Mathematics for Machine Learning

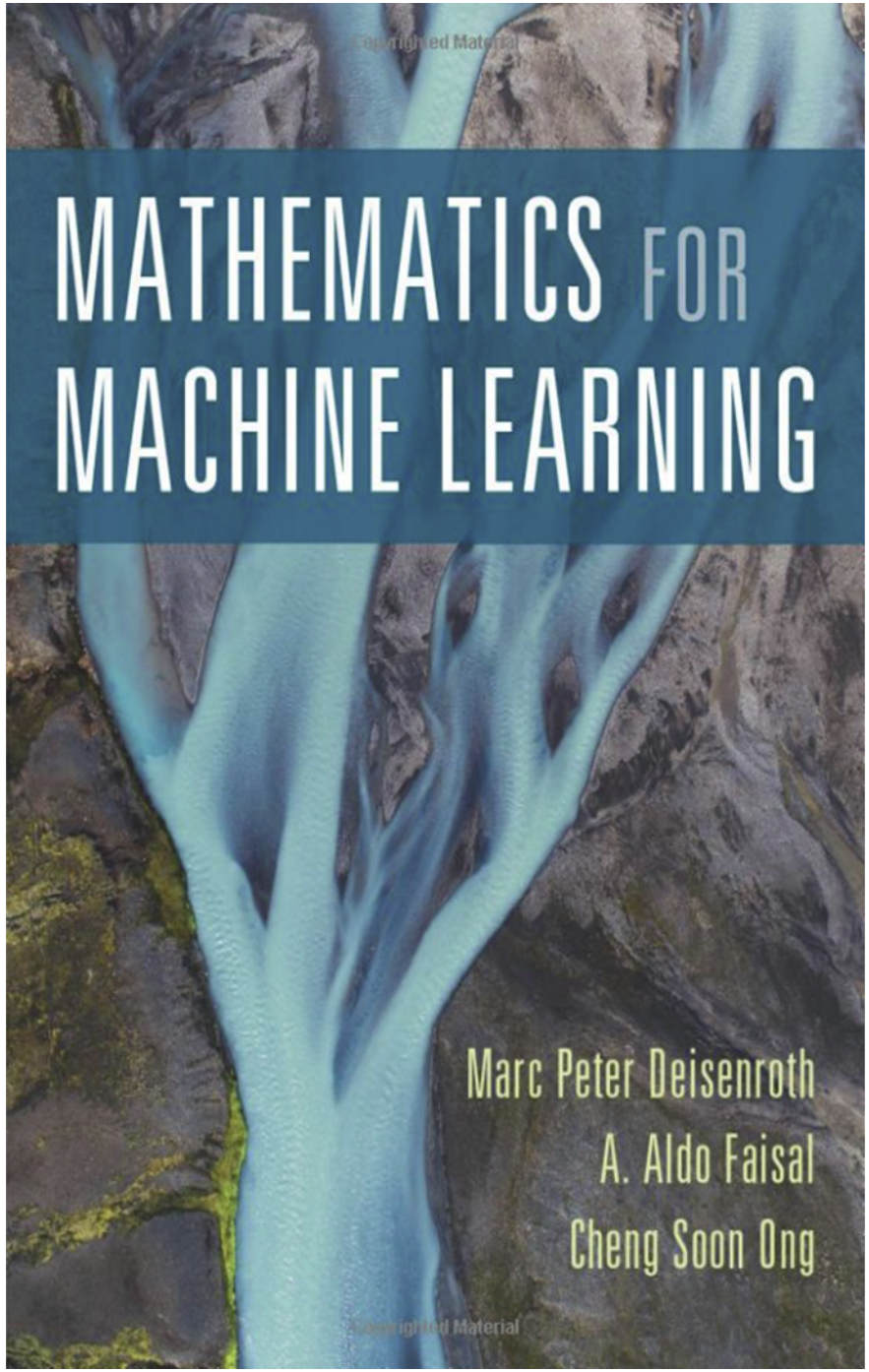
In this session ...

- Calculus
- Linear Algebra
- Probability
- Other methods

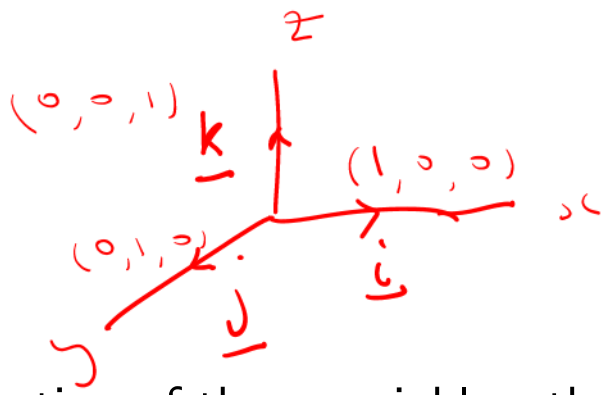
There is no ML in this session!

This module presents a number of themes associated with Machine Learning (ML). The lectures in ~~this~~ these modules cover several advanced topics in this field.

This lecture only covers the advanced mathematical methods that are/to be employed by other tutors in their Machine Learning lectures. We will not be doing ML or applications of the maths to ML.


Copyrighted Material
MATHEMATICS FOR MACHINE LEARNING
Marc Peter Deisenroth
A. Aldo Faisal
Cheng Soon Ong
Copyrighted Material

Extrema



1D

$$\frac{df}{dx} = \left(\frac{d}{dx}\right) f$$

Vector Calculus

If $f = f(x, y, z)$ is a function of three variables, then the *gradient of* f is

$$\nabla f(x, y, z) = f_x(x, y, z)\,\mathbf{i} + f_y(x, y, z)\,\mathbf{j} + f_z(x, y, z)\,\mathbf{k}.$$

The symbol ∇, called the **del** operator, is a vector differential operator symbolised by

$\mathbf{i} = (1,0,0)$ $\mathbf{k} = (0,0,1)$ $\mathbf{j} = (0,1,0)$

$$\nabla = \mathbf{i}\frac{\partial}{\partial x} + \mathbf{j}\frac{\partial}{\partial y} + \mathbf{k}\frac{\partial}{\partial z}.$$

(f_x, f_y, f_z)

It has properties similar to the operator d/dx. Standing alone it is meaningless; however, if it operates on $f(x, y, z)$ it produces the three-dimensional vector function given above.

In applications, the gradient $\nabla f(x, y, z)$ is sometimes denoted by **grad** $f(x, y, z)$.

∇: to find derivs in higher dim$^{\text{n}}$s. Gives rate of increase in each direction

Also called Hamilton operator

Example 1: Calculate ∇f for $f(x, y, z) = x^2 + yz$.

$$f_x = 2x; \; f_y = z; \; f_z = y,$$

partial differentiation

to give

$$\nabla f(x, y, z) = 2x\mathbf{i} + z\mathbf{j} + y\mathbf{k}$$

$$\nabla f = \mathbf{i}\frac{\partial f}{\partial x} + \mathbf{j}\frac{\partial f}{\partial y}$$

Example 2: If $f(x,y,z) = yz^3 - 2x^2$; find the gradient of f at the point $P(2,-3,1)$.

$$f_x = -4x;\ f_y = z^3;\ f_z = 3yz^2,$$

to give

$$\nabla f = (f_x, f_y, f_z)$$

$$\nabla f(x,y,z) = -4x\mathbf{i} + z^3\mathbf{j} + 3yz^2\mathbf{k}$$

Hence, at $P(2,-3,1)$

$$\nabla f(x,y,z) = -8\mathbf{i} + \mathbf{j} - 9\mathbf{k}.$$

$$\nabla \cdot \nabla \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \right)$$

Scalar

$\nabla \cdot \underline{f}$ divergence (vector fields source of each point)

vector product

$\nabla \times \underline{f}$ curl (infinitesimal circulation) $=$

$$\underline{f} = f_1\underline{i} + f_2\underline{j} + f_3\underline{k}$$

$$\begin{vmatrix} \underline{i} & \underline{j} & \underline{k} \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ f_1 & f_2 & f_3 \end{vmatrix}$$

Calculus Osowski

Constrained Optimisation

Lagrange Multipliers

Let $f(x, y, z)$ and $g(x, y, z)$ have continuous first order partial derivatives, and suppose f has an extremum $f(x_0, y_0, z_0)$ when (x, y, z) is subject to the constraint $g(x, y, z) = 0$. If $\nabla g(x_0, y_0, z_0) \neq \mathbf{0}$ then there is a real number λ such that

$$\nabla f(x_0, y_0, z_0) = \lambda \nabla g(x_0, y_0, z_0)$$

λ is called a *Lagrange multiplier*.

∇ Theorem

g is some fn.

$\underbrace{R \cdot v - ? = 0}_{g}$ $\quad R \cdot v = ?$

≥ 1

Example 3: Find the extrema of $f(x,y) = xy$; if (x,y) is restricted to the ellipse $4x^2 + y^2 = 4$.

Solution: In this example the constraint is $g(x,y) = 4x^2 + y^2 - 4 = 0$.

Setting $\nabla f(x,y) = \lambda \nabla g(x,y)$, we obtain

$$y\mathbf{i} + x\mathbf{j} = \lambda(8x\mathbf{i} + 2y\mathbf{j}).$$

Equating coefficients

$$y = 8\lambda x; \; x = 2\lambda y,$$

together with $4x^2 + y^2 - 4 = 0$. A number of ways to solve, here we eliminate y.

$$x - 2\lambda(8\lambda x) = 0.$$

So solve

$$x\left(1 - 16\lambda^2\right) = 0$$

to get the values $x = 0$ or $\lambda = \pm 1/4$.

$x = 0$: using $4x^2 + y^2 - 4 = 0 \to y = \pm 2$.

Possible choices for extrema: $(0, 2)$; $(0, -2)$.

If $\lambda = \pm 1/4$, then $y = 8\lambda x = \pm 2x$. Substitute in $4x^2 + y^2 - 4 = 0$ to get $x = \pm\sqrt{2}/2$.

The corresponding y values are $y = \pm\sqrt{2}$.

Exercise : Verify for yourself.

This gives the following points

$$\left(\sqrt{2}/2, \pm\sqrt{2}\right)$$
$$\left(-\sqrt{2}/2, \pm\sqrt{2}\right)$$

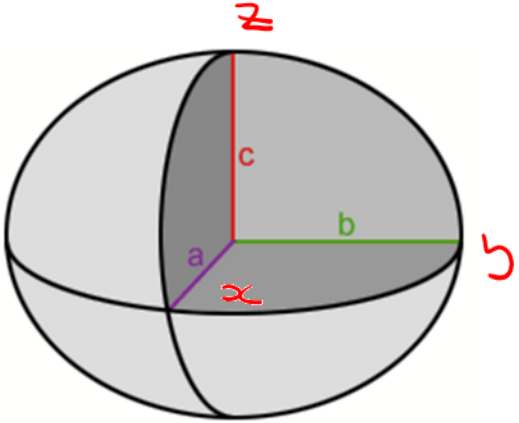
(x, y)	$f(x, y)$
$(0, 2)$	0
$(0, -2)$	0
$\left(\sqrt{2}/2, \sqrt{2}\right)$	1
$\left(\sqrt{2}/2, -\sqrt{2}\right)$	-1
$\left(-\sqrt{2}/2, \sqrt{2}\right)$	-1
$\left(-\sqrt{2}/2, -\sqrt{2}\right)$	1

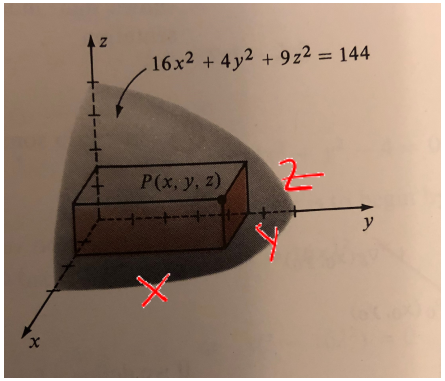
So we see $f(x, y) = 1$ is the maximum value at 1 at either $\left(\sqrt{2}/2, \sqrt{2}\right)$ or $\left(-\sqrt{2}/2, -\sqrt{2}\right)$.

A minimum value of $f(x, y) = -1$ is attained at $\left(\sqrt{2}/2, -\sqrt{2}\right)$ or $\left(-\sqrt{2}/2, \sqrt{2}\right)$.

Example 4 (harder): Find the volume of the largest rectangular box with faces parallel to the coordinate planes that can be inscribed in the ellipsoid $16x^2 + 4y^2 + 9z^2 = 144$. Constraint

Start by considering a sketch of the problem.





f not given

$V = 2x \times 2y \times 2z$

$\left(\frac{x}{a}\right)^2 + \left(\frac{y}{b}\right)^2 + \left(\frac{z}{c}\right)^2 = 12^2$

We wish to maximise

$$V = f(x, y, z) = 8xyz.$$

$\left(\frac{x}{12a}\right)^2 + \left(\frac{y}{12b}\right)^2 + \left(\frac{z}{12c}\right)^2 = 1$

subject to the constraint

$$g(x, y, z) = 16x^2 + 4y^2 + 9z^2 - 144 = 0$$

$\nabla f = \lambda \nabla g$

- Consider

$$\nabla f(x,y,z) = \lambda \nabla g(x,y,z)$$

or

$$8yz\mathbf{i} + 8xz\mathbf{j} + 8xy\mathbf{k} = \lambda\left(32x\mathbf{i} + 8y\mathbf{j} + 18z\mathbf{k}\right)$$

- Together with $g(x,y,z) = 0$ gives the following system of 4 equations:

$$8yz = 32\lambda x,\ 8xz = 8\lambda y,\ 8xy = 18\lambda z,\ 16x^2 + 4y^2 + 9z^2 - 144 = 0.$$

- After some algebra we have $xyz = 12\lambda$

- The first equation gives

$$32\lambda\left(3 - x^2\right) = 0$$

so either $\lambda = 0$ or $x = \sqrt{3}$.

- Similarly, second gives

$$8\lambda\left(12-y^2\right)=0$$

hence $y=2\sqrt{3}$.

- Finally multiplying the third equation $8xy=18\lambda z$ by z

$$8xyz=18\lambda z^2$$

and using the values found earlier

$$8\left(12\lambda\right)=18\lambda z^2$$

which gives $z=4/\sqrt{3}$.

Hence we obtain the desired volume

$$V=8\left(\sqrt{3}\right)\left(2\sqrt{3}\right)\left(4/\sqrt{3}\right)=64\sqrt{3}\text{unit}^3.$$

Constrained Extrema

Finding relative max./min of $f(x,y,z)$ s.t constraints condition $g(x,y,z)=0$ consists of the formation of the Auxiliary Function

A.F. Lagrangian $\boxed{G(x,y,z;\lambda) = f(x,y,z) + \lambda g(x,y,z)}$ λ – Lagrange multiplier

subject to conditions $G_x = 0$; $G_y = 0$; $G_z = 0$; $G_\lambda = 0$

Ex: Optimise the function $f(x,y) = 4x^2 + 3xy + 6y^2$ s.t constraint $x+y=56$. g

Lagrangian $G(x,y) = f + \lambda g = 4x^2 + 3xy + 6y^2 + \lambda(56 - x - y)$ (*)

$$\left.\begin{aligned} G_x &= 8x + 3y - \lambda = 0 \\ G_y &= 3x + 12y - \lambda = 0 \\ G_\lambda &= 56 - x - y = 0 \end{aligned}\right\} \text{3 eq}^{\text{n}}\text{s in 3 unknowns.}$$

$\boxed{x_0 = 36 \; ; \; y_0 = 20 \; ; \; \lambda = 348}$

Subst in $G(x,y)$.

Exercise: ① Re-do this problem using $\nabla f = \lambda \nabla g$

② Redo earlier examples by expressing as a Lagrangian.

$$\nabla f = \lambda \nabla g$$

$$\nabla\left(4x^2+3xy+6y^2\right) = \lambda \nabla (x+y-56)$$

$$(8x+3y,\ 3x+12y) = \lambda(1,1)$$

Eq 1: $8x+3y = \lambda$

Eq 2: $3x+12y = \lambda$

Eq 3: $x+y-56 = 0$

Essential Linear Algebra

Given $\mathbf{x} \in \mathbb{R}^N; \mathbf{y} \in \mathbb{R}^N$. Define the inner (scalar) product:

$$\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^\mathsf{T}\mathbf{y}$$

generalising

$$= \sum_{i=1}^{N} x_i y_i.$$

Inner products and norms are closely related

$$\|\mathbf{x}\| = \sqrt{\mathbf{x}^\mathsf{T}\mathbf{x}}$$

$\mathbf{x}^\mathsf{T}\mathbf{x} = (x_1, \ldots, x_N)\begin{pmatrix} x_1 \\ \\ x_N \end{pmatrix} = x_1^2 + \ldots + x_N^2$

Recall $\|\underline{x}\|_2 = \sum_{i=1}^{N} x_i^2$

l_2 norm

$$\langle \mathbf{x}, \mathbf{y} \rangle = \|\mathbf{x}\| \, \|\mathbf{y}\| \cos\theta$$

This gives

$$\cos\theta = \frac{\mathbf{x}^\mathsf{T}\mathbf{y}}{\sqrt{\mathbf{x}^\mathsf{T}\mathbf{x}\mathbf{y}^\mathsf{T}\mathbf{y}}}$$

$\langle \, , \, \rangle$ represents a mathematical op.

$$\langle f, g \rangle = \int_0^1 fg$$

Field

$\mathbb{R}^N$ $\sum_{i=1}^{m} \lambda_i \underline{v}_i$ l.c

Linear Dependence/Independence

Set of vectors $\underline{v}_1,\ \underline{v}_2,\ \underline{v}_3, \ldots\ldots\ldots,\ \underline{v}_m \in V$ ← λ_i

1. The set $\{\underline{v}_1,\ \underline{v}_2,\ \underline{v}_3, \ldots\ldots\ldots,\ \underline{v}_m\}$ is said to be linearly dependent (l.d)
   ⇌ if $\exists$ scalars $\{\lambda_1, \ldots\ldots\ldots,\ \lambda_m\} \in \mathbb{F}$ not all zero s.t

$$\lambda_1\underline{v}_1 + \lambda_2\underline{v}_2 + \ldots\ldots + \lambda_m\underline{v}_m = \underline{0}$$

Special set

The left hand side is called a LINEAR COMBINATION of $\underline{v}_1, \ldots\ldots\ldots,\ \underline{v}_m$.

2. If $\{\underline{v}_1, \ldots\ldots\ldots,\ \underline{v}_m\}$ are not l.d then they are linearly independent (l.i). In this case an equation such as

$\mathbb{F}$

$$\lambda_1\underline{v}_1 + \lambda_2\underline{v}_2 + \ldots\ldots\ldots\ldots + \lambda_m\underline{v}_m = \underline{0}$$

$$\Rightarrow \lambda_1 = \lambda_2 = \ldots\ldots\ldots = \lambda_m = 0$$

$\mathbb{Z}^3$ $\mathbb{Z}$

$\mathbb{R}^N$ Space of vectors

Vector Space $\mathbb{R}$ → field.

Examples $\underline{v}_i, \underline{w}_i \in \mathbb{R}^3$; $\lambda_i \in \mathbb{R}$ (field)

① $\underline{v}_1 = \begin{pmatrix} 0 \\ -1 \\ 1 \end{pmatrix}$; $\underline{v}_2 = \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}$; $\underline{v}_3 = \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix}$

linear system

$$\sum_{i=1}^{3} \lambda_i \underline{v}_i = \underline{0} = \lambda_1 \underline{v}_1 + \lambda_2 \underline{v}_3 + \lambda_3 \underline{v}_3 \leftarrow$$

$$= \lambda_1 \begin{pmatrix} 0 \\ -1 \\ 1 \end{pmatrix}_{\underline{v}_1} + \lambda_2 \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}_{\underline{v}_2} + \lambda_3 \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix}_{\underline{v}_3} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

augmented matrix

$$\left(\begin{array}{ccc|c} 0 & 1 & -1 & 0 \\ -1 & 0 & 1 & 0 \\ 1 & -1 & 0 & 0 \end{array}\right)$$

$\lambda_2 = \lambda_3 = \alpha$ (say) free var.

$\lambda_1 = \lambda_3 \longrightarrow \alpha \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$ w.l.o.g put $\alpha = 1$

$\therefore$ $\boxed{\lambda_i = 1 \quad \forall\, i}$ $\left(\underline{v}_1 + \underline{v}_2 + \underline{v}_3 = \underline{0}\right)$

e.g. $\underline{v}_1 = -\underline{v}_2 - \underline{v}_3$ ←

(2) $\underline{w}_1 = \begin{pmatrix} 2 \\ 2 \\ 1 \end{pmatrix}$; $\underline{w}_2 = \begin{pmatrix} -1 \\ 6 \\ 5 \end{pmatrix}$; $\underline{w}_3 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$ $\boxed{\sum_{i=1}^{3} \lambda_i \underline{w}_i = \underline{0}}$

$$\left(\begin{array}{ccc|c} 2 & -1 & 1 & 0 \\ 2 & 6 & 0 & 0 \\ 1 & 5 & 0 & 0 \end{array}\right) \sim \left(\begin{array}{ccc|c} 2 & -1 & 1 & \\ 0 & 7 & -1 & \\ 0 & 11 & -1 & \end{array}\right) \begin{matrix} \\ R_2 - R_1 \\ 2R_3 - R_2 \end{matrix} \sim \left(\begin{array}{ccc|c} 2 & -1 & 1 & \\ 0 & 7 & -1 & \\ 0 & 4 & 0 & \end{array}\right) \begin{matrix} \\ \\ R_3 - R_2 \end{matrix}$$

$4\lambda_2 = 0 \rightarrow \lambda_2 = 0$; $7\lambda_2 - \lambda_3 = 0 \rightarrow \lambda_3 = 0$ $\therefore$ $\lambda_1 = 0$

$\underline{w}_1, \underline{w}_2, \underline{w}_3$ are linearly indep $\therefore$ all λ's $= 0$

(3) A special set $\underline{e}_1 = (1,0,0)^T$; $\underline{e}_2 = (0,1,0)^T$; $\underline{e}_3 = (0,0,1)^T$

$$\left(\begin{array}{ccc|c} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{array}\right) \Rightarrow \lambda_1 = 0 = \lambda_2 = \lambda_3 \quad \therefore \text{ Lin. Indep}$$

$\underline{e}_i \in \mathbb{R}^3$ Standard basis for $\mathbb{R}^3$

(4) $\underline{u} = (1, 6, 9)$; $\underline{v} = (2, 4, 8)$; $\underline{w} = (0, 0, 0) \in \mathbb{R}^3$

Write $\lambda_1 \underline{u} + \lambda_2 \underline{v} + \lambda_3 \underline{w} = \underline{0}$

Choose $\lambda_1 = 0 = \lambda_2 \qquad \lambda_3 = 6$

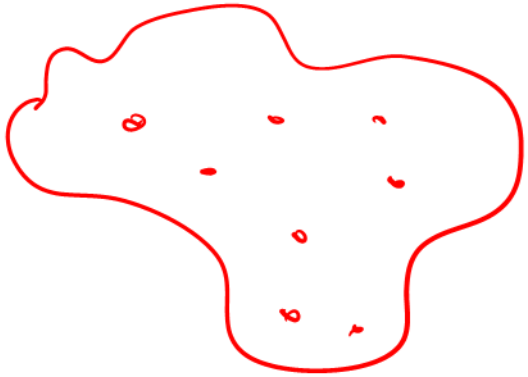
Then $0 \times \underline{u} + 0 \times \underline{v} + 6 \times \underline{w} = \underline{0}$ i.e λ_i's not all zero

$\therefore$ $\underline{u}, \underline{v}, \underline{w}$ lin. dep.

Moral: If n vectors in $\mathbb{R}^n$ and contain zero vector then set is linearly dependent.

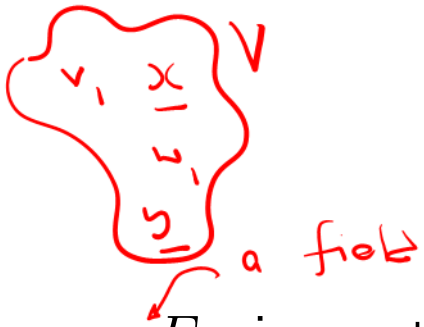
$\underline{v}_1, \underline{v}_2, \ldots, \underline{v}_n \in \mathbb{R}^n$ one of the vectors is $\underline{0}$

Vector Spaces



We are interested in the non-abstract treatment of this subject. Throughout we will use the term *field* denoted F to refer to a set of scalars.

Definition:



A $vector\ \ space\ \ V$ over F is a set with a binary operation called *Vector Addition,* denoted $V \times V \to V,\ \ (x,\ y) \mapsto x + y,$ and a function $F \times V \to V,\ \ (c,\ x) \mapsto cx,\ \ (c \in F)\,,$ called *Scalar Multiplication* such that the following eight rules hold:

Don't worry.

+1) $+$ is associative $\quad \forall x,\ y,\ z \in V \quad (x+y)+z = x+(y+z)$

+2) $+$ is commutative $\quad \forall x,\ y \in V \quad x+y = y+x$

+3) $+$ has a neutral $\quad \exists\ 0 \in V \quad \forall x \in V \quad x+0 = x$

+4) $+$ has inverse $\quad \forall x \in V \quad \exists\ y \in V \quad x+y = 0 \quad y$ is denoted $(-x)$

·1) $\cdot$ is associative $\quad \forall c,\ d \in F, \quad \forall x \in V \quad c\,(dx) = (cd)\,x$

·2) · is commutative $\forall x,\ y \in V \quad xy = yx$

·3) · has a neutral $\forall x \in V \quad 1 \cdot x = x \qquad (1 \neq 0)$

·4) · has an inverse $\forall x \in V \quad (x \neq 0) \Rightarrow \quad (\exists\, y \in V \quad xy = 1) \quad y$ is denoted $\left(x^{-1}\right)$

+·1) Right distributive $\forall c \in F \quad \forall x,\ y \in V \quad c\,(x + y) = cx + cy$ scalar multiplication is distributive over vector addition

+·2) Left distributive $(c + d)\,x = cx + dx$

Remarks:

$\mathbb{R}$ vector space

$\mathbb{R}^N$

field

1. Elements of F are called SCALARS and elements of V are called VECTORS.

2. If $F = \mathbb{R}$ we say V is a *real* vector space

   $= \mathbb{C}$ *complex*

   $= \mathbb{Q}$ *rational*

3. At this stage we have

2	$+$'s	addition
2	$\cdot$'s	multiplication
2	0's	neutrals
2	$-$'s	inverses

(and things usually get a lot worse) even in 1st year university algebra

4. The axioms can be used to deduce various rules.

Examples:

1. Let $m,\ n \in \mathbb{N}^+$ then ${}^{m}F^{n}$ is a vector space over F with respect to operations of matrix addition and scalar multiplication.

$$\begin{pmatrix} a & b & c \\ d & e & f \end{pmatrix} + \begin{pmatrix} A & B & C \\ D & E & F \end{pmatrix} = \begin{pmatrix} a+A & b+B & c+C \\ d+D & e+E & f+F \end{pmatrix}$$

$M1$ $M2$

$m=2,\ n=3$

$M1, M2 \in {}^{2}F^{3}$; $\lambda \in F$

$$\lambda M1 = \begin{pmatrix} \lambda a & \lambda b & \lambda c \\ \lambda d & \lambda e & \lambda f \end{pmatrix} \in {}^{2}F^{3}$$

2. Let $V = \mathbb{R}[x]$ denote the set of all polynomials

$$\sum_{n=0}^{N} a_n x^n \quad n \in \mathbb{N}, \quad a_i\,(i = 1, ..., N) \in \mathbb{R}$$

Then V is a vector space over $\mathbb{R}$ w.r.t. addition of polynomials and multiplication by a constant.

$$\sum_{n=0}^{N} a_n x^n + \sum_{n=0}^{N} b_n x^n = \sum_{n=0}^{N} (a_n + b_n) x^n = \sum_{n=0}^{N} c_n x^n$$

$$\lambda \sum_{n=0}^{N} a_n x^n$$

3. Let F be an arbitrary field and V the set of all n dimensional vectors with vector addition

$$(a_1,\ .....,\ a_n) + (b_1,\ .....,\ b_n) = (a_1 + b_1,\ ....., \ a_n + b_n)$$

and scalar multiplication

$$k\,(a_1,\ a_2,\ .....,\ a_n) = (ka_1,\ ka_2,\ .....,\ ka_n)$$

where $a_i,\ b_i,\ k \in F$. Then V is a vector space over F.

3.1 Subspaces

Special subset of a vector space

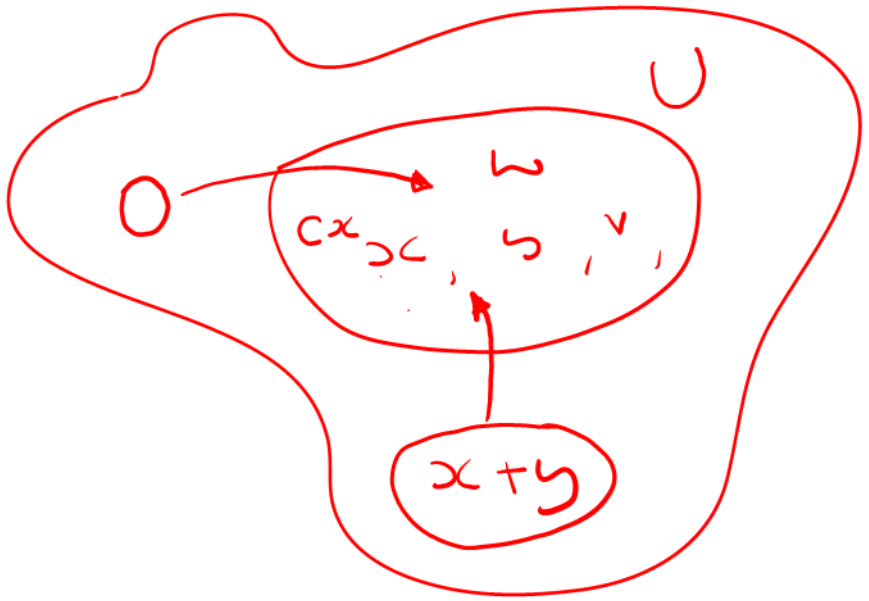
Definition:

A subspace U of a vector space V (over F) is a subset , i.e. $U \subset V$ such that

(i) $0 \in U$

(ii) $\forall x,\ y \in U \quad x + y \in U$

(iii) $\forall c \in F \quad \forall x \in U \quad cx \in U$



Note:

U is then a vector space with vector addition and scalar multiplication calculated in V.

$U \times U \to U$ is a binary operation on U

$$(x,\ y) \mapsto x + y$$

$F \times U \to U,$ is a function

$$(c,\ x) \mapsto cx$$

The eight rules obviously hold because they are satisfied in the larger space and will automatically be satisfied in every subspace.

Examples:

$\mathbb{R}^3$

(1) Let $V =$ set of all 3×3 matrices and $U,\ W \subset V$ such that

$U =$ set of lower triangular matrices

$$\begin{pmatrix} a & 0 & 0 \\ d & b & 0 \\ e & f & c \end{pmatrix}$$

$W =$ set of symmetric matrices.

$$\begin{pmatrix} A & \alpha & \beta \\ \alpha & B & \gamma \\ \beta & \gamma & C \end{pmatrix}$$

Suppose $A,\ B \in U;\ \ C,\ D \in W,$ then $A + B \in U$ and $cA \in U$ where $c \in \mathbb{R}$. The sums $A + B$ and cA inherit properties of A and B; and

similarly $C + D \in W$ and $kC \in W$ where $k \in \mathbb{R}$. 0 is in both spaces. Hence U and W are subspaces of V. ✓

quad I

(2) Consider the vector space $V = \mathbb{R}^2$ over $F = \mathbb{R}$. $U = \{(x,\ y)\ /\ x,\ y\ \in [0,\ \infty)$ is the subset consisting of vectors whose components are ≥ 0, i.e the first quadrant, all co-ordinates $(x,\ y) \geq 0$. Now $\underline{0} \in U$ and U is closed under vector addition $(x + y) \in U$. What about closure under scalar multiplication? Suppose $\underline{u} = (1,\ 1) \in U$ and $c = -1 \in \mathbb{R}$, then $c\underline{u} = (-1,\ -1) \notin U \therefore$ scalar multiplication fails. Hence U is not a subspace of $\mathbb{R}^2$.

Now suppose $W = \{(-a,\ -b)\ /\ a,\ b\ \in [0,\ \infty)\} \subset V$, i.e. the third quadrant. Define a new subset of V such that $S = U + W \subset V$. We see that for any vector $\underline{w}$ in S, $k\underline{w} \in S$ where $k \in \mathbb{R}$, so closure under scalar multiplication. However now addition fails because $(2,\ 1) + (-1,\ -3) = (1,\ -2)$

U W IV

which is in neither quadrant.

So the smallest subspace containing the 1st quadrant is the whole space $\mathbb{R}^2$.

Vector Spaces Summary

$\underline{a}, \underline{b}, \underline{c}, \ldots, \underline{x}, \underline{y}, \underline{t}$ vectors

$\alpha, \beta, \gamma, \lambda, \ldots$ scalars

Vector Space: set V on which are defined ops

① scalar mult^{n} ② vector add^{n}

scalar multiplication: $\underline{x} \in V \quad \lambda \in \underline{\underline{F}}$ (field)

$\underline{w} = \lambda \underline{x} \in V$

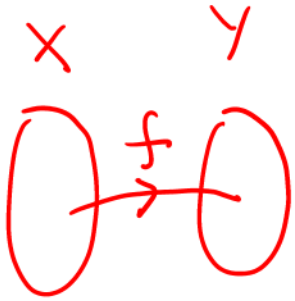
vector addition: $\underline{x}, \underline{y} \in V$

$\underline{v} = \underline{x} + \underline{y} \in \underline{\underline{V}}$

8 properties satisfied:
comm.; associative; $\exists$ neutral element (e.g. $0 (+1)$ $1 (\times 1)$);
inverse, distributive

Linear Mappings:

V, W vector spaces over some field F.



Linear mapping f such that

$$f: V \to W \qquad f: X \to Y$$

is a mapping satisfying

(2) $\forall\ \underline{v}, \underline{w} \in V$ $\quad f(\underline{v} + \underline{w}) = f(\underline{v}) + f(\underline{w})$ ←

(1) any scalar $\lambda \in F$ and any $\underline{v} \in V$

$$f(\lambda \underline{v}) = \lambda f(\underline{v})$$

Linear map also preserves $\underline{0}$ onto itself ←

f is called a linear transf^{n}

V — departure space of f

W — arrival space of f

$$\begin{pmatrix} 0 \\ 0 \end{pmatrix} \to \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

2D 3D

Example 1: $f : \mathbb{R}^2 \to \mathbb{R}^3$ ($V \to W$) $\quad f(x,y) = (2x+y,\ 0,\ y-x)$

(i)

$$f(\lambda \underline{v}) = f\left(\lambda \begin{pmatrix} x \\ y \end{pmatrix}\right) = f\begin{pmatrix} \lambda x \\ \lambda y \end{pmatrix} = \begin{pmatrix} 2\lambda x + \lambda y \\ 0 \\ \lambda y - \lambda x \end{pmatrix} = \lambda \begin{pmatrix} 2x+y \\ 0 \\ y-x \end{pmatrix} = \lambda f\begin{pmatrix} x \\ y \end{pmatrix}$$

(ii)

$$f(\underline{v}+\underline{w}) = f\left(\begin{pmatrix} x_1 \\ y_1 \end{pmatrix} + \begin{pmatrix} x_2 \\ y_2 \end{pmatrix}\right) = \begin{pmatrix} 2x_1+y_1 + 2x_2+y_2 \\ 0 + 0 \\ y_1 - x_1 + y_2 - x_2 \end{pmatrix}$$

($\underline{v}$, $\underline{w}$)

$$= \begin{pmatrix} 2x_1+y_1 \\ 0 \\ y_1 - x_1 \end{pmatrix} + \begin{pmatrix} 2x_2+y_2 \\ 0 \\ y_2 - x_2 \end{pmatrix} = f\begin{pmatrix} x_1 \\ y_1 \end{pmatrix} + f\begin{pmatrix} x_2 \\ y_2 \end{pmatrix} = f(\underline{v}) + f(\underline{w})$$

($\underline{v}$, $\underline{w} \in \mathbb{R}^2$)

(iii)

$$f\begin{pmatrix} 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 2\times 0 + 0 \\ 0 \\ 0 - 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

$\therefore \underline{0}$ preserved

$\underline{0} \in \mathbb{R}^2$ $\qquad$ $\underline{0} \in \mathbb{R}^3$

i.e $f(\underline{0}_{\mathbb{R}^2}) := \underline{0}_{\mathbb{R}^3}$

Example 2: $g: \mathbb{R}^2 \to \mathbb{R}^3$ $\quad g\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x+3y \\ 1 \\ y \end{pmatrix}$!!

$g\begin{pmatrix} 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \neq \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$ $\underline{0}$ is not preserved $\therefore$ f not a linear mapping.

Linear maps can be represented by matrices.

Consider standard basis for $\mathbb{R}^2$ $\left\{\underline{e}_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \underline{e}_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}\right\}$

and earlier $f(x,y) = (2x+y, 0, y-x)$

$f(\underline{e}_1) = f\begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 2 \\ 0 \\ -1 \end{pmatrix}$ $\qquad f(\underline{e}_2) = f\begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}$

$$\begin{pmatrix} 2 & 1 \\ 0 & 0 \\ -1 & 1 \end{pmatrix}$$

This is the matrix rep. of f wrt the standard basis of $\mathbb{R}^2$.

$\underline{e}_i$ are lin. indep and orthogonal

All vectors in $\mathbb{R}^n$ can be written uniquely as a lin. comb.

Matrix Diagonalisation

$f = x^2$ $(x_1+y_1^2)^2 \neq x_1^2+y_1^2$

Given $A \in \mathbb{R}^n$. $|A - \lambda I| = 0$ characteristic eqn

Form the matrix $\underline{\underline{P}}$ from the eigenvectors of A. Then A can be factorised as $P^{-1} = P^T$

$$A = PDP^{-1} \qquad PDP^T$$

Where $D = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}$ Check: $D = P^{-1}AP$

We wish to calculate A^n.

$\prod_{i=1}^{n} D^i$

n lots of D

$A^n = (PDP^{-1})(PDP^{-1})(PDP^{-1}) \ldots\ldots\ldots (PDP^{-1})$

$= PD \ldots\ldots DP^{-1} = (PD^nP^{-1})$

Example:

Given $M = \begin{pmatrix} 0 & -1 \\ -5 & 4 \end{pmatrix}$ (cute) with $\lambda_1 = 5$, $\underline{v}_1 = \begin{pmatrix} -1 \\ 5 \end{pmatrix}$, $\lambda_2 = -1$, $\underline{v}_2 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$

$\underline{v}_1 \quad \underline{v}_2$

$$P = \begin{pmatrix} -1 & 1 \\ 5 & 1 \end{pmatrix} \qquad P^{-1} = -\frac{1}{6}\begin{pmatrix} 1 & -1 \\ -5 & -1 \end{pmatrix} \qquad D = \begin{pmatrix} 5 & 0 \\ 0 & -1 \end{pmatrix} \begin{matrix} \leftarrow \lambda_1 \\ \leftarrow \lambda_2 \end{matrix}$$

$$M^{\wedge} = (PDP^{-1})^{\wedge} = (PD^{\wedge}P^{-1})$$

$$= -\frac{1}{6}\begin{pmatrix} -1 & 1 \\ 5 & 1 \end{pmatrix}\begin{pmatrix} 5^{\wedge} & 0 \\ 0 & (-1)^{\wedge} \end{pmatrix}\begin{pmatrix} 1 & -1 \\ -5 & -1 \end{pmatrix}$$

(P = first matrix; P^{-1} = last matrix)

$$= -\frac{1}{6}\begin{pmatrix} -1 & 1 \\ 5 & 1 \end{pmatrix}\begin{pmatrix} 5^{\wedge} & -(5^{\wedge}) \\ -5(-1)^{\wedge} & -(-1)^{\wedge} \end{pmatrix} = -\frac{1}{6}\begin{pmatrix} -5^{\wedge} - 5(-1)^{\wedge} & 5^{\wedge} - (-1)^{\wedge} \\ 5^{\wedge+1} - 5(-1)^{\wedge} & -5(5^{\wedge}) - (-1)^{\wedge} \end{pmatrix}$$

$$M^{\wedge} = \frac{1}{6}\begin{pmatrix} 5^{\wedge} + 5(-1)^{\wedge} & (-1)^{\wedge} - 5^{\wedge} \\ 5(-1)^{\wedge} - 5^{\wedge+1} & 5(5^{\wedge}) + (-1)^{\wedge} \end{pmatrix}$$

$\wedge \Rightarrow$ Try this by induction

Principal Component Analysis

PCA is a technique that takes a dataset $\mathbf{X}$ (consisting of a set of N tuples in $p-$dimensional space) and finds the directions along which the tuples line up best.

In short, a dimension reduction technique:

1. Taking linear combinations of the p original variables. Say $\mathbf{Z} = \mathbf{X}\alpha_j$. Note: α_j is a vector.

2. Each of these linear combinations explains as much as possible of the variance in the data:

$$\max_{\alpha_j} V_j = \mathbb{V}\left(\mathbf{X}\alpha_j\right), \text{ with } V_j \geq V_k \text{ if } j < k.$$

3. Each linear combination is uncorrelated with all the others: $\alpha_j^T \alpha_k = 0 \; \forall j \neq k$.

4. Each linear combination has unit length: $\alpha_j^T \alpha_j = 1 \; \forall j$.

also think in vector norm form

$$\|X\| = \sqrt{X^T X}$$

PCA Derivation

Similar to last maths.

Denote the variance-covariance matrix of our dataset $\mathbf{X}$ as $\mathbf{\Sigma}$.

Let's focus on the first linear combination with maximum variance. In order to find this, we need to solve the following constrained optimization problem:

Take the first derivative and set to zero:

$G = f + \lambda(\cdot) = 0$

Note that we want to do maximisation

$$\max_{\alpha_1} \mathbb{V}(\mathbf{X}\alpha_1) \text{ s.t. } \alpha_1^\top \alpha_1 = 1$$

Constrained.

constraint

The Lagrangian is

$$\begin{aligned} L &= (\mathbf{X}\alpha_1)^\top \mathbf{X}\alpha_1 - \lambda_1\left(\alpha_1^\top\alpha_1 - 1\right) \\ &= \alpha_1^\top \mathbf{\Sigma}\alpha_1 - \lambda_1\left(\alpha_1^\top\alpha_1 - 1\right) \end{aligned}$$

$X^\top X = \Sigma$

done with ses.

Treat as α_1^2 $= \alpha_1^2 \Sigma - \lambda_1(\alpha_1^2 - 1)$

"Power Method"

Take first derivative and set to zero:

$$\frac{\partial L}{\partial \alpha_1} = \Sigma\alpha_1 - \lambda_1\alpha_1 = 0 \Rightarrow (\Sigma - \lambda_1 I_p)\,\alpha_1 = 0$$

factorise

Note we want to maximise $\alpha_1^T \Sigma \alpha_1 = \lambda_1 \alpha_1^T \alpha_1 = \lambda_1$.

Solution: $\lambda_1 =$ largest eigenvalue, $\alpha_1 =$ corresponding eigenvector.

To find the second linear combination, we need to take a further constraint into account

$$\max_{\alpha_2} \mathbb{V}(\mathbf{X}\alpha_2) \quad \text{s.t.} \quad \alpha_2^T\alpha_2 = 1, \quad \alpha_2^T\alpha_1 = 0$$

earlier def'n

The Lagrangian is

$$L = \alpha_2^T\Sigma\alpha_2 - \lambda_2\left(\alpha_2^T\alpha_2 - 1\right) - \theta_2\alpha_2^T\alpha_1$$

behaves like $\Sigma\alpha_2^2 - \lambda_2(\alpha_2^2 - 1) - \theta\alpha_2^2$

Or use "product rule"

Optimising yields $\frac{\partial \mathcal{L}}{\partial \alpha_2} = 2\Sigma\alpha_2 - 2\lambda_2\alpha_2 - 2\theta\alpha_2$

$$\frac{\partial L}{\partial \alpha_2} = \Sigma\alpha_2 - \lambda_2\alpha_2 - \theta_2\alpha_1 = 0$$

Left-multiplying α_1^T : $\underbrace{\alpha_1^T\Sigma\alpha_2}_{=0} - \underbrace{\lambda_2\alpha_1^T\alpha_2}_{=0} - \theta_2\underbrace{\alpha_1^T\alpha_1}_{=1} = 0 \longrightarrow \theta_2 = 0$

Therefore, ← to factorise

$$(\Sigma - \lambda_2 I_p)\,\alpha_2 = 0 \longrightarrow \Sigma = \lambda_2 I_p$$

$\Rightarrow \lambda_2 =$ second largest eigenvalue, $\alpha_2 =$ corresponding eigenvector.

$\left\{ L = \alpha_2^T \Sigma \alpha_2 = \alpha_2^T \lambda_2 I_p \alpha_2 = \lambda_2 \right\}$

differentiating wrt vectors ⇒ in module 2 More later this module

Some background calculus

for earlier work on extrema.

If f is a function of two variables x, y such that $f(x,y) \begin{cases} \leq \\ \geq \end{cases} f(x_0, y_0)$ at all points (x,y) inside some sufficiently small circle, centre (x_0, y_0), then $f(x,y)$ has a local $\begin{cases} \text{maximum} \\ \text{minimum} \end{cases}$ at (x_0, y_0).

We define a *critical point* (x_0, y_0) at which either

$$f_x(x_0, y_0) = 0 \text{ or } f_y(x_0, y_0) = 0$$

or

$$f_x(x_0, y_0) \text{ or } f_y(x_0, y_0) \text{ do not exist.}$$

For example consider $f(x,y) = 1 - x^2 + y^2$. Then

$$f_x = -2x \text{ and } f_y = 2y.$$

So if $f_x = f_y = 0$ then $x = 0,\ y = 0$. But the point $(0,0)$ corresponds neither to a local maxima nor a minima. Such a point is called a *saddle point*.

Testing for Local Extrema of a function of Two Variables

Suppose $f(x, y)$ has continuous partial derivatives of order one and two within some disc with centre (x_0, y_0). Let $f_x(x_0, y_0) = 0$ and $f_y(x_0, y_0) = 0$ and not all the partial derivatives of f are zero at (x_0, y_0). Define

$$\Delta = \begin{vmatrix} f_{xx}(x_0, y_0) & f_{xy}(x_0, y_0) \\ f_{xy}(x_0, y_0) & f_{yy}(x_0, y_0) \end{vmatrix}$$
$$= f_{xx} f_{yy}|_{(x_0,y_0)} - f_{xy}^2|_{(x_0,y_0)}$$

Then we have the following

- $\Delta > 0$ { Local maximum at (x_0, y_0) if $f_{xx}(x_0, y_0) < 0$; Local minimum at (x_0, y_0) if $f_{xx}(x_0, y_0) > 0$ }
- $\Delta < 0$ The function has a saddle point at (x_0, y_0)
- $\Delta = 0$ More work required to determine the nature of (x_0, y_0)

Probability

Maximum Likelihood Estimators (MLE)

- Suppose we draw $x_1,\ x_2, \ldots\ldots, x_n$ iid with density $p\left(\cdot;\theta\right)$.

- The probability of drawing a number in $[x_1, x_1 + dx]$ is $p\left(x_1;\theta\right) dx$

- Similarly the probability of drawing a number from x_2 to $x_2 + dx$ is $p\left(x_2;\theta\right) dx$ and so forth.

- The probability of all these events occurring is

$$p\left(x_1;\theta\right) \cdots p\left(x_n;\theta\right) dx^n = L\left(\theta; x_1, x_2, \ldots., x_n\right) dx^n$$

- The *likelihood function* is

$$L(\theta; x_1, x_2, \ldots, x_n) = L(\theta, \mathbf{x}),$$

which we think of as a function of θ.

parameter θ vector $\mathbf{x}$

Vector

- We ignore dx^n as this term is 'constant'.

- Maximum likelihood estimation works on the principle that the 'best' estimate of the parameters θ is given by maximising $L(\theta; \mathbf{x})$.

- As likelihoods are positive, and given independence are of the form

$$L(\theta, \mathbf{x}) = \prod_{i=1}^{n} p(x_i; \theta) = p(x_1;\theta) \cdots p(x_n;\theta)$$

it is often easier working with the log-likelihood

maths is easier.

$$\ell(\theta, \mathbf{x}) = \log L(\theta, \mathbf{x}) = \sum_{i=1} \log p(x_i; \theta)$$

if L then $\ell = \log L$

- Score Function $\frac{\partial \ell}{\partial \theta} = S(\Theta, \underline{X})$

- If ℓ is differentiable in θ, then the maximum is when the Score function (derivative of ℓ) is zero,

$$S(\theta, \mathbf{x}) = \frac{d\ell}{d\theta} = 0.$$

- Usually, ℓ cannot have local minima or saddle points, so there is a maximum. Theorem from calculus.

Suppose we pick n i.i.d random variables x_i such that $X_i \sim N\left(\mu, \sigma^2\right)$. Then the probability of drawing a number in the interval

$$x_i \text{ to } x_i + dx = p\left(x_i; \mu, \sigma\right) \quad \text{for} \quad i = 1, ..., n$$

where $p\left(x_i; \mu, \sigma^2\right)$ is the pdf for the normal distribution.

$$\begin{aligned} & L\left(x_1, x_2, ..., x_n; \mu, \sigma\right) \\ = \; & p\left(x_1; \mu, \sigma\right) ..... \times p\left(x_n; \mu, \sigma\right) (dx)^n . \end{aligned}$$

can omit

Recall, we ignore the term $(dx)^n$ as it is a constant. Hence

$$L\left(x_1, x_2, ..., x_n; \mu, \sigma\right) := \prod_{i=1}^{n} p\left(x_i; \mu, \sigma\right)$$

We now choose the parameters μ & σ in a way that maximises the likelihood of observing this. This is done by setting the **Score function** equal to zero

2 eqns ; 2 unknowns $S\left(\mu, \mathbf{x}\right) = \dfrac{\partial \ell}{\partial \mu} = 0, \quad S\left(\sigma, \mathbf{x}\right) = \dfrac{\partial \ell}{\partial \sigma} = 0.$

We have already seen (in the calculus course), in principle that these values can be local maxima/minima or saddle points.

However due to the nature of this problem, i.e. L is a likelihood function ensures that L can only have a maximum.

The problem can be simplified by noting that L attains a maximum iff $\ell = \log L$ has a maximum. In this case ℓ is called the *log-Likelihood function.*

So we have

$$\log L = \ell\left(x_1,\ x_2, ..., x_n;\ \mu,\ \sigma\right) = \sum_{i=1}^{n} \log p\left(x_i;\ \mu,\ \sigma\right)$$

and our problem becomes solving

$$\frac{\partial \ell}{\partial \mu} = 0, \quad \frac{\partial \ell}{\partial \sigma} = 0.$$

Example 1:

Suppose we draw n positive numbers X_i $(i = 1, ..., n)$ from i.i.d samples with distribution

$$f(x) = \begin{cases} 2\lambda^3 \sqrt{\frac{x}{\pi}} e^{-\lambda^2 x} & \text{for } x \geq 0 \\ 0 & \text{for } x < 0 \end{cases}$$

where the unknown parameter $\lambda > 0$.

How can we obtain a maximum Likelihood estimator $\widehat{\lambda}$ for λ?

$\widehat{\lambda}$ estimate

$$\frac{2\lambda^3}{\sqrt{\pi}} \sqrt{x_1}\, e^{-\lambda^2 x_1} \cdot \frac{2\lambda^3}{\sqrt{\pi}} \sqrt{x_2}\, e^{-\lambda^2 x_2} \cdot \cdots \cdots \frac{2\lambda^3}{\sqrt{\pi}} \sqrt{x_n}\, e^{-\lambda^2 x_n}$$

$$\left(\frac{2\lambda^3}{\sqrt{\pi}}\right)^n (x_1 x_2 \cdots x_n)^{1/2}\, e^{-\lambda^2 x_1 - \lambda^2 x_2 - \cdots - \lambda^2 x_n}$$

Math Stats. G. Hoel

The Likelihood function is

$$L(x_1,\ x_2, \ldots, x_n;\ \lambda) = \prod_{i=1}^{n} f(x_i;\lambda)$$

$$= \frac{2^n \lambda^{3n}}{\pi^{n/2}} \sqrt{(x_1 x_2 \ldots\ldots x_n)}\, e^{-\lambda^2 (x_1 + x_2 + \ldots\ldots + x_n)},$$

Test of basic algebra

and for $\lambda \to 0$ & ∞, $L \to 0$, so there must be at least one maximum.

We use $\ell = \log L$ (much easier), as ℓ has maximum iff L has a maximum.

$$\ell = \log \frac{2^n}{\pi^{n/2}} + 3n \log \lambda + \frac{1}{2} \log (x_1 x_2 \ldots\ldots x_n) - \lambda^2 (x_1 + x_2 + \ldots\ldots + x_n).$$

$\log \left(\frac{2^n}{\pi^{n/2}} \cdot \lambda^{3n} \cdot (x_1 x_2 \ldots\ldots x_n)^{1/2} \cdot e^{-\lambda^2 (x_1 + \ldots\ldots + x_n)} \right)$:= sum of logs

Score fn. $\frac{\partial \ell}{\partial \lambda} = 0$

So

$$\begin{aligned} \frac{\partial \ell}{\partial \lambda} &= \frac{3n}{\lambda} - 2\lambda\left(x_1 + x_2 + ... + x_n\right) = 0 \\ &\Rightarrow 2\widehat{\lambda}\left(x_1 + x_2 + ... + x_n\right) \\ &= \frac{3n}{\widehat{\lambda}} \end{aligned}$$

Hence

$$\widehat{\lambda}^2 = \frac{3n}{2\left(x_1 + x_2 + ... + x_n\right)}$$

and

$$\widehat{\lambda} = \sqrt{\frac{3n}{2\left(x_1 + x_2 + ... + x_n\right)}}.$$

The Maximum Likelihood estimator method is a popular method due to its simplicity when the form of the probability density function is known.

Example 2: Suppose $X_1, ..., X_n \sim N\left(\mu, \sigma^2\right)$ iid. What is the MLE of $\left(\mu, \sigma^2\right)$?

Repeat for $\frac{\partial l}{\partial \sigma} =>$

In this case,

pdf $$p\left(x_i; \mu, \sigma^2\right) = \left(2\pi\sigma^2\right)^{-1/2} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right)$$

so

Likelihood $$L\left(x_i; \mu, \sigma^2\right) = \left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\sum_i \frac{(x_i - \mu)^2}{2\sigma^2}\right)$$

log Likelihood $$l\left(x_i; \mu, \sigma^2\right) = -\frac{n}{2} \log\left(2\pi\sigma^2\right) - \sum_i \frac{(x_i - \mu)^2}{2\sigma^2}$$

Note that for μ, this is just least squares.

$$\frac{\partial l}{\partial \mu} = \frac{1}{2\sigma^2} \sum_i (x_i - \mu) \times -2 = 0 = \sum_{i=1}^{\hat{}} x_i - \sum_i^{\hat{}} \mu$$

$$\sum_i x_i - \mu \sum_i = 0$$

$n\mu = \sum x_i$ $\qquad \hat{\mu} = \frac{1}{n}\sum x_i$

Example 3: Suppose $X_1, ..., X_n \sim \exp(-\lambda x_i)$ iid. What is the MLE of λ?

In this case, -ve exp. dist^n

$$p(x_i; \lambda) = \lambda \exp(-\lambda x_i)$$

so

$$\begin{aligned} L(x_i; \lambda) &= \lambda^n \exp\left(-\lambda \sum_i x_i\right) \\ l(x_i; \lambda) &= n \log(\lambda) - \lambda \sum_i x_i \end{aligned}$$

hence the score function

$$S(x_i; \lambda) = \frac{n}{\lambda} - \sum_i x_i \qquad \text{diff} = 0$$

which gives the MLE

$$\hat{\lambda} = \frac{n}{\sum_i x_i} = \overline{x}^{-1} \qquad \hat{\lambda} = \frac{1}{\bar{x}}$$

Loss Functions

- We have an outcome measurement, which can be quantitative (such as stock price). Let's denote these **outcomes** as $\mathbf{y}$.

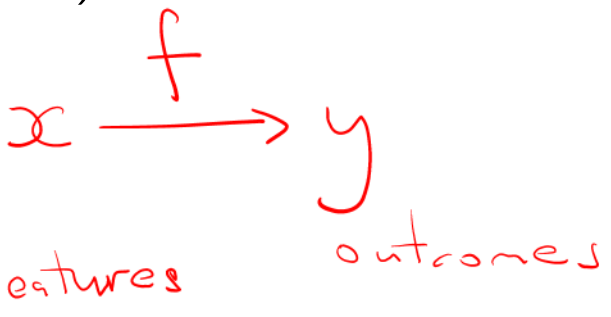
- We wish to predict the outcome based on a set of **features**. Let's denote these features as $\mathbf{x}$.

- We have a **training dataset**, where we observe the features and the outcome for a set of objects.

- Using the training data we build a **prediction model**, or **learner**, which we denote as $f(x)$. The model is a (possibly highly non-linear) transformation of features into outcomes:

$$y = f(x) + \varepsilon$$

$f(x)$ – true fn. to approximate



Other types of notation are

$$\begin{aligned} y &= f_\beta(\mathbf{x}) = f(\mathbf{x};\boldsymbol{\beta}) = \boldsymbol{\beta}^{\mathsf{T}}\mathbf{x} \\ &= (\beta_1,\ldots,\beta_N)\begin{pmatrix} x_1 \\ \vdots \\ x_N \end{pmatrix} = \sum_{i=1}^{N}\beta_i x_i. \end{aligned}$$

input– Calculus $x \in \mathbb{R}$. Vector Calculus: $\mathbf{x} \in \mathbb{R}^D$. ML: can also work for discrete inputs, strings, tree, graphs, ...

output– Classification: $y \in \{0,1\}$. Regression: $y \in \mathbb{R}$

methods– Linear classifiers, neural networks, decision trees, ensemble models,..

probabilistic classifiers, . . .

β– parameters/weights: $\beta \in \mathbb{R}$. $\boldsymbol{\beta} \in \mathbb{R}^D$

To set the scene: Here's a typical scenario of interest

Outcome y (this is the output) which is what we wish to predict

y : output

- ↗ continuous variable (e.g price) (regression)
- ↘ discrete variable (credit scoring, prediction of default) (classification)

$\mathbf{x}$ (set of features): input. So $\mathbf{x}$ can be a scalar or vector.

p : number of features - size of vector, i.e. $\mathbf{x} \in \mathbb{R}^p$.

Set of input and output pairs $\mathcal{S} = \left\{ \left(x^{(i)}, y^{(i)} \right) : i = 1, \ldots N \right\}$;

data set for training

$x^{(i)} \in \mathbb{R}$, $y^{(i)} \in \mathbb{R}$ we want to estimate the parameter (weights) vector $\beta = (\beta_1, \ldots, \beta_n)$ of unknown parameters.

$$\left(x^{(i)}, y^{(i)}\right) : i = 1, \ldots N \text{ pairs/2-tuples}$$

GOAL: Predict the relationship between the input and output.

So how is $x \longrightarrow y$ achieved?

So output is some (unknown) function of the input, together with some noise. Let's write

$$y = f(x) + \varepsilon$$

The estimate for f, which is sometimes denoted as $\widehat{f}$.

$\mathbb{E}[\varepsilon] = 0 \quad \mathbb{V}[\varepsilon] = \sigma^2$

$\mathbb{E}[y] = \mathbb{E}[f + \varepsilon] = \mathbb{E}[f] + \mathbb{E}[\varepsilon] = f$

The Loss Function L

Here we define the *loss function*

$$L\left(y, f\left(x\right)\right)$$

as a measure of how far our prediction is from the outcome. A popular definition is

$$L\left(y, f\left(x\right)\right) = \left(y - f\left(x\right)\right)^2$$

$y-$ actual outcome. $f\left(x\right)-$ predicted outcome.

We want to minimise L so that the distance between the two outcomes is the smallest value possible.

Loss functions are used in ML to estimate how badly models are performing - a measure of how wrong the model is in terms of its ability to estimate the relationship between input and output.

Look for $f(x)$ which minimises this 'error function' L, which can be written as

$$\underset{f(x)}{\arg\min}\{\mathbb{E}\left[L\left(y, f\left(x\right)\right)\right]\}$$

or in component form

$$\underset{f(x)}{\arg\min}\left\{\frac{1}{N}\sum_{i=1}^{N}\left(y^{(i)}-f\left(x^{(i)}\right)\right)^{2}\right\}$$

MSE for the training set is

$$\frac{1}{N}\sum_{i=1}^{N}\left(y^{(i)}-f\left(x^{(i)}\right)\right)^{2}$$

Loss fn. M.L.

In OR a classic optimisation problem is to maximise/minimise an objective function subject to a system of constraints The function we want to minimize or maximize is called the objective function. However when we are minimizing it, we may also call it the **cost function**, **loss function**, or error function.

• suppose that X is a $p-$dimensional real-valued random input vector and Y is a real valued random output variable.

joint pdf

• The (unknown) joint distribution of X and Y is $P(X, Y)$.

Y | X

• We seek to find a function $f(x)$ for predicting Y given X.

• To find the "optimal" $f(x)$, we need an objective (or loss) function, which we assume to be quadratic here

$$L(Y, f(X)) = (Y - f(X))^2$$

Econometrics

The expected loss function — math def. of an Expectation

$$\mathbb{E}[L] = \int_{\mathbb{R}}\int_{\mathbb{R}} (y - f(x))^2 P(x,y)\, dxdy$$

$$\begin{aligned}(y - f(x))^2 &= \{y - \mathbb{E}[y|x] + \mathbb{E}[y|x] - f(x)\}^2 \\ &= \{y - \mathbb{E}[y|x]\}^2 + 2\{y - \mathbb{E}[y|x]\}\{\mathbb{E}[y|x] - f\} + \{y - f\}^2\end{aligned}$$

① ② (=0) ③ — $\mathbb{E}$

$\mathbb{E}[①] = \mathbb{E}[(y-\mu_y)^2] = \mathbb{V}[y]$ ← variance.

$\mathbb{E}[②] = 0 \quad \because \mathbb{E}[y - \mathbb{E}[y]] = \mathbb{E}[y] - \mathbb{E}[\mu_y] = \mu_y - \mu_y = 0$

$\mathbb{E}[③] = \mathbb{E}[(\mathbb{E}[y] - \hat{f})^2] = \text{bias}^2$

$B(y,\hat{f})$

$B(\theta) = \mathbb{E}[\hat{\theta}] - \theta$

MSE of estimator = var. of est. + bias^2

Bias-Variance decomposition

There is a trade-off between bias and variance, with very flexible models having low bias and high variance, and relatively rigid models having high bias and low variance. The model with the optimal predictive capability is the one that leads to the best balance between bias and variance.

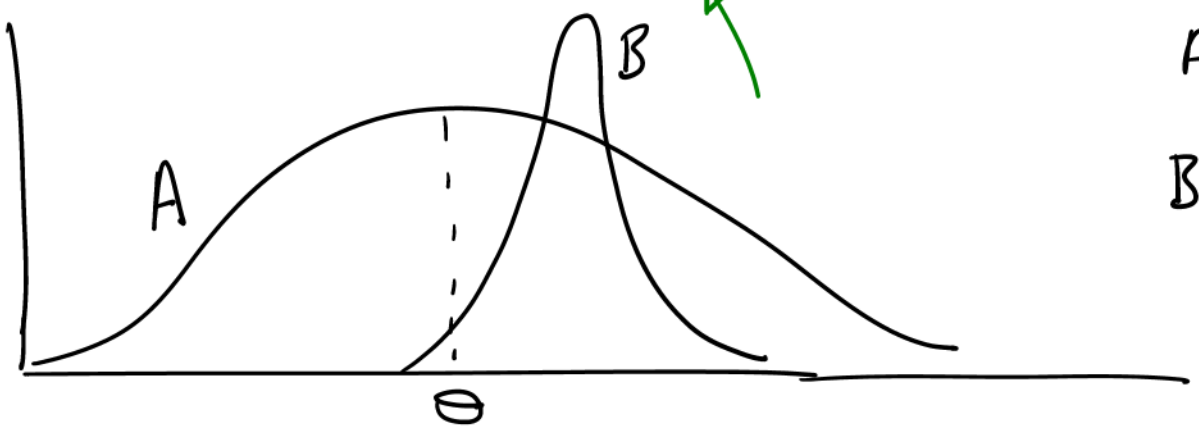
$y = f + \varepsilon$

Writing $y - \hat{f} = f + \varepsilon - \hat{f}$

$\mathbb{E}[(f + \varepsilon - \hat{f})] \longrightarrow \mathbb{E}[\varepsilon^2] = \sigma^2$ i.e $\mathbb{V}[\text{noise}]$

noise

$\mathbb{E}[L] = \mathbb{V}(\hat{f}) + \sigma^2 + \text{Bias}^2[\hat{f}]$



A: unbiased but larger var.

B: smaller var