A short mathematics preliminaries

Lecture 3: 2DV516

Jonas Nordqvist

jonas.nordqvist@lnu.se

2021

Agenda

- Some math preliminaries for the course
  - Linear algebra
  - Calculus
  - Statistics and probability theory

Why do we need mathematics in ML?

Most of this course can be done without explicit knowledge about higher mathematics, but in order to fully comprehend and appreciate the material and to develop it further mathematics is needed.

There are several aspects of machine learning which is closely related to mathematics

- Finding the optimal parameters for a particular algorithm to the given data set (calculus, optimization)
- Representing our algorithms in vectorized form (linear algebra)
- Different notions of distance between vectors (data), norms, etc. (*e.g.* linear algebra)
- Predicting probabilities, learning distributions (statistics, probability theory)
- Clustering, dimensionality reduction (linear algebra)
- Measures of central tendency, spread of data, hypothesis testing (statistics)
- Estimating the uncertainty of an algorithm (statistics, probability theory)

Linear algebra

A basic course in Linear algebra is sufficient for this course. However, you might need to refreshen a bit, and what do you need to refresh?

- Matrix algebra
- Matrix transpose and inverse
- Norms (notion of lengths of vectors)
- Scalar product (dot product)
- Projections
- Vectorization
- Eigenvalues and eigenvectors

Matrix algebra

I'm assuming this to be known but I'm including this here for reference, and refer to any text on Linear algebra for anyone who needs a more extended update.

Definition

Let $A = (a_{ij})_{p\times n}$ and $B = (b_{ij})_{n\times q}$. We define the product $A \cdot B$ as the matrix $C = (c_{ij})_{p\times q}$ where,

$$c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj} = a_{i1} b_{1j} + a_{i2} b_{2j} + \ldots + a_{in} b_{nj}.$$

To be able to multiply the matrices A and B we require that the number of columns of A is equal to the number of rows of B. As a help for the memory we can note that the element c_{ij} in the product $C = AB$ is the element-wise multiplication of the ith row in A and the jth column of B.

Matrix algebra

Example

Let

$$A = \begin{pmatrix} 1 & 0 & 2 \\ 4 & -1 & 2 \end{pmatrix}, \quad B = \begin{pmatrix} 3 & 1 \\ 5 & 6 \\ 1 & 0 \end{pmatrix}.$$

Then

$$\begin{aligned} AB &= \begin{pmatrix} 1 & 0 & 2 \\ 4 & -1 & 2 \end{pmatrix} \begin{pmatrix} 3 & 1 \\ 5 & 6 \\ 1 & 0 \end{pmatrix} \\ &= \begin{pmatrix} 1\cdot 3 + 0\cdot 5 + 2\cdot 1 & 1\cdot 1 + 0\cdot 6 + 2\cdot 0 \\ 4\cdot 3 + (-1)\cdot 5 + 2\cdot 1 & 4\cdot 1 + (-1)\cdot 6 + 2\cdot 0 \end{pmatrix} = \begin{pmatrix} 5 & 1 \\ 9 & -2 \end{pmatrix}. \end{aligned}$$

Matrix algebra

Note that BA from the previous example is

$$BA = \begin{pmatrix} 7 & -1 & 8 \\ 29 & -6 & 22 \\ 1 & 0 & 2 \end{pmatrix}.$$

Hence, matrix multiplication is obviously *not* commutative. But however, it is associative and we also have that multiplication is distributive over addition. More specifically let A be of size $p \times n$ and B of size $n \times m$ and C of size $m \times q$. Then the associative law holds, i.e.

$$A(BC) = (AB)C.$$

Furthermore, if A and B are of size $p \times n$ and C of size $n \times q$ and D of size $m \times p$. Then the distributive laws hold

$$(A + B)C = AC + BC,$$

and

$$D(A + B) = DA + DB.$$

Transpose and inverse

Let $A = (a_{ij})$ be an $m \times n$ matrix the transpose of A denoted A^T is the $n \times m$ matrix such that $A^T = (a_{ji})$.

Example

$$\text{If } A = \begin{pmatrix} 1 & 2 & 3 \\ 3 & 4 & 5 \end{pmatrix}, \quad \text{then } A^T = \begin{pmatrix} 1 & 3 \\ 2 & 4 \\ 3 & 5 \end{pmatrix}.$$

The inverse of a $n \times n$ matrix A is a matrix B such that

$$AB = BA = I.$$

The inverse is denoted by A^{-1}.

Vector norms

A norm is a function which assigns a strictly positive value (length) for every nonzero vector in a vector space. That is for a vector space $\mathcal{V}$:

$$||\cdot||\colon \mathcal{V} \to [0, +\infty).$$

For $a \in \mathbb{R}$ and $\mathbf{u}, \mathbf{v} \in \mathcal{V}$ we have

$$||\,\mathbf{u} + \mathbf{v}\,|| \leq ||\,\mathbf{u}\,|| + ||\,\mathbf{v}\,||$$

$$||a\,\mathbf{u}\,|| = |a|||\,\mathbf{u}\,||,$$

and

$$||\,\mathbf{u}\,|| = 0 \implies \mathbf{u} = 0.$$

Example

The *Euclidean norm* is a vector norm on $\mathbb{R}^n$, *i.e.* if $n = 2$ and $\mathbf{u} = (u_1, u_2)$ then

$$||\,\mathbf{u}\,|| = \sqrt{u_1^2 + u_2^2}.$$

Scalar product and vector projections

Scalar product is a binary operation taking two vectors and returning one scalar value. Let $\mathbf{u} = (u_1, \ldots, u_n)^T$ and $\mathbf{v} = (v_1, \ldots, v_n)^T$, then $\mathbf{u} \bullet \mathbf{v}$ is defined as

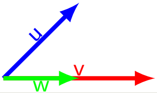
$$\mathbf{u} \bullet \mathbf{v} = \sum_{i=1}^{n} u_i v_i = \mathbf{u}^T \mathbf{v}.$$

Among other things this implies that if $||\cdot||$ is the Euclidean norm then $||\,\mathbf{u}\,|| = \sqrt{\mathbf{u} \bullet \mathbf{u}}$.

Let $\mathbf{u}, \mathbf{v}$ be vectors in $\mathbb{R}^2$, then the (orthogonal) projection of $\mathbf{u}$ onto $\mathbf{v}$ is the vector $\mathbf{w} = a\,\mathbf{v}$ for some real number a.

These projections are given by

$$\mathbf{w} = \frac{\mathbf{u} \bullet \mathbf{v}}{||\,\mathbf{v}\,||}.$$



Vectorization

Vectorization is a style of programming aiming to apply algorithms on entire arrays (or vectors) instead of doing operations element-wise. One (obvious) upside of this is speed, and another is readability and structure for your code.

Two simple examples from Python are

Example

Compute the sum of the elements of a vector **x**. We can use either of the following

- `s = 0 for i in range(len(x)): s = s + x[i]`
- `s = numpy.sum(x).`

Example

Compute the sum of the (element-wise) product of two vectors **u** and **v**. Any of the following would solve the problem

- `s = 0; for i in range(len(x)): s = s + u[i]⋆v[i]`
- `s = np.dot(u,v)`

Eigenvalues and eigenvectors

Let M be a matrix in $n \times n$-matrix. Then the eigenvectors of M are n-dimensional vectors $\mathbf{u}$ (not the zero vector) satisfying the equation

$$M\,\mathbf{u} = \lambda\,\mathbf{u},$$

for some λ known as the eigenvalue of $\mathbf{u}$.

Example

Let

$$M = \begin{pmatrix} 1 & 2 \\ 4 & 3 \end{pmatrix}.$$

The eigenvectors of M are $\mathbf{u} = (1,2)^T$ and $\mathbf{v} = (-1,1)^T$ with corresponding eigenvalues $\lambda_{\mathbf{u}} = 5$ and $\lambda_{\mathbf{v}} = -1$ as

$$M\,\mathbf{u} = \begin{pmatrix} 5 \\ 10 \end{pmatrix} \quad \text{and} \quad M\,\mathbf{v} = \begin{pmatrix} 1 \\ -1 \end{pmatrix}.$$

Calculus

The following are central notions used in several parts of the course, of which the last is perhaps what is new to some people.

- Function and functions of several variables
- Derivative
- Gradients

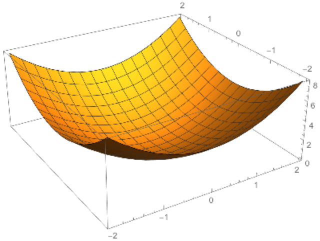
We will dedicate this section to understand the concept of a gradients for functions of several variables, and how this tell us about the greatest change in a function.

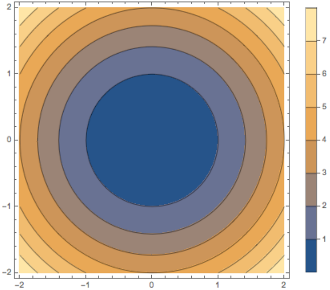
Functions of several variables

A function of several variables is a function that takes more than one input argument. Since, this is a function there is a single output value.

An example of such a function in two variables is

$$f(x, y) = x^2 + y^2.$$





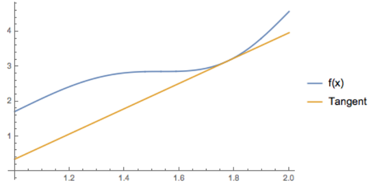
To the left we have the plot of $f(x, y)$ in $\mathbb{R}^3$, and to the right is the contour plot.

Univariate case: Derivative

We recall that the derivative of a function $f\colon\ \mathbb{R} \to \mathbb{R}$, denoted $f'(x)$, is itself a function of x given by

$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}. \tag{1}$$

In the one-dimensional case the derivative at a given point a is the slope of the *tangent* of the graph $f(x)$ at the point $x = a$.



Existence

A function f is said to be differentiable at a point a if the limit

$$\lim_{h \to 0} \frac{f(a+h) - f(a)}{h}$$

exists.

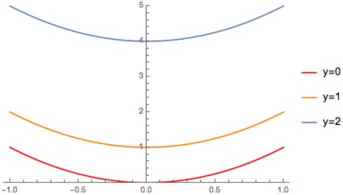
Some examples of non-differentiable functions are $f(x) = |x|$ at $x = 0$, or $f(x) = \max\{x, 0\}$ at $x = 0$.

Multidimensional case: Partial derivatives

A partial derivative of a function $f(x_1, \ldots, x_n)$ is its derivative with respect to one of the variables $x_1, \ldots, x_n$.

As opposed to the univariate case. To each point of the surface of a function of several variables there are *infinitely* many tangent lines.

The partial derivative at $(x_1, \ldots, x_n) = (a_1, \ldots, a_n) = \mathbf{a}$ of the variable x_1 is for instance the slope of the tangent line $\mathbf{a}$ in the direction parallel x_1.



The partial derivative of a function f with respect to a variable x_j is typically denoted $\frac{\partial f}{\partial x_j}$.

Gradient

Let $f\colon \mathbb{R}^n \to \mathbb{R}$ be a function in the variables $x_1, \ldots, x_n$. Then the *gradient* of f denoted ∇f is the n-dimensional vector given by

$$\nabla f = \left(\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_n} \right),$$

where each $\frac{\partial f}{\partial x_i}$ is the partial derivative of f with respect to x_i.

The direction of the vector ∇f is the direction of greatest change of f, and the size of ∇f is proportional to the size of the change.

Gradient example

Note that in practice when we differentiate f with respect to x then y should be treated as a constant.

Example

Let $f(x, y) = \exp(2xy) + x^2y^3 - x\sin(y)$. Compute ∇f.

Gradient example

Note that in practice when we differentiate f with respect to x then y should be treated as a constant.

Example

Let $f(x,y) = \exp(2xy) + x^2y^3 - x\sin(y)$. Compute ∇f.

Solution.

We have

$$\frac{\partial f}{\partial x} = 2y\exp(2xy) + 2xy^3 - \sin(y)$$

and

$$\frac{\partial f}{\partial y} = 2x\exp(2xy) + 3x^2y^2 - x\cos(y).$$

Thus,

$$\nabla f = \left(2y\exp(2xy) + 2xy^3 - \sin(y), 2x\exp(2xy) + 3x^2y^2 - x\cos(y)\right).$$

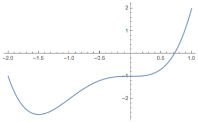
□

Minimizing a function

Assuming that a differentiable function has a *local or global* minimum at $x = a$. At the point $x = a$ we know that the derivative at said point must be 0.

Example

Let $P(x) = x^4 + 2x^3 - 1$. Then $P'(x) = 4x^3 + 6x^2 = 0$ if and only if $x = x_1 = 0$ (double root) or $x = x_2 = -3/2$. The points x_1 and x_2 are either a maximum, minimum or a saddle point of the function $P(x)$. Studying the plot we see that x_1 is a saddle point and x_2 is a minimum.



We can find the minimum iteratively by using the method of gradient descent. For integers $n \geq 1$ let x_0 be a starting point

$$x_{n+1} = x_n - \alpha f'(x_n),$$

where α is called the learning rate.

Gradient descent example 1D

Let $f(x) = x^2$. This function has its (unique) minimum at $x = 0$.

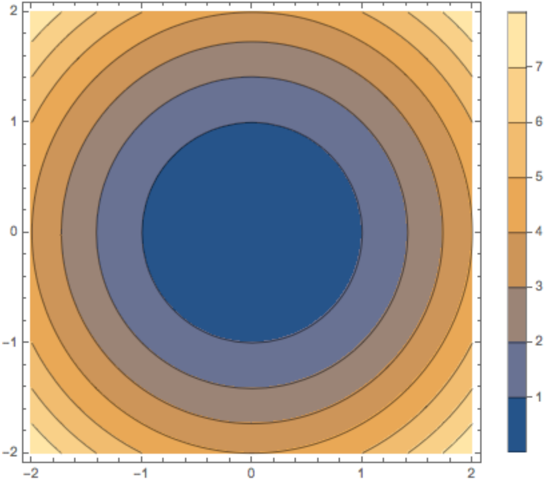
We will illustrate the gradient descent method by choosing an initial $x_0 := 5$, and we will use a learning rate $\alpha := 0.1$.

The first few steps yields

$$\begin{aligned} x_1 &= 5 - 0.1 \cdot f'(5) = 5 - 0.1 \cdot 10 = 4 \\ x_2 &= 4 - 0.1 \cdot f'(4) = 3.2, \\ x_3 &= 3.2 - 0.1 \cdot f'(3.2) = 2.56. \\ &\dots \\ x_{100} &\approx 1.09 \cdot 10^{-9}. \end{aligned}$$

The multivariable case of gradient descent involves the gradient instead of the derivative.

Gradient descent brief example 2D



Some notions from statistics

Let $x = (x_1, \ldots, x_n)$.

- Empirical mean

  $$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

  In Python (numpy): `numpy.mean(x)`

- Empirical variance

  $$\mathtt{var}(x) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2.$$

  We reduce the denominator by 1 to obtain an unbiased estimate.
  In Python (numpy): `numpy.var(x)`

Some notions from statistics

- Empirical standard deviation

  $$\mathtt{std}(x) = \sqrt{\mathtt{var}(x)}$$

  Standard deviation will be used to standardize data. For instance, if $x' = \frac{x-\bar{x}}{\mathtt{std}(x)}$ then x' is a standardization of x.
  In Python (numpy): `numpy.std(x)`
- Empirical covariance

  $$\mathsf{cov}(x, y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}).$$

  In Python (numpy): `numpy.cov(x,y)` returns the matrix

  $$\begin{pmatrix} \mathsf{cov}(x, x) & \mathsf{cov}(x, y) \\ \mathsf{cov}(y, x) & \mathsf{cov}(y, y) \end{pmatrix}.$$

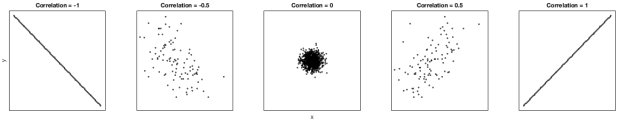
  Note that $\mathsf{cov}(x, x) = \mathtt{var}(x)$.

Some notions from statistics

- Correlation is a measure of linear relation defined by

$$\mathrm{corr}(x,y) = \frac{\mathrm{cov}(x,y)}{\mathrm{std}(x)\,\mathrm{std}(y)}$$

In Python (numpy): `numpy.corrcoef(x,y)`



Note that: $\left|\mathrm{corr}(x,y)\right| = 1 \implies x = ay + b$

Beware: $x = ay + b$ ***only*** **implies correlation** ***not*** **causation!**

`http://www.tylervigen.com/spurious-correlations`

Example

We measure the height of 8 students divided into two classes.

height class 1 (x)	height class 2 (y)
180	199
179	210
182	150
179	161

The mean of the students' heights in the two classes are the same $\bar{x} = \bar{y} = 180$, but the standard deviation – then spread of the data – differs substantially. We have

$$\mathrm{std}(x) = \sqrt{2}, \qquad \mathrm{std}(y) \approx 28.$$

Some probability theory

Definition

A sample space (often denoted Ω) is a set of all possible outcomes of a random experiment.

Example

- The roll of a dice: $\Omega = \{1, 2, 3, 4, 5, 6\}$
- The toss of a coin: $\Omega = \{H, T\}$
- A persons length at the age of 7: $\Omega = \mathbb{R}^+$.

Definition

An event of a sample space is a subset $A \subset \Omega$

Example

- An dice throw which is less than or equal to 3: $A = \{1, 2, 3\}$
- Seven-year-olds longer than 120 cm: $A = [120, \infty)$

Some probability theory

Classical probability is defined as

> *The probability of an event is the ratio of the number of cases favorable to it, to the number of all cases possible when nothing leads us to expect that any one of these cases should occur more than any other, which renders them, for us, equally possible.*

- What is the probability that a throw of a dice is six?
- What is the probability that a throw of a dice is either four or six?
- What is the probability that a throw of a dice is either odd or larger than or equal to 5?

Assume that you have two dices both with six sides, one red dice with the sides $\{1, 2, 3, 4, 5, 6\}$ and one blue with sides $\{1, 2, 2, 4, 4, 3\}$.

- What is the probability of 3 if the dice is blue?
- What is the probability of a blue dice if the side shows 2?

Some probability theory

- Sum rule: The probability that two mutually exclusive events x or y occur is the sum of its probabilities

$$P(X = x \text{ or } Y = y) = P(x) + P(y).$$

- Product rule: The probability that both event x and y occur is the product of the conditional probability of x given y and the probability of y

$$P(X = x \text{ and } Y = y) = P(x|y)P(y).$$

  In the case that these probabilities are independent, then $P(X = x \text{ and } Y = y) = P(X = x)P(Y = y)$.

- Bayes' rule: The probability that x occurs given that y have occurred

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

Some probability theory

- What is the probability that a throw of a dice is six? $P(X = 6) = 1/6$

Some probability theory

- What is the probability that a throw of a dice is six? $P(X = 6) = 1/6$
- What is the probability that a throw of a dice is either four or six? $P(X = 4 \vee X = 6) = 1/6 + 1/6 = 1/3$

Some probability theory

- What is the probability that a throw of a dice is six? $P(X = 6) = 1/6$
- What is the probability that a throw of a dice is either four or six? $P(X = 4 \vee X = 6) = 1/6 + 1/6 = 1/3$
- What is the probability that a throw of a dice is either odd or larger than or equal to 5? No longer mutually exclusive hence equals $P(X = \text{odd}) + P(X \geq 5) - P(X = \text{odd and larger than } 5) = 1/2 + 1/3 - 1/6 = 2/3$.

Some probability theory

- What is the probability that a throw of a dice is six? $P(X = 6) = 1/6$
- What is the probability that a throw of a dice is either four or six? $P(X = 4 \lor X = 6) = 1/6 + 1/6 = 1/3$
- What is the probability that a throw of a dice is either odd or larger than or equal to 5? No longer mutually exclusive hence equals $P(X = \text{odd}) + P(X \geq 5) - P(X = \text{odd and larger than } 5) = 1/2 + 1/3 - 1/6 = 2/3.$

Some probability theory

- What is the probability that a throw of a dice is six? $P(X = 6) = 1/6$
- What is the probability that a throw of a dice is either four or six? $P(X = 4 \vee X = 6) = 1/6 + 1/6 = 1/3$
- What is the probability that a throw of a dice is either odd or larger than or equal to 5? No longer mutually exclusive hence equals $P(X = \text{odd}) + P(X \geq 5) - P(X = \text{odd and larger than } 5) = 1/2 + 1/3 - 1/6 = 2/3$.

Assume that you have two dices both with six sides, one red dice with the sides $\{1, 2, 3, 4, 5, 6\}$ and one blue with sides $\{1, 2, 2, 4, 4, 3\}$.

- What is the probability of 3 if the dice is blue? $P(X = 3|\text{blue}) = 1/6$

Some probability theory

- What is the probability that a throw of a dice is six? $P(X = 6) = 1/6$
- What is the probability that a throw of a dice is either four or six? $P(X = 4 \vee X = 6) = 1/6 + 1/6 = 1/3$
- What is the probability that a throw of a dice is either odd or larger than or equal to 5? No longer mutually exclusive hence equals $P(X = \text{odd}) + P(X \geq 5) - P(X = \text{odd and larger than } 5) = 1/2 + 1/3 - 1/6 = 2/3$.

Assume that you have two dices both with six sides, one red dice with the sides $\{1, 2, 3, 4, 5, 6\}$ and one blue with sides $\{1, 2, 2, 4, 4, 3\}$.

- What is the probability of 3 if the dice is blue? $P(X = 3|\text{blue}) = 1/6$
- What is the probability of a blue dice if the side shows 2?

Some probability theory

- What is the probability that a throw of a dice is six? $P(X = 6) = 1/6$
- What is the probability that a throw of a dice is either four or six? $P(X = 4 \vee X = 6) = 1/6 + 1/6 = 1/3$
- What is the probability that a throw of a dice is either odd or larger than or equal to 5? No longer mutually exclusive hence equals $P(X = \text{odd}) + P(X \geq 5) - P(X = \text{odd and larger than } 5) = 1/2 + 1/3 - 1/6 = 2/3$.

Assume that you have two dices both with six sides, one red dice with the sides $\{1, 2, 3, 4, 5, 6\}$ and one blue with sides $\{1, 2, 2, 4, 4, 3\}$.

- What is the probability of 3 if the dice is blue? $P(X = 3|\text{blue}) = 1/6$
- What is the probability of a blue dice if the side shows 2?

Some probability theory

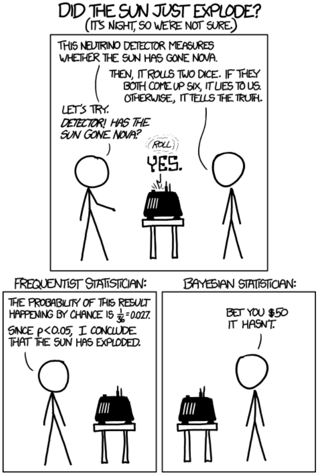
- What is the probability that a throw of a dice is six? $P(X = 6) = 1/6$
- What is the probability that a throw of a dice is either four or six? $P(X = 4 \vee X = 6) = 1/6 + 1/6 = 1/3$
- What is the probability that a throw of a dice is either odd or larger than or equal to 5? No longer mutually exclusive hence equals $P(X = \text{odd}) + P(X \geq 5) - P(X = \text{odd and larger than } 5) = 1/2 + 1/3 - 1/6 = 2/3$.

Assume that you have two dices both with six sides, one red dice with the sides $\{1, 2, 3, 4, 5, 6\}$ and one blue with sides $\{1, 2, 2, 4, 4, 3\}$.

- What is the probability of 3 if the dice is blue? $P(X = 3|\text{blue}) = 1/6$
- What is the probability of a blue dice if the side shows 2?

$$P(\text{blue}|2) = \frac{P(2|\text{blue})P(\text{blue})}{P(2)} = \frac{2/6 \cdot 1/2}{3/12} = \frac{2/12}{3/12} = \frac{2}{3}.$$

Should we be worried?


DID THE SUN JUST EXPLODE?
(IT'S NIGHT, SO WE'RE NOT SURE.)
THIS NEUTRINO DETECTOR MEASURES WHETHER THE SUN HAS GONE NOVA.
THEN, IT ROLLS TWO DICE. IF THEY BOTH COME UP SIX, IT LIES TO US. OTHERWISE, IT TELLS THE TRUTH.
LET'S TRY.
DETECTOR! HAS THE SUN GONE NOVA?
ROLL
YES.
FREQUENTIST STATISTICIAN:
THE PROBABILITY OF THIS RESULT HAPPENING BY CHANCE IS $\frac{1}{36}=0.027$. SINCE $p<0.05$, I CONCLUDE THAT THE SUN HAS EXPLODED.
BAYESIAN STATISTICIAN:
BET YOU $50 IT HASN'T.

Should you be worried?

A medical test for a given disease
The test correctly identifies the disease 99% of the time
The test incorrectly turns out positive 2% of the time

Given that
1% of the population suffers from the disease.

Let D and p denote the disease and positive answer respectively

$$P(D|p) = \frac{P(p|D)P(D)}{P(p)} = \frac{P(p|D)P(D)}{P(p|D)P(D) + P(p|\text{not}\,D)P(\text{not}\,D)} =$$

$$= \frac{0.99 \cdot 0.01}{0.99 \cdot 0.01 + 0.02 \cdot 0.99} = 1/3.$$

Bayes classifier

Based on the idea of Bayes' theorem we can construct the a so-called *Bayes classifier*.

The idea is to compute the probability $P(C_k|x_1, \dots, x_n)$, *i.e.* the probability of a data point $\mathbf{x} = (x_1, \dots, x_n)$ belonging to a class C_k.

In general these probabilities are intractable to compute, but using Bayes' theorem and some further assumptions it can be used to create a valid classifier.

Some probability theory

A random variable X denotes a quantity that is uncertain, e.g. the roll of a dice, or the exact time it takes to get to work.

A probability distribution $p(X)$ of a random variable X gives information about the possible outcomes of a X, and also which values are expected to occur more often than others, i.e. are assigned a higher probability.

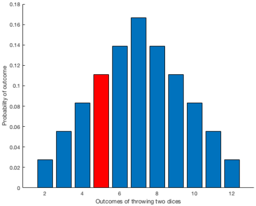
There are discrete and continuous probability distributions.

[Whiteboard roll of a dice, two dices]

Some probability theory

Discrete case: probability mass function

$$P(X = 5)$$



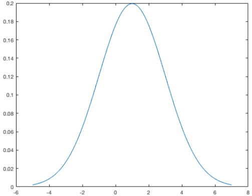
We note that

$$\sum_{i \in \text{all possible events}} P(X = i) = 1.$$

Some probability theory

Continuous case: probability density function

$$P(a \leq X \leq b) = \int_a^b p(X)dX .$$



We note that

$$\int_{-\infty}^{+\infty} p(X)dX = 1.$$

Some probability theory

The expected value of a random variable can be thought of as the average of the random variable. The expected values of a random variable can be computed as

- Discrete random variable

$$\mathbb{E}[X] = \sum_i x_i P(X = x_i) := \mu$$

- Continuous random variable

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

Example

The expected outcome of throwing a dice is given by

$$1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = \frac{21}{6} = 3.5.$$

Some probability theory

Covariance of two random variables X and Y

$$\mathrm{cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$$

Variance of a random variable

$$\mathrm{var}(X) = \mathrm{cov}(X, X) = \mathbb{E}[(X - \mathbb{E}[X])^2] := \sigma^2$$

Note: we can estimate μ and σ for a dataset x by their empirical counterparts, i.e. $\hat{\mu} = \bar{x}$ and $\hat{\sigma}^2 = \mathrm{var}(x)$

Gaussian distribution

A continuous random variable is said to be a Gaussian random variable if its probability density function is of the form

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad x, \mu \in \mathbb{R}, \quad \sigma > 0.$$

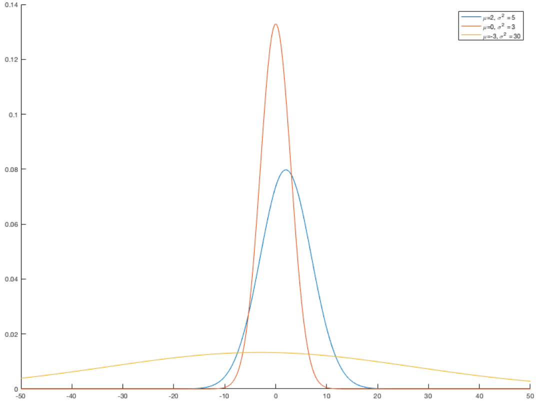
We denote this by $X \sim N(\mu, \sigma^2)$.

The Gaussian distribution is also known as the normal distribution.

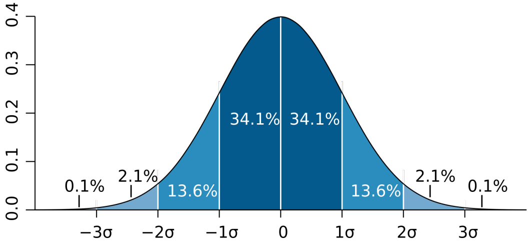
The Gaussian distribution has a prominent position in the literature as it fits many natural phenomena.

Gaussian distribution

The probability density function for three Gaussian distributions with different parameters



Gaussian distribution


0.4
0.3
0.2
0.1
0.0
0.1%
2.1%
13.6%
34.1%
34.1%
13.6%
2.1%
0.1%
−3σ
−2σ
−1σ
0
1σ
2σ
3σ