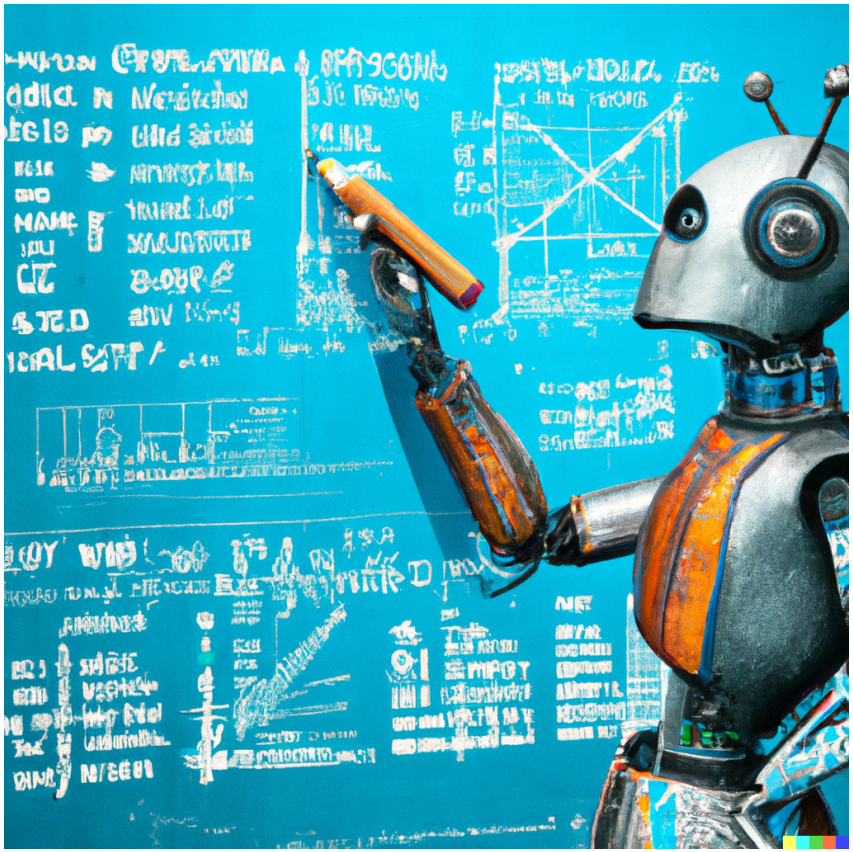
Mathematics and Statistics for Machine Learning

Elizaveta Semenova,
University of Oxford

Deep Learning Indaba
Tunis, 2022

Outline

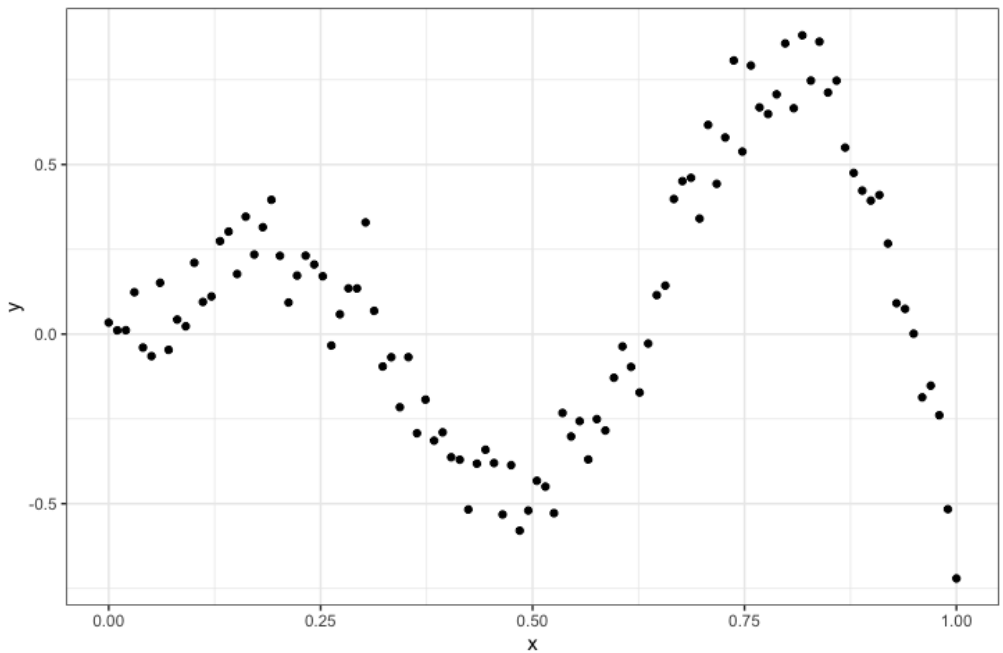
Modern applications of machine learning



"Robot mixing statistics, calculus and linear algebra, digital art", DALLE-2

Motivation: curve fitting

- Problem setting: given observed data pairs $(x_i, y_i), i = 1, ..., n$, learn to predict y from x.



Motivation: curve fitting

- Problem setting: given observed data pairs $(x_i, y_i), i = 1, ..., n$, learn to predict y from x.
- Model f, depending on parameters θ: $y = f(x, \theta)$.

Motivation: curve fitting

- Problem setting: given observed data pairs $(x_i, y_i), i = 1, ..., n$, learn to predict y from x.
- Model f, depending on parameters θ: $y = f(x, \theta)$.

linear model	$f(x, \theta) = \theta^T x$	parametric, linear
neural network	$f(x, \theta) = NN_\theta(x, \theta)x$	parametric, non-linear
Gaussian process	$f(x, \theta) = GP_\theta(x)$	non-parametric

Motivation: curve fitting

- Problem setting: given observed data pairs $(x_i, y_i), i = 1, ..., n$, learn to predict y from x.
- Model f, depending on parameters θ: $y = f(x, \theta)$.
- Training a model, means finding parameters θ^*, such that $f(x_i, \theta^*) \approx y_i$, e.g.:

$$\sum_{i=1}^{n} (f(x_i, \theta^*) - y_i)^2 \rightarrow \min_{\theta}$$

Motivation: curve fitting

$$\underbrace{\sum_{i=1}^{n} (f(x_i, \theta^*) - y_i)^2}_{\mathcal{L}(x,y,\theta) \text{ - loss function}} \to \min_{\theta}$$

Mathematical foundations of machine learning

- How to derive an appropriate loss function for each problem?
  $\rightarrow$ Statistics
- How to optimise a loss functions?
  $\rightarrow$ Calculus
- How to express computations efficiently?
  $\rightarrow$ Linear algebra

Mathematical fields that Machine Learning builds on

- Probability theory and Statistics,
- Calculus,
- Linear algebra.

Outline

Systems of linear equations

Linear algebra studies vectors, vector spaces and mapping between vector spaces. It emerged from the study of systems of linear equations.

$$\begin{cases} 2x_1 + 4x_2 = 5, \\ 10x_1 + 3x_2 = 14. \end{cases}$$

We can rewrite the system as

$$\underbrace{\begin{pmatrix} 2 & 4 \\ 10 & 3 \end{pmatrix}}_{A} \underbrace{\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}}_{x} = \underbrace{\begin{pmatrix} 5 \\ 14 \end{pmatrix}}_{b}.$$

$Ax = b$ gives a compact representation of the linear system.

Vectors over real numbers

$$v = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix} \in \mathbb{R}^{n \times 1}$$

Dot product

For two vectors

$$v = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix} \in \mathbb{R}^{n \times 1}, \quad w = \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{pmatrix} \in \mathbb{R}^{n \times 1}$$

we define the dot product as

$$v \cdot w = \sum_{i=1}^{n} v_i w_i.$$

Vector norms

A norm of a vector $x \in \mathbb{R}^n$ is a function, mapping the vector into the set of real numbers

$$||\cdot|| : x \to \mathbb{R}.$$

It allows to measure the length of a vector $||x|| \in \mathbb{R}$ and satisfies the following conditions for all $\lambda \in \mathbb{R}$ and vectors $x, y \in \mathbb{R}^n$:

1. $||\lambda x|| = |\lambda| ||x||$ - homogeneity,
2. $||x + y|| \leq ||x|| + ||y||$ - triangle inequality,
3. $||x|| \geq 0$, and $||x|| = 0 \iff x = 0$ - positive definiteness.

Examples of vector norms

- The l_1-norm ("Manhattan" or the "Taxicab" norm):

$$||x||_{l_1} = \sum_{i=1}^{n} |x_i|,$$

Examples of vector norms

- The l_1-norm ("Manhattan" or the "Taxicab" norm):

$$||x||_{l_1} = \sum_{i=1}^{n} |x_i|,$$

- The l_2-norm ("Euclidean" norm):

$$||x||_{l_2} = \sqrt{\sum_{i=1}^{n} x_i^2},$$

Examples of vector norms

- The l_1-norm ("Manhattan" or the "Taxicab" norm):

$$||x||_{l_1} = \sum_{i=1}^{n} |x_i|,$$

- The l_2-norm ("Euclidean" norm):

$$||x||_{l_2} = \sqrt{\sum_{i=1}^{n} x_i^2},$$

- The l_p-norm:

$$||x||_{l_p} = \left(\sum_{i=1}^{n} x_i^p \right)^{1/p},$$

Examples of vector norms

- The l_1-norm ("Manhattan" or the "Taxicab" norm):

$$||x||_{l_1} = \sum_{i=1}^{n} |x_i|,$$

- The l_2-norm ("Euclidean" norm):

$$||x||_{l_2} = \sqrt{\sum_{i=1}^{n} x_i^2},$$

- The l_p-norm:

$$||x||_{l_p} = \left(\sum_{i=1}^{n} x_i^p \right)^{1/p},$$

- The l_∞-norm:

$$||x||_{l_\infty} = \max_{1 \leq i \leq n} |x_i|.$$

Loss function as a norm

$$e_i := y_i - f(x_i, \theta^*) - \text{error at observation } i,$$

$$\begin{aligned}\mathcal{L}(x, y, \theta) &= \sum_{i=1}^{n} (y_i - f(x_i, \theta^*))^2 \\ &= ||y - f(x, \theta)||_{l_2} \\ &= ||e||_{l_2}\end{aligned}$$

Matrices over real numbers

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1m} \\ a_{21} & a_{22} & a_{23} & \dots & a_{2m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \dots & a_{nm} \end{pmatrix} \in \mathbb{R}^{n \times m}$$

Identity matrix

$$I_n = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{pmatrix} \in \mathbb{R}^{n \times n}$$

Matrix summation

For matrices $A \in \mathbb{R}^{n \times m}$ and $B \in \mathbb{R}^{n \times m}$ the sum is defined as

$$A + B = \begin{pmatrix} a_{11} + b_{11} & a_{12} + b_{12} & \dots & a_{1m} + b_{1m} \\ a_{21} + b_{21} & a_{22} + b_{22} & \dots & a_{2m} + b_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} + b_{n1} & a_{n2} + b_{n2} & \dots & a_{nm} + b_{nm} \end{pmatrix}$$

Matrix multiplication

For matrices $A \in \mathbb{R}^{nxk}$ and $B \in \mathbb{R}^{kxm}$ the product $C \in \mathbb{R}^{nxm}$ can be calculated

$$AB = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1k} \\ a_{21} & a_{22} & \dots & a_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nk} \end{pmatrix} \begin{pmatrix} b_{11} & b_{12} & b_{13} & \dots & b_{1m} \\ b_{21} & b_{22} & b_{23} & \dots & b_{2m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ b_{k1} & b_{k2} & b_{k3} & \dots & b_{km} \end{pmatrix} = C$$

Matrix multiplication

$$c_{ij} = \underbrace{\begin{pmatrix} a_{i1} & a_{i2} & \dots & a_{ik} \end{pmatrix} \begin{pmatrix} b_{1j} \\ b_{2j} \\ \vdots \\ b_{km} \end{pmatrix}}_{\text{dot product } a_{i,.} \cdot b_{.,j}} = \sum_{p=1}^{k} a_{ip} b_{pj}$$

Matrix transpose

For a matrix $A \in \mathbb{R}^{nxm}$ its transposed $A^T \in \mathbb{R}^{mxn}$ is found by reflecting the elements with respect to the main diagonal: $(A^T)_{ij} = A_{ji}$, i.e.

$$A^T = \begin{pmatrix} a_{11} & a_{21} & a_{31} & \dots & a_{m1} \\ a_{12} & a_{22} & a_{32} & \dots & a_{m2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{1n} & a_{n2} & a_{3n} & \dots & a_{mn} \end{pmatrix} \in \mathbb{R}^{mxn}$$

Matrix inverse

For a square matrix $A \in \mathbb{R}^{n \times n}$, matrix $A^{-1} \in \mathbb{R}^{n \times n}$ gives its inverse if

$$AB = BA = I_n.$$

Some important properties

Addition:

$$A + B = B + A,$$
$$(A + B)^T = A^T + B^T,$$
$$(A + B)^{-1} \neq A^{-1} + B^{-1},$$

Multiplication:

$$AB \neq BA,$$
$$(AB)^T = B^T A^T,$$
$$(AB)^{-1} = B^{-1} A^{-1}.$$

Outline

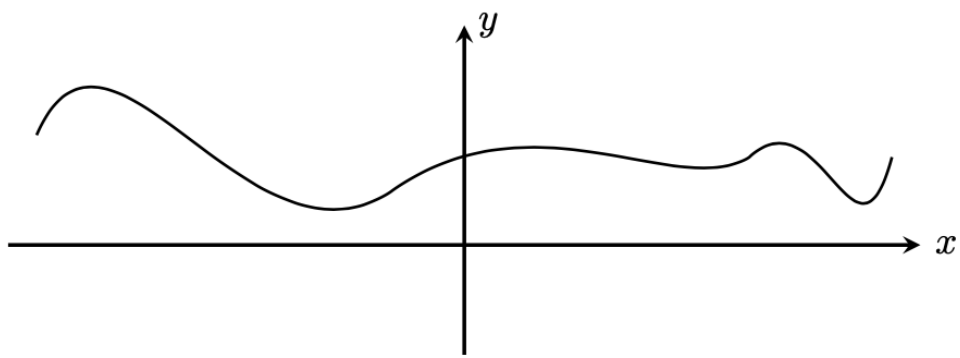
Univariate calculus

$$f : \mathbb{R} \to \mathbb{R}$$

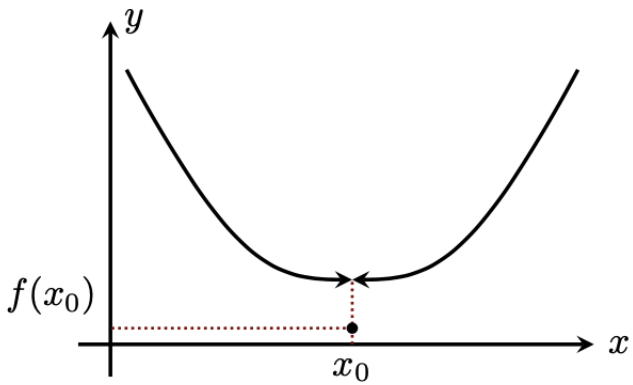
Continuous functions

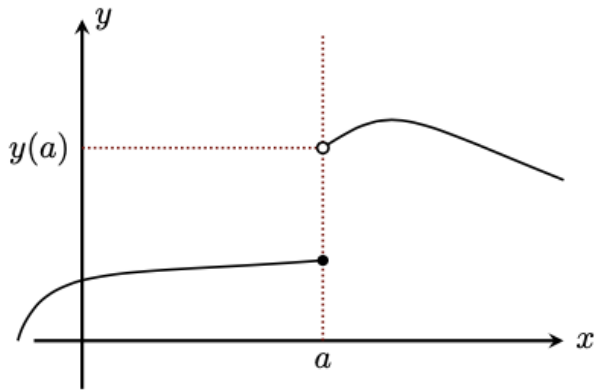
A function is called continuous if small changes in the argument x lead to small changed in the function values y (an informal definition).

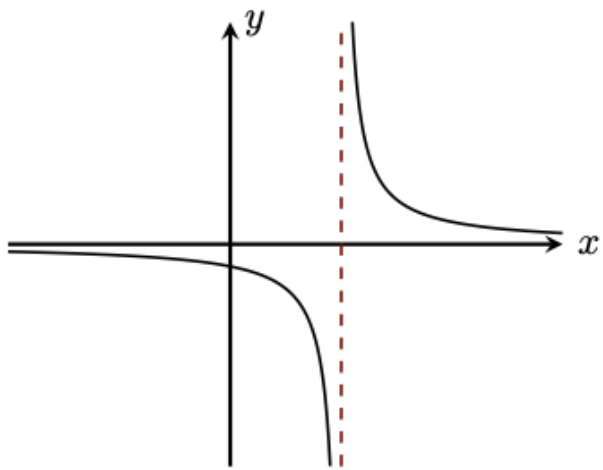
$$f : \mathbb{R} \to \mathbb{R},$$
$$y = f(x)$$



Discontinuity types







Continuous functions and limits

A function f is called continuous at a point x_0 if

$$\lim_{x \to x_0} f(x) = f(\lim_{x \to x_0} x) = f(x_0).$$

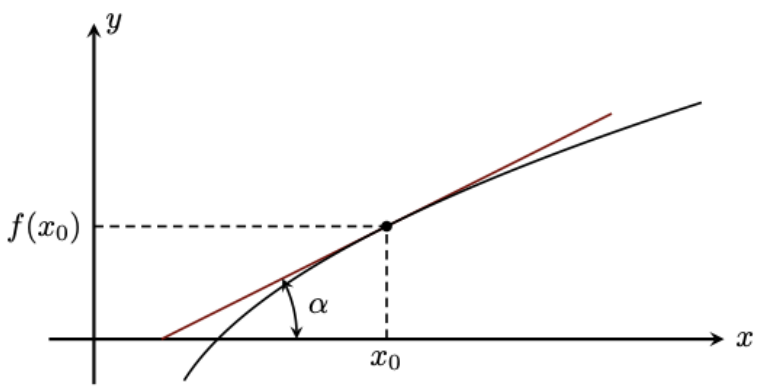
Differentiation

Derivative of a continuous function f is defined as

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

Geometrical interpretation of derivative

The derivative is equal to the slope (or gradient) of the tangent line to the curve $y = f(x)$.



Differentiation rules

Denote $u = u(x), v = v(x)$. Then

Sum:	$(u+v)' = u'(x) + v'(x),$
Product:	$(uv)' = u'v + v'u,$
Quotient:	$\left(\frac{u}{v}\right)' = \frac{u'v - v'u}{v)^2},$
Chain:	$(u \circ v)' = v'u'(v).$

Examples

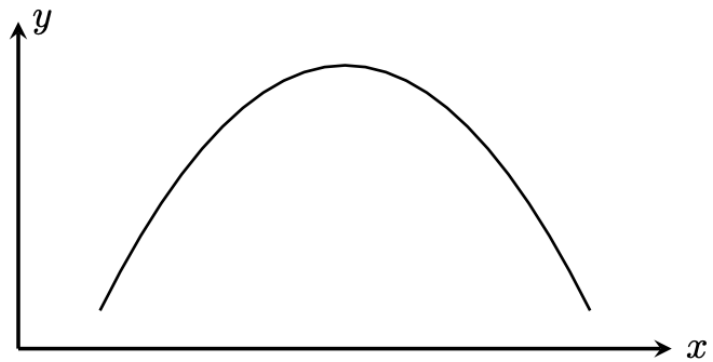
$$
\begin{aligned}
f(x) &= x^a, & f'(x) &= a x^{a-1}, \\
f(x) &= e^x, & f'(x) &= e^x, \\
f(x) &= \ln(x), & f'(x) &= \frac{1}{x}, \\
f(x) &= \sin(x), & f'(x) &= \cos(x), \\
f(x) &= \cos(x), & f'(x) &= -\sin(x), \\
f(x) &= \tanh(x), & f'(x) &= 1 - \tanh^2(x).
\end{aligned}
$$

Exercise

Calculate derivatives of the following functions:

$$f(x) = \sin(3x) + \frac{1}{x},$$
$$f(x) = x\sin(3x),$$
$$f(x) = \frac{\sin(3x)}{x},$$
$$f(x) = \sin\left(\frac{3}{x}\right).$$

Condition for an extremum



$f'(x) > 0 \Rightarrow$ the function is increasing,

$f'(x) < 0 \Rightarrow$ the function is decreasing,

$f'(x) = 0 \Rightarrow$ sufficient and necessary condition for an extremum.

Linear regression example

- Assume that b is known, and $\theta = a$ is unknown:

$$y = ax + b,$$
$$\mathcal{L}(x, y, \theta) = \sum_{i=1}^{n} (y_i - ax_i - b)^2,$$
$$\theta^* = \text{argmin}_a \mathcal{L}(x, y, a).$$

Then

$$\frac{d\mathcal{L}}{d\theta} = \frac{d\mathcal{L}}{da} = 0$$

will provide the solution.

Linear regression example

- Assume that b is known, and $\theta = a$ is unknown:

$$y = ax + b,$$
$$\mathcal{L}(x, y, \theta) = \sum_{i=1}^{n} (y_i - ax_i - b)^2,$$
$$\theta^* = \text{argmin}_a \mathcal{L}(x, y, a).$$

  Then
$$\frac{d\mathcal{L}}{d\theta} = \frac{d\mathcal{L}}{da} = 0$$
  will provide the solution.

- What if both a and b are unknown, i.e. $\theta = \begin{pmatrix} a \\ b \end{pmatrix}$?

Multivariate calculus

$$f : \mathbb{R}^n \to \mathbb{R}$$

Partial differentiation

For a function of variables $x_1, ..., x_n$ partial derivatives are defined as

$$\frac{\partial f}{\partial x_1} = \lim_{h \to 0} \frac{f(x_1 + h, x2, ..., x_n) - f(x_1, ..., x_n)}{h},$$

$$\frac{\partial f}{\partial x_2} = \lim_{h \to 0} \frac{f(x1, x_2 + h, x3, ..., x_n) - f(x_1, ..., x_n)}{h},$$

...

$$\frac{\partial f}{\partial x_n} = \lim_{h \to 0} \frac{f(x1, ..., x_n + h) - f(x_1, ..., x_n)}{h}.$$

Gradient

We collect partial derivatives into a row-vector[1], called gradient:

$$\nabla f(x) = \frac{df}{dx} = \left[\frac{\partial f(x)}{\partial x_1}, ..., \frac{\partial f(x)}{\partial x_n}\right] \in \mathbb{R}^{1xn}$$

[1] This convention differs, sometimes gradient is being composed as a column-vector

Exercise

Find gradient of the function

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

Differentiation rules apply

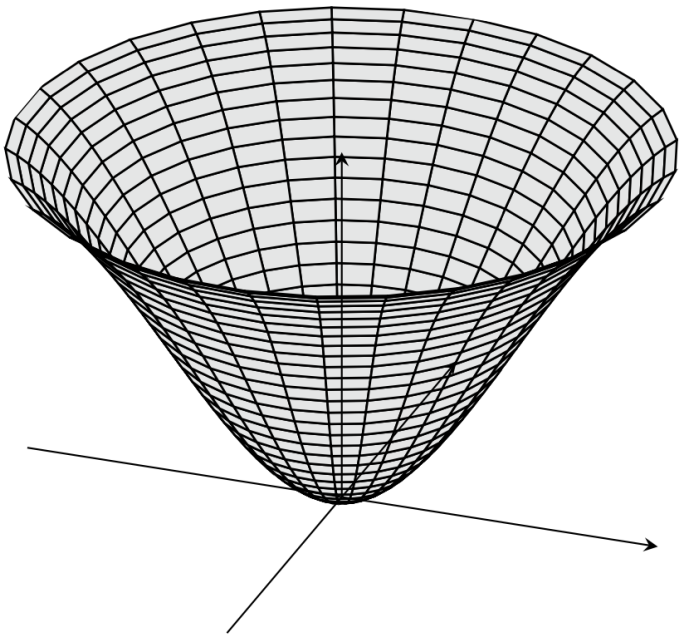
$$\text{Sum:} \quad \frac{\partial(u+v)}{\partial x} = \frac{\partial u}{\partial x} + \frac{\partial v}{\partial x},$$

$$\text{Product:} \quad \frac{\partial(uv)}{\partial x} = v\frac{\partial u}{\partial x} + u\frac{\partial v}{\partial x},$$

$$\text{Quotient:} \quad \frac{\partial}{\partial x}\left(\frac{u}{v}\right) = \frac{v\frac{\partial u}{\partial x} - u\frac{\partial v}{\partial x}}{v^2},$$

$$\text{Chain:} \quad \frac{\partial}{\partial x}(u \circ v) = \frac{\partial u}{\partial v}\frac{\partial v}{\partial x}.$$

Sufficient condition for an extremum



$$\nabla f(x) = 0.$$

Multivariate calculus

$$f : \mathbb{R}^n \to \mathbb{R}^m$$

Multivariate calculus

- $f : \mathbb{R}^n \to \mathbb{R}^m$
  $y = f(x), \quad x \in \mathbb{R}^n, \quad y \in \mathbb{R}^m$

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix} = \begin{pmatrix} f_1(x) \\ f_2(x) \\ \vdots \\ f_m(x) \end{pmatrix} = \begin{pmatrix} f_1(x_1, \ldots, x_n) \\ f_2(x_1, \ldots, x_n) \\ \vdots \\ f_m(x_1, \ldots, x_n) \end{pmatrix}$$

- Jacobian matrix (collection of all partial derivatives):

$$\nabla f(x) = \begin{pmatrix} \frac{dy_1}{dx} \\ \frac{dy_2}{dx} \\ \vdots \\ \frac{dy_m}{dx} \end{pmatrix} = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{pmatrix} \in \mathbb{R}^{m \times n}$$

Integration

$$\int f(x)dx$$

Riemann integral

- $f(x)$ is a function defined on the interval $[a, b]$.
- We partition $[a, b]$ into n intervals of equal width $\Delta x = \frac{b-a}{n}$:

$$[x_0, x_1], ..., [x_{n-1}, x_n],$$

  where $x_0 = a, x_n = b$.
- Denote by x_i^* any sample point in the interval $[x_{i-1}, x_i]$.
- The Riemann definite integral is then defined as

$$\int_a^b f(x)dx = \lim_{n\to\infty} \sum_{i=1}^{n} f(x_i^*)\Delta x$$

Fundamental theorem of calculus

- Given any function $f(x)$ on $[a, b]$, an antiderivative is any function $F(x)$ on $[a, b]$ such that $F'(x) = f(x)$.

- Assume that f is continuous on the interval $[a, b]$, then the function

$$F(x) = \int_a^x f(t)dt, \quad a \leq x \leq b$$

is differentiable on (a, b) and

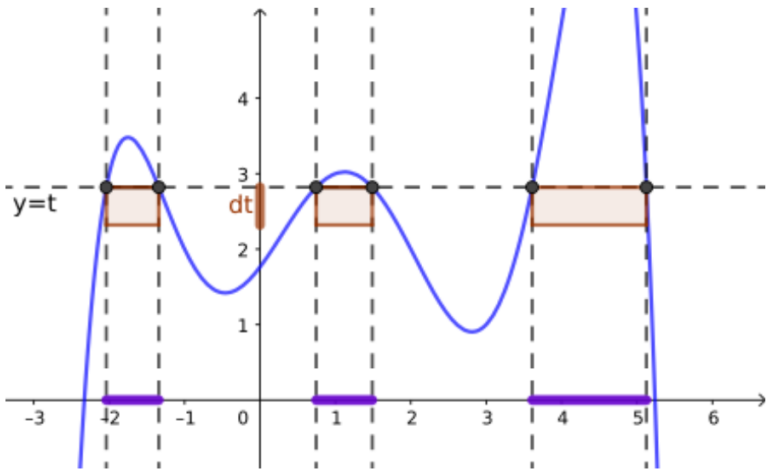
$$F'(x) = f(x),$$

i.e. $F(x)$ is an antiderivateive of $f(x)$.

Lebesgue integral

- In the definition of the Riemann integral of a function $f(x)$, the x-axis is partitioned and the integral is defined in terms of limits of the Riemann sum.

- The idea behind the Lebesgue integral is to partition the y-axis, which contains the range of f, rather than the x-axis.



Credit: Slawomir Bialy at English Wikipedia, CC BY-SA 4.0

Outline

What is probability?

A branch of mathematics that deals with calculating the likelihood of a given event to occur.

Probability theory and statistics

Probability theory studies models of random variables and their properties.

Statistics aims to estimate the properties of random variables, given finite data.

Basics of probability

What is the probability of getting a number greater than 10 when we roll a die?

$\rightarrow 0$

What is the probability of getting a number less than 10 when we roll a die?

$\rightarrow 1$

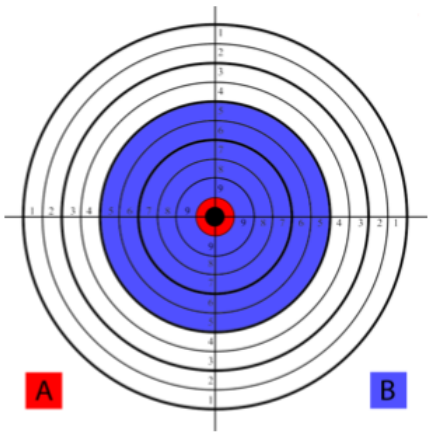
Probability of any event lies between 0 and 1: $p \in (0, 1)$.

For any event A it holds

$$\begin{aligned} 0 \leq P(A) &\leq 1, \\ P(\emptyset) &= 0, \\ P(\bar{A}) + P(A) &= 1. \end{aligned}$$

Relationship of two events

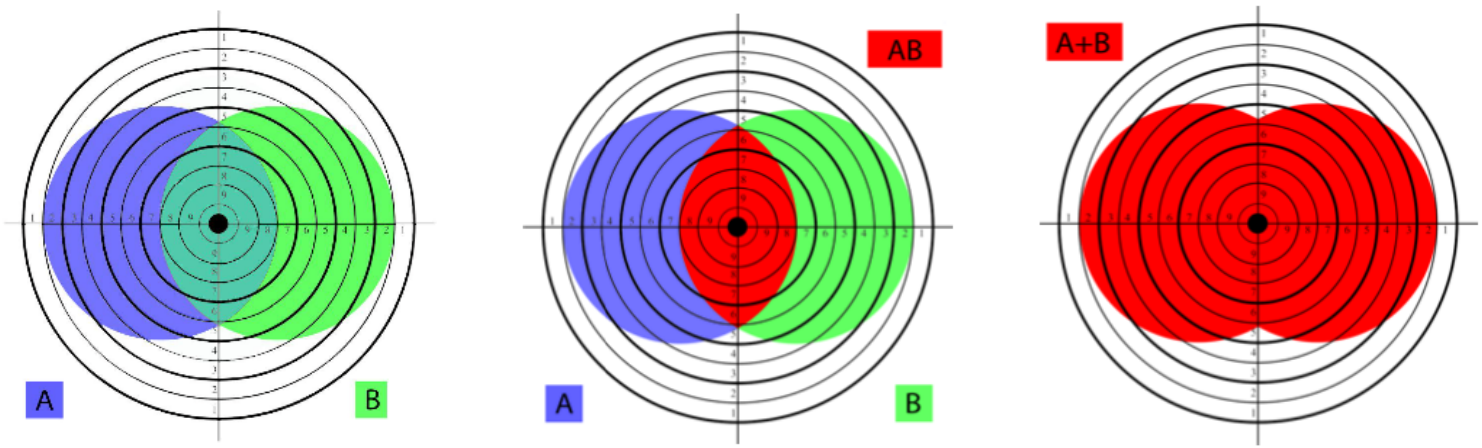
Inclusion: $A \subseteq B \Rightarrow P(A) \leq P(B)$



Relationship of two events

Product and sum (intersection and union):

$$A \cap B, \quad A \cup B$$



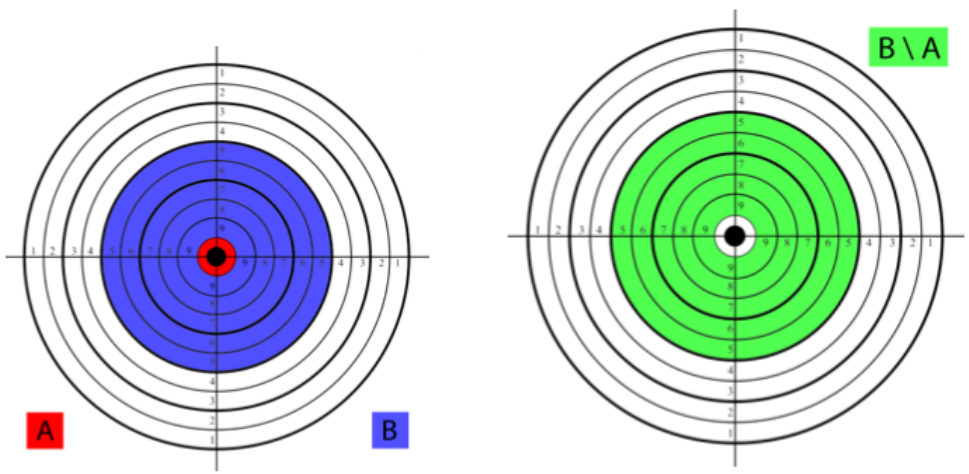
$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

Relationship of two events

Difference: $B \backslash A$

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

If $A \subseteq B$, the formula simplifies to $P(B \backslash A) = P(B) - P(A)$



Independent events

Two events A and B are independent if

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

Conditional probability

Conditional probability $P(A|B)$ is the probability of event A provided evidence B:

$$P(A|B) = \frac{P(AB)}{P(B)}$$

Bayes' formula

$$P(A|B) = \frac{P(AB)}{P(B)} \quad \Rightarrow \quad P(AB) = P(A|B)P(B),$$

$$P(B|A) = \frac{P(AB)}{P(A)} \quad \Rightarrow \quad P(AB) = P(B|A)P(A),$$

$$\Rightarrow \quad P(A|B) = \frac{P(B|A)P(A)}{P(B)}.$$

Random variables

- A random variable X is a numerical description of the outcomes of random events.
- In other words, a random variable maps the outcomes of random events to numerical values.

Random event: head or tail,
random variable: 0 or 1.

Discrete random variables

A random variable X is discrete if its possible realisations come from a countable set

$$A = \{a_1, a_2, a_3, ..., \}$$

with probabilities

$$p_1, p_2, p_3, \ldots$$

such that

$$\sum_{i=1}^{\infty} p_i = 1, \quad p_i \geq 0.$$

Then we write for the probability mass function (PMF)

$$P(X = a_i) = p_i.$$

Discrete random variables - examples

Some common discrete distributions:

- Bernoulli with success probability p:

$$X \sim Bern(p),$$
$$P(X = 1) = p,$$
$$P(X = 0) = 1 - p;$$

- Binomial describes number of successes out of n independent experiments, each with success probability p:

$$X \sim Binom(n, p),$$
$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k},$$
$$k = 0, 1, 2, ..., n.$$

Continuous random variables

A continuous random variable X can be defined via a distribution function

$$F(x) = P(X \leq x),$$

or, alternatively, via a probability density function (PDF)

$$f(x): \quad \int_a^b f(x)dx = P(a \leq X \leq b).$$

Distribution function and distribution density are closely related:

$$F(x) = \int_{-\infty}^{x} f(x)dx.$$

Continuous random variables - examples

Some common continuous distributions:

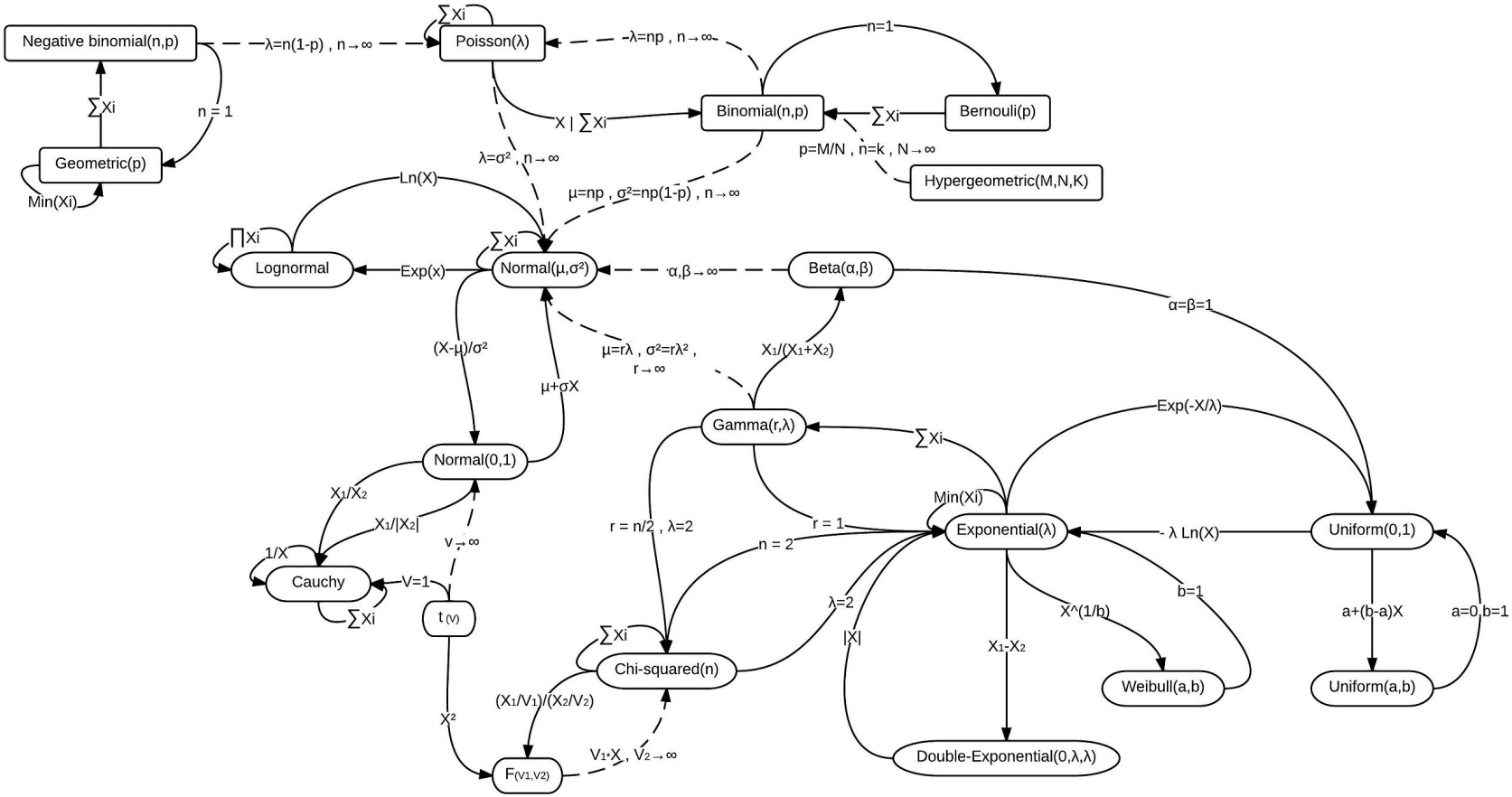
- Uniform on the interval $[a, b]$:

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

- Normal with mean μ and variance σ^2:

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

Relationships among some probability distributions



Credit: Ehsan Azhdari at English Wikipedia, CC BY-SA 3.0

Summary statistics

A statistic of a random variable is a deterministic function of the random variable, providing a useful view of how a random variable behaves.

Expected value

- Expected value of a function $g(x)$ with respect to random variable $X \sim p(x)$:

$$E[g(X)] = \int_{\mathcal{X}} g(x)p(x)dx \qquad \text{- continuous } X,$$
$$= \sum_{\mathcal{X}} g(x)p(x) \qquad \text{- discrete } X.$$

Mean and moments

- Mean of a random variable X is calculated using $g(x) = x$:

$$E[X] = \int_{\mathcal{X}} xp(x)dx \qquad \text{- continuous } X,$$
$$= \sum_{\mathcal{X}} xp(x) \qquad \text{- discrete } X.$$

- A moment of order k is the expectation of $g(x) = x^k$:

$$E[X^k] = \int_{\mathcal{X}} x^k p(x)dx \qquad \text{- continuous } X,$$
$$= \sum_{\mathcal{X}} x^k p(x) \qquad \text{- discrete } X.$$

Covariance

For two random variables, X, Y with $E[X] = \mu_x, E[Y] = \mu_y$, covariance is defined as

$$\text{Cov}[X, Y] = E\left[(x - \mu_x)(y - \mu_y)\right].$$

It can be shown that

$$\text{Cov}[X, Y] = E[XY] - \mu_x \mu_y.$$

Variance

For a random variable X with $E[X] = \mu_x$, variance is defined as

$$\begin{aligned} \text{Var}[X] &= \text{Cov}[X, X] \\ &= E\left[(x - \mu_x)^2\right] \\ &= E[x^2] - \mu_x^2. \end{aligned}$$

Summary statistics

Statistics can help us characterise relationships between random variables, such as correlation and independence.

Outline

It takes ~~two~~ three to tango.

It takes ~~two~~ three to tango.

How can statistics, linear algebra and calculus, combined, help us perform machine learning tasks?

Types of inference

- Frequentist $\rightarrow$ optimization
- Bayesian $\rightarrow$ integration

The principle of maximum likelihood

From probabilistic viewpoint, we believe that the observed data $x_1, ..., x_n$ is generated by a certain PMF or PDF

$$f(x|\theta)$$

with unknown parameters θ. If all observations are independent, we can write down the likelihood

$$L(\theta|x_1, ..., x_n) = \prod_{i=1}^{n} f(x_i|\theta).$$

Maximum likelihood estimator

The higher the likelihood, the closer observed data corresponds to the chosen model.

$$\hat{\theta}_{\text{MLE}} : L(\theta|x_1, ..., x_n) \to \max_{\theta}.$$

Equivalently,

$$\hat{\theta}_{\text{MLE}} : -\ln L(\theta|x_1, ..., x_n) \to \min_{\theta}.$$

The latter formulation is preferable as it turns the product of small numbers into the sum of not small numbers:

$$-\ln L(\theta|x_1, ..., x_n) = -\sum_{i=1}^{n} \ln f(x_i|\theta).$$

MLE and losses - Bernoulli distribution

$$
\begin{aligned}
x_i & \sim Ber(p), \quad i = 1, ..., n, \\
f(x_i|p) & = p^{x_i}(1-p)^{1-x_i}, \\
L(p|x_1, ..., x_n) & = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}, \\
l(p|x_1, ..., x_n) = \ln L(p|x_1, ..., x_n) & = \sum_{i=1}^{n} \underbrace{(x_i \ln p + (1-x_i)\ln(1-p))}_{\text{binary classification loss!}}
\end{aligned}
$$

Exercise: derive MLE estimate of parameter p.

MLE and losses - Bernoulli distribution

$$
\begin{aligned}
l(p) = \ln L(p|x_1, ..., x_n) &= \sum_{i=1}^{n} (x_i \ln p + (1 - x_i) \ln(1 - p)) \\
&= \ln p \sum_{i=1}^{n} x_i + \ln(1 - p) \sum_{i=1}^{n} (1 - x_i), \\
\frac{\partial l(p)}{\partial p} &= \frac{\sum_{i=1}^{n} x_i}{p} - \frac{\sum_{i=1}^{n} (1 - x_i)}{1 - p} \quad \Rightarrow \\
\hat{p}_{\text{MLE}} &= \frac{\sum_{i=1}^{n} x_i}{n}.
\end{aligned}
$$

MLE and losses - univariate Normal

$$
\begin{aligned}
x_i &\sim Normal(\mu, \sigma^2), \quad i = 1, ..., n, \\
f(x_i|p) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right), \\
L(\mu, \sigma^2|x_1, ..., x_n) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right), \\
l(\mu, \sigma^2|x_1, ..., x_n) &= \sum_{i=1}^{n} \left[-\ln(\sqrt{2\pi\sigma^2}) - \frac{1}{2\sigma^2}(x_i-\mu)^2\right] \\
&= -n\ln(\sqrt{2\pi\sigma^2}) - \frac{1}{2\sigma^2} \underbrace{\sum_{i=1}^{n}(x_i-\mu)^2}_{\text{MSE loss}}
\end{aligned}
$$

MLE and losses - univariate Normal

Exercise: derive MLE estimates of parameters μ and σ^2.

MLE and losses - univariate Normal

Exercise: derive MLE estimates of parameters μ and σ^2.

$$\hat{\mu}_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

$$\hat{\sigma^2}_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu}_{\text{MLE}})^2.$$

Gradients in a neural network

- Output y is computed as a many-level function composition

$$\begin{aligned} y &= (f_k \circ f_{k-1} \circ \cdots \circ f_1)(x) \\ &= f_k(f_{k-1} \ldots f_2(f_1(x))) \end{aligned}$$

- Each layer i is represented by an activation function

$$f_i(x_{i-1}) = \sigma(W_{i-1}x_{i-1} + b_{i-1})$$

- $\sigma(.)$ can be a sigmoid, tanh, ReLU or others.
- To train this model, we need to find derivatives of a loss function $L(\theta)$ with respect to all model parameters $\theta_i = (W_{i-1}, b_{i-1})$, $i = 1, \ldots k$.

Gradients in a neural network

$$
\begin{aligned}
f_i &= \sigma(f_{i-1}, \theta_{i-1}) \\
y &= f_k(f_{k-1} \dots f_{i-1} \dots f_2(f_1(x))) \\
&= f_k(f_{k-1,\theta_{k-1}} \dots f_{i-1,\theta_{i-1}} \dots f_2(f_1(x))) \\
\frac{\partial L}{\partial \theta_{k-1}} &= \frac{\partial L}{\partial f_k} \frac{\partial f_k}{\partial \theta_{k-1}}, \\
\frac{\partial L}{\partial \theta_{k-2}} &= \frac{\partial L}{\partial f_k} \frac{\partial f_k}{\partial f_{k-1}} \frac{\partial f_{k-1}}{\partial \theta_{k-2}}, \\
\frac{\partial L}{\partial \theta_{k-3}} &= \frac{\partial L}{\partial f_k} \frac{\partial f_k}{\partial f_{k-1}} \frac{\partial f_{k-1}}{\partial f_{k-2}} \frac{\partial f_{k-2}}{\partial \theta_{k-3}}.
\end{aligned}
$$

Bayesian inference

- Bayesian computation:

$$p(\theta) - \text{prior distribution},$$

$$p(y|x,\theta) = \int p(y|x,\theta)p(\theta)d\theta$$

Bayesian inference

- Advantages
  - allows to incorporate prior knowledge,
  - prevents overfitting,
  - inherently characterises uncertainty.
- Maximum *a posteriori* estimate (MAP):

$$\theta^*_{\text{MAP}} : p(\theta|y) \rightarrow \max_{\theta}.$$

- Numerical estimation
  - Markov Chain Monte Carlo algorithms,
  - Expectation propagation,
  - Variational Bayes,
  - Laplace approximation.

Thank You!

www.elizaveta-semenova.com

@liza_p_semenova

DEEP LEARNING INDABA