UCL COMP0078: Supervised Learning
Mathematical Introduction

Antonin Schrab

If you find any mistakes/typos, please contact sl-support@cs.ucl.ac.uk.

1 Notations

Sets.

- Natural numbers: $\mathbb{N} = \{0, 1, 2, 3, \dots\}$ (sometimes excluding 0 depending on context)
- Positive integers up to n: $[n] = \{1, 2 \dots, n-1, n\}$
- Integers: $\mathbb{Z} = \{\dots, -3, -2, -1, 0, 1, 2, 3, \dots\}$
- Rational numbers: $\mathbb{Q} = \{a/b : a, b \in \mathbb{Z}, b \neq 0\}$
- Real numbers: $\mathbb{R}$ includes both rational (*e.g.* fractions) and irrational numbers (*e.g.* $\pi, e, \dots$)
- Complex numbers: $\mathbb{C} = \{a + ib : a, b \in \mathbb{R}\}$ where i is an element satisfying $i^2 = -1$

For a set S, the notation $x \in S$ means the element x is part of the set S.

For sets $S_1, \dots, S_n$, the notation $(x_1, \dots, x_n) \in S_1 \times \dots \times S_n$ means that $x_i \in S_i$ for $i \in [n]$.

If all elements of a set S' also in the set S, we say that S' is a subset of S and write $S' \subseteq S$. If in addition $S' \neq S$, we write $S' \subset S$.

We write $x \in S \cup S'$ (called union) if $x \in S$ or $x \in S'$.

We write $x \in S \cap S'$ (called intersection) if $x \in S$ and $x \in S'$.

We say that u is an upper bound of S if $x \leq u$ for all $x \in S$. The smallest upper bound of S is called the supremum and is denoted $\sup S$.

We say that ℓ is an lower bound of S if $x \geq \ell$ for all $x \in S$. The largest lower bound of S is called the infimum and is denoted $\inf S$.

Quantifiers.

- $\forall$ means 'For all' (also 'For any')
- $\exists$ means 'There exists'
- $\nexists$ means 'There does not exist'
- : means 'such that'

For sets S and S', a function $f \colon S \to S'$ has the property that the function output $f(x)$ belongs to S' for every input x from S, that is: $f(x) \in S' \ \forall x \in S$. If the maximum of the function on S is 1, then $\exists\, x \in S : f(x) = 1$ and $\forall y > 1 \nexists x \in S : f(x) = y$. The first statement means that there exists a value x in S for which the function $f(x)$ attains its maximum 1. The second statement means that for any value y strictly greater than 1 there is no x in S such that the function $f(x)$ takes the value y.

2 Linear algebra

Matrices. A matrix $A = (a_{ij})_{i=1,j=1}^{m,n} \in \mathbb{R}^{m\times n}$ is a rectangular array of shape m by n with values in $\mathbb{R}$

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \dots & a_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & a_{m3} & \dots & a_{mn} \end{pmatrix}.$$

If $m = n$, the matrix is said to be square.

A square matrix is said to be diagonal if $a_{ij} = 0$ for $1 \leq i \neq j \leq n$.

The identity matrix, denoted I, is the diagonal matrix with only ones on the diagonal, that is, $a_{ii} = 1$ for $i \in [n]$.

Vectors. A vector $v = (v_i)_{i=1}^{m} \in \mathbb{R}^m = \mathbb{R}^{m\times 1}$ is a matrix of shape m by 1. By convention, a vector is a matrix with one column

$$v = \begin{pmatrix} v_1 \\ \vdots \\ v_m \end{pmatrix}.$$

Transposes. The transpose of a matrix $A = (a_{ij})_{i=1,j=1}^{m,n} \in \mathbb{R}^{m\times n}$, denoted $A^\top$, is the matrix $A = (a_{ji})_{j=1,i=1}^{n,m} \in \mathbb{R}^{n\times m}$

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \dots & a_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & a_{m3} & \dots & a_{mn} \end{pmatrix} \in \mathbb{R}^{m\times n}, \qquad A^\top = \begin{pmatrix} a_{11} & a_{21} & \dots & a_{m1} \\ a_{12} & a_{22} & \dots & a_{m2} \\ a_{13} & a_{23} & \dots & a_{m3} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1n} & a_{2n} & \dots & a_{mn} \end{pmatrix} \in \mathbb{R}^{n\times m}.$$

A square matrix A is said to be symmetric if $A^\top = A$.

Hadamard products. Let $A = (a_{ij})_{i=1,j=1}^{m,n} \in \mathbb{R}^{m\times n}$ and $B = (b_{ij})_{i=1,j=1}^{m,n} \in \mathbb{R}^{m\times n}$ be matrices, the Hadamard product $A \circ B$ is defined as the matrix $(a_{ij}b_{ij})_{i=1,j=1}^{m,n} \in \mathbb{R}^{m\times n}$

$$A \circ B = \begin{pmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \dots & a_{mn} \end{pmatrix} \circ \begin{pmatrix} b_{11} & \dots & b_{1n} \\ \vdots & \ddots & \vdots \\ b_{m1} & \dots & b_{mn} \end{pmatrix} = \begin{pmatrix} a_{11}b_{11} & \dots & a_{1n}b_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1}b_{m1} & \dots & a_{mn}b_{mn} \end{pmatrix}.$$

Matrix products. Let $A = (a_{ij})_{i=1,j=1}^{m,n} \in \mathbb{R}^{m\times n}$ and $B = (b_{ij})_{i=1,j=1}^{m,n} \in \mathbb{R}^{n\times \ell}$ be matrices, the matrix product AB is defined as the matrix $C = (c_{ij})_{i=1,j=1}^{m,\ell} \in \mathbb{R}^{m\times \ell}$ where $c_{ij} = \sum_{k=1}^{n} a_{ik}b_{kj}$ for $i \in [m], j \in [\ell]$. For example, with $m = n = 2$, we have

$$AB = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix} = \begin{pmatrix} a_{11}b_{11} + a_{12}b_{21} & a_{11}b_{12} + a_{12}b_{22} \\ a_{21}b_{11} + a_{22}b_{21} & a_{21}b_{12} + a_{22}b_{22} \end{pmatrix}.$$

In general, for $A, B \in \mathbb{R}^{n\times n}$, we have $AB \neq BA$.

Matrix inverses. A square matrix $A \in \mathbb{R}^{n\times n}$ is invertible if there exists another matrix $B \in \mathbb{R}^{n\times n}$ such that

$$AB = BA = I_n.$$

If it exists, the matrix B is unique, is called the inverse of A, and is denoted A^{-1}.

Only square matrices can be inverted, for rectangular matrices there exists a concept of left and right inverses.

Eigendecomposition. Let $A \in \mathbb{R}^{n\times n}$ be a symmetric matrix. An eigenvalue/eigenvector pair of A is a pair $(v, \lambda) \in \mathbb{R}^n \times \mathbb{R}$ satisfying

$$Av = \lambda v.$$

As A is symmetric, it can be proved that there are exactly n eigenvalue/eigenvector pairs $(v_i, \lambda_i)_{i=1}^n$ with $v_i^\top v_j$ being 1 if $i = j$ or 0 otherwise (orthonormality). Let $V \in \mathbb{R}^{n\times n}$ be the matrix with v_i as i-th column for $i \in [n]$, and let $\Lambda \in \mathbb{R}^{n\times n}$ be the diagonal matrix with λ_i as (i, i)-th entry for $i \in [n]$. Then the eigendecomposition of A is

$$A = V\Lambda V^\top.$$

Singular Value Decomposition (SVD). This is a generalisation of the eigendecomposition for a rectangular matrix $A \in \mathbb{R}^{m\times n}$ which takes the form

$$A = U\Sigma V^\top$$

with $U \in \mathbb{R}^{m\times m}$ and $V \in \mathbb{R}^{n\times n}$ orthogonal matrices (*i.e.*, $UU^\top = U^\top U = I_m$ and $VV^\top = V^\top V = I_n$), and $\Sigma \in \mathbb{R}^{m\times n}$ having diagonal entries $\sigma_1 \geq \cdots \geq \sigma_p$ where $p = \min(m, n)$. These are called the singular values and correspond to the square roots of the nonzero eigenvalues of both $AA^\top$ and $A^\top A$, they satisfy

$$Av_i = \sigma_i u_i$$

where u_i and v_i are the i-th columns of U and V, respectively. Equivalently, the singular value decomposition can be written as

$$A = \sum_{i=1}^{p} \sigma_i u_i v_i^\top.$$

The number of strictly positive singular values corresponds to the rank of A.

3 Analysis

Continuity. A function $f \colon \mathbb{R} \to \mathbb{R}$ is continuous if $\forall \epsilon > 0\, \exists \delta > 0 : |x - y| < \delta \implies |f(x) - f(y)| < \epsilon$. The symbol $\implies$ means 'implies'. Intuitively, the definition means that if x and y are close (*i.e.*, $|x - y| < \delta$) then the function values $f(x)$ and $f(y)$ must also be close (*i.e.*, $|f(x) - f(y)| < \epsilon$), and no matter how close we want $f(x)$ and $f(y)$ to be (*i.e.*, for any small ϵ), this can be achieved by requiring x and y to be close enough (*i.e.*, $\exists \delta > 0$).

Here, the absolute value $|x|$ is defined as x if $x \geq 0$ and $-x$ if $x < 0$.

Derivatives. A continuous function $f \colon \mathbb{R} \to \mathbb{R}$ is differentiable at $x \in \mathbb{R}$ if the limit

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

exists, in which case this limit is called the derivative of f and is denoted by $f'(x)$ or $\frac{df}{dx}$. Intuitively, the derivative of f at x can be thought of as the slope of the tangent to the f curve at x.

- $\frac{d}{dx}x^n = nx^{n-1}$
- $\frac{d}{dx}e^x = e^x$
- $\frac{d}{dx}\ln(x) = 1/x$
- $\frac{d}{dx}\sin(x) = \cos(x)$
- $\frac{d}{dx}\cos(x) = -\sin(x)$
- Multiplication rule: $\frac{d}{dx}(uw) = \frac{du}{dx}w + u\frac{dw}{dx}$
- Chain rule: $\frac{dy}{dx} = \frac{du}{dx}\frac{dy}{du}$

Derivatives can be generalised to functions of multiple inputs, as well as to vectors and matrices. See The Matrix Cookbook for details.

Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, $\mathbf{A} \in \mathbb{R}^{n\times n}$ and $u \in \mathbb{R}$.

- $\frac{\partial \mathbf{x}}{\partial u} = \left(\frac{\partial x_1}{\partial u}, \dots, \frac{\partial x_n}{\partial u}\right) \in \mathbb{R}^n$
- $\frac{\partial u}{\partial \mathbf{x}} = \left(\frac{\partial u}{\partial x_1}, \dots, \frac{\partial u}{\partial x_n}\right) \in \mathbb{R}^n$
- $\frac{\partial \mathbf{x}}{\partial \mathbf{y}} = \left(\frac{\partial x_i}{\partial y_j}\right)_{1\leq i,j\leq n} \in \mathbb{R}^{n\times n}$
- $\frac{\partial \mathbf{x}^\top \mathbf{y}}{\partial \mathbf{x}} = \frac{\partial \mathbf{y}^\top \mathbf{x}}{\partial \mathbf{x}} = \mathbf{y}$ which can be seen as for $j = 1, \dots, n$ we have

$$\frac{\partial \mathbf{x}^\top \mathbf{y}}{\partial x_j} = \frac{\partial}{\partial x_j}\sum_{i=1}^{n} x_i y_i = x_j$$

- $\frac{\partial \mathbf{A}^\top \mathbf{x}}{\partial \mathbf{x}} = \frac{\partial \mathbf{x}^\top \mathbf{A}}{\partial \mathbf{x}} = \mathbf{A}$
- $\frac{\partial \mathbf{x}^\top \mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = (\mathbf{A} + \mathbf{A}^\top)\mathbf{x} = 2\mathbf{A}\mathbf{x}$ if $\mathbf{A}$ is symmetric
- $\frac{\partial \mathbf{x}^\top \mathbf{x}}{\partial \mathbf{x}} = 2\mathbf{x}$

Measures. A measure is a function which assigns to sets some non-negative values. The empty set is necessarily assigned the value 0. A measure also has the property that breaking down a set in disjoint subsets and summing up the measures of these subsets is equivalent to measuring the set in the first place.

The most common measure on $\mathbb{R}$ is called the Lebesgue measure, for which a real interval (a, b) for $b > a$ has measure its length $b - a$. The Lebesgue measure of the whole space $\mathbb{R}$ is infinite.

Measures for which the whole space (say $\mathbb{R}$) has finite measure equal to 1 are called probability measures.

Integrals. For a set S, the indicator function $\mathbf{1}_S(x)$ is defined to be 1 if $x \in S$ or 0 if $x \in S$. The integral of the indicator function $\mathbf{1}_S(x)$ with respect to a measure μ is defined to be

$$\int \mathbf{1}_S \, d\mu = \mu(S).$$

Consider a linear combination of indicator functions $s = \sum_{i=1}^n c_i \mathbf{1}_{S_i}$ with $c_i \geq 0$ for $i \in [n]$, which we refer to as a simple function. Enforcing lineariey, we define the integral of s to be

$$\int s \, d\mu \int \left(\sum_{i=1}^n c_i \mathbf{1}_{S_i} \right) \, d\mu = \sum_{i=1}^n c_i \int \mathbf{1}_{S_i} \, d\mu = \sum_{i=1}^n c_i \mu(S_i).$$

The integral of s over a subset A is defined as

$$\int_A s \, d\mu = \int \mathbf{1}_A f \, d\mu = \sum_{i=1}^n c_i \int \mathbf{1}_A \mathbf{1}_{S_i} \, d\mu = \sum_{i=1}^n c_i \int \mathbf{1}_{A \cap S_i} \, d\mu = \sum_{i=1}^n c_i \mu(A \cap S_i).$$

Suppose f is a non-negative function, then its integral over A is defined to be

$$\int_A f \, d\mu = \sup \left\{ \int_A s \, d\mu : s \text{ simple and } 0 \leq s \leq f \right\}.$$

Finally, for a signed function f, we first decompose it as $f = f^+ - f^-$ where $f^+ = \max(f, 0)$ and $f^- = \max(-f, 0)$ are two non-negative functions, and we define the integral of f over A to be

$$\int_A f \, d\mu = \int_A f^+ \, d\mu - \int_A f^- \, d\mu.$$

Using derivatives and the Lebesgue measure, we can perform a change of variable of the integration

$$\int_A f \, dx = \int_A \left(f \frac{dx}{dy} \right) \, dy.$$

4 Probability

Random variable. A random variable (often denoted X) is a variable whose possible values are numerical outcomes of a random phenomenon. This randomness can be characterised by defining a probability measure (often denoted $\mathbb{P}$).

Expectations. If X is a random variable with probability measure $\mathbb{P}$, then the expectation/mean of X is

$$\mathbb{E}[X] = \int x \, \mathrm{d}\mathbb{P}.$$

If the probability measure $\mathbb{P}$ is on a discrete space $\{x_i : i \in \mathbb{N}\}$ this gives rise to a probability mass function (p.m.f.)

$$p_X(x) = \mathbb{P}(X = x)$$

and the expectation becomes

$$\mathbb{E}[X] = \int x \, \mathrm{d}\mathbb{P} = \sum_{i \in \mathbb{N}} x_i \, \mathbb{P}(X = x_i) = \sum_{i \in \mathbb{N}} x_i \, p_X(x_i).$$

If the probability measure $\mathbb{P}$ is on a continuous space, then in some cases the probability measure $\mathbb{P}$ can be characterised by a probability density function (p.d.f.) $p_X(x)$ such that

$$\mathbb{P}(a \leq X \leq b) = \int_a^b p_X(x) \, dx$$

(can be generalised to more dimensions) and the expectation becomes

$$\mathbb{E}[X] = \int x \,\mathrm{d}\mathbb{P} = \int x \,\mathrm{d}p_X(x) = \int x p_X(x)\,\mathrm{d}x.$$

If it exists, the random variable is entirely characterised by its probability mass/density function.

Variances. The variance of a random variable X is defined as

$$\operatorname{var}(X) = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2.$$

The standard deviation is defined as $\sqrt{\operatorname{var}(X)}$. The variance and standard deviation measure the spread of the distribution of X around its mean $\mathbb{E}[X]$.

Common distributions. See Table 1.

Name	Symbol	Parameters	Sample space	p.m.f./p.d.f.	Mean	Variance
Bernoulli	$\mathrm{Ber}(p)$	$p \in [0,1]$	$\{0,1\}$	$p_X(0) = 1-p,\ p_X(1) = p$	p	$p(1-p)$
Binomial	$\mathrm{Bin}(n,p)$	$n \in \mathbb{N}, p \in [0,1]$	$\{0,1,\ldots,n\}$	$p_X(k) = \binom{n}{k}p^k(1-p)^{n-k}$	np	$np(1-p)$
Geometric	$\mathrm{Geom}(p)$	$p \in (0,1]$	$\mathbb{N} \setminus \{0\}$	$p_X(k) = p(1-p)^{k-1}$	$1/p$	$\frac{1-p}{p^2}$
Poisson	$\mathrm{Pois}(\lambda)$	$\lambda > 0$	$\mathbb{N}$	$p_X(k) = \lambda^k e^{-\lambda}/k!$	λ	λ
Uniform	$\mathrm{Unif}\{a,b\}$	$a,b \in \mathbb{Z} : a < b$	$\{a,\ldots,b\}$	$p_X(k) = 1/(b-a+1)$	$\frac{a+b}{2}$	$\frac{(b-a+1)^2-1}{12}$
Uniform	$\mathrm{Unif}[a,b]$	$a,b \in \mathbb{R} : a < b$	$[a,b]$	$p_X(x) = 1/(b-a)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Gaussian	$\mathcal{N}(\mu,\sigma)$	$\mu > 0, \sigma > 0$	$\mathbb{R}$	$\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-(x-\mu)^2/2\sigma^2\right)$	μ	σ^2
Gaussian	$\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$	$\boldsymbol{\mu} \in \mathbb{R}^d, \boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$	$\mathbb{R}^d$	$\frac{\exp\left(-(\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})/2\right)}{(2\pi)^{d/2}\det(\boldsymbol{\Sigma})^{1/2}}$	$\boldsymbol{\mu}$	$\boldsymbol{\Sigma}$

Table 1: Common probability distributions.

Markov's inequality. If X is a non-negative random variable, then for all $t > 0$

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[X]}{t},$$

and equivalently, we have for all $s > 0$

$$\mathbb{P}(X \geq s\,\mathbb{E}[X]) \leq \frac{1}{s}.$$

Chebyshev's inequality. If X is a random variable with finite mean μ and non-zero finite variance σ^2, then for all $t > 0$

$$\mathbb{P}(|X-\mu| \geq t\sigma) \leq \frac{1}{t^2},$$

and equivalently, we have for all $s > 0$

$$\mathbb{P}(|X-\mu| \geq s) \leq \frac{\sigma^2}{s^2}.$$

5 Other important basic concepts

See Complexity Notes.

- ‘Big O’ notation
- ‘Big Theta’ notation
- ‘Big Omega’ notation
- Time complexity of algorithms
- P versus NP

See Week 2 Kernel Slides.

- Vector spaces
- Inner products
- Norms