements of the matrix are denoted by $a_{i,j}$, where i is the row number and j the column number.

Some special matrices

- A **vector** can be seen as an $N \times 1$ matrix, i.e. a matrix consisting of just one column
- A **square matrix** has the same number of rows and columns, i.e. its an $N \times N$ matrix. Its **diagonal** consists of the elements where both subscripts are identical, i.e. $a_{1,1}, a_{2,2}, \ldots$
- A **diagonal matrix** is a square matrix such that all its entries (except the diagonal) are zero. Diagonal entries can be nonzero or zero
- The **identity matrix** $\mathbf{I}$ is the diagonal matrix with 1's on the diagonal and 0's elsewhere:

$$\mathbf{I}_{3\times 3} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Transpose

The transpose $\mathbf{A}^T$ of a matrix $\mathbf{A}$ is a matrix created by interchanging rows and columns, reading the rows from **left to right** and the columns from **top to bottom**.

For instance, if $\mathbf{A} = \begin{bmatrix} 1 & 4 \\ 5 & 6 \\ 0 & 3 \end{bmatrix}$, then $\mathbf{A}^T = \begin{bmatrix} 1 & 5 & 0 \\ 4 & 6 & 3 \end{bmatrix}$

Note that the transpose of an $N \times M$ matrix is an $M \times N$ matrix and hence the transpose of a $N \times 1$ vector is a single-row $1 \times M$ matrix.

Matrix addition and multiplication by scalar

We can add two matrices $\mathbf{A}$ and $\mathbf{B}$ as long as they have the **same shape**, by adding element-wise, for instance:

$$\begin{bmatrix} 1 & 4 \\ 5 & 6 \\ 0 & 3 \end{bmatrix} + \begin{bmatrix} 2 & 0 \\ -1 & 2 \\ 0 & -2 \end{bmatrix} = \begin{bmatrix} 3 & 4 \\ 4 & 8 \\ 0 & 1 \end{bmatrix}$$

By using the mathematical notation that we have developed, the elements $c_{i,j}$ of the matrix $\mathbf{C} = \mathbf{A} + \mathbf{B}$ are defined as $c_{i,j} = a_{i,j} + b_{i,j}$.

Similarly, if b is a scalar, the elements $c_{i,j}$ of the matrix $\mathbf{C} = b \times \mathbf{A}$ are defined as $c_{i,j} = b \times a_{i,j}$.

Matrix multiplication

The product of matrices represent **linear transformations**. The product $\mathbf{C} = \mathbf{AB}$ exists if the number of columns of $\mathbf{A}$ is the same as the number of rows of $\mathbf{B}$. The entries of $\mathbf{C}$ are defined as the linear combination:

$$c_{i,j} = \sum_k a_{i,k} b_{k,j}$$

Visually:

$$B = \begin{pmatrix} b_{1,1} & b_{1,2} & \cdots & b_{1,M} \\ b_{2,1} & b_{2,2} & \cdots & b_{2,M} \\ \vdots & \vdots & \vdots & \\ b_{P,1} & b_{P,2} & \cdots & b_{P,M} \end{pmatrix}$$

$$A = \begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,P} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,p} \\ \vdots & \vdots & \vdots & \\ a_{N,1} & a_{N,2} & \cdots & a_{N,P} \end{pmatrix} \quad \begin{pmatrix} c_{1,1} & c_{1,2} & \cdots & c_{1,M} \\ c_{2,1} & c_{2,2} & \cdots & c_{2,M} \\ \vdots & \vdots & \vdots & \\ c_{N,1} & c_{N,2} & \cdots & c_{N,M} \end{pmatrix} = AB$$

For instance, $c_{1,2} = \sum_K a_{1,K} b_{K,2} = a_{1,1}b_{1,2} + a_{1,2}b_{2,2} + \cdots + a_{1,P}b_{P,2}$. If $\mathbf{A}$ is $N \times P$ and $\mathbf{B}$ is $P \times M$, $\mathbf{C} = \mathbf{AB}$ is $N \times M$.

Matrix inversion

The inverse $\mathbf{A}^{-1}$ of a **square** matrix $\mathbf{A}$ is a matrix such that there product is the identity matrix $\mathbf{I}$:

$$\mathbf{A}^{-1}\mathbf{A} = \mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$$

Not all the matrices have an inverse.

Agenda

The notion of function

Given two sets of objects, called domain and codomain, a function is a **rule** that associates (or maps) each element in the domain to exactly one element in the codomain.

The notation

$$y = f(x)$$

reads *function* f *maps* x *to* y. The scalar x is known as the **independent variable** and y as the **dependent variable**.

Functions can be represented graphically as for example, curves in a 2D space or surfaces in a 3D space, and in turn, functions can be used to describe curves and surfaces.

The straight line

The simplest function that maps one scalar value x to another scalar value y is the rule represented by the linear operation

$$y = w_0 + w_1 x$$

which corresponds to a straight line with slope w_1 and intercept point w_0.

Planes and hyperplanes

Plane surfaces in 3D spaces map two scalar values x_1 and x_2 to one scalar value y and are represented by the linear combination

$$y = w_0 + w_1 x_1 + w_2 x_2$$

The notions of straight line and plane can be readily extended to any number of independent variables:

$$y = w_0 + w_1 x_1 + \cdots + w_n x_n$$

The corresponding geometrical object is known as a hyperplane.

Straight lines, planes and hyperplanes in vector notation

Straight lines, planes and hyperplanes are defined by a **linear equation**, i.e. the dependent variable is a linear combination of the dependent variables.

Using vector notation, straight lines, planes and hyperplanes can be defined by the **same equation**:

$$y = \boldsymbol{w}^T \boldsymbol{x} = w_0 + w_1 x_1 + \cdots + w_P x_P$$

where $\boldsymbol{w} = [w_0, w_1, \ldots, w_P]^T$ is the weight vector, $\boldsymbol{x} = [1, x_1, \ldots, x_P]^T$ is a vector containing all the independent variables and P is the number of independent variables.

Straight lines, planes and hyperplanes in vector notation

If we now define $x_{P+1} = y$, then

$$\begin{aligned} x_{P+1} &= \boldsymbol{x}^T \boldsymbol{w} \\ 0 &= \boldsymbol{x}^T \boldsymbol{w} - x_{P+1} \\ 0 &= \boldsymbol{x}'^T \boldsymbol{w}' \end{aligned}$$

where $\boldsymbol{w}' = [w_0, w_1, \ldots, w_P, -1]^T$ and $\boldsymbol{x}' = [1, x_1, \ldots, x_P, x_{P+1}]^T$.
Therefore, straight lines, planes and hyperplanes will sometimes be defined by the equation

$$\boldsymbol{x}^T \boldsymbol{w} = 0$$

This equation should be read as follows: *every point $\boldsymbol{x}$ such that $\boldsymbol{x}^T \boldsymbol{w} = 0$ belongs to the hyperplane defined by the weights $\boldsymbol{w}$*.

Agenda

The dataset

Linear algebra

Linear functions

Basic probability and statistics

Random variables

- A **random variable** is a variable that can take on different values randomly.
- An **experiment** is the act of producing such value, which we call **outcome**. Example: rolling a die or tossing a coin.
- Random variables can be **discrete** (e.g. a die) or **continuous** (e.g. tomorrow's temperature).
- An **event** is a set of values that a random variable can take. For instance, the values $\{1, 4, 5\}$ constitute an event in the case of the die, *heads* is an elementary event in the case of the coin.

Probability

A **probability** $P(x)$ allows to quantify how likely a random variable is to take on the values in an event x:

- $P(x) = 1$ indicates that it is certain that the random variable will take on one of the values in x.
- $P(x) = 0$ indicates that it is impossible.

Given two random variables, the **joint probability** $P(x, y)$ describes how likely the first random variable is to take on a value in x and the second random variable is to take on a values in y.

Finally, a **conditional probability** $P(x|y)$ is the probability that the first variable takes on a value in x **given that we know** that the second has taken on a value in y.

Bayes' Theorem

Bayes' Theorem gives us a simple way to calculate a conditional probability $P(x|y)$ from a probability $P(y|x)$:

$$P(x|y) = \frac{P(y|x)P(x)}{P(y)}$$

In the context of Bayes' Theorem, we will call $P(x)$ and $P(y)$ **priors**, and $P(x|y)$ and $P(y|x)$ **posteriors**.

Probability distributions

- In general, the probability that a continuous random variable takes on a specific value is 0. Continuous random variables are best described by **probability distributions**, which quantify the density of probability, rather than the probability itself.
- The probability of an event can be calculated by **integrating** the distribution (i.e. obtaining its area or volume).

Gaussian distribution

An example of probability distribution is the Gaussian or normal distribution, which is defined as

$$N(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$