CS5014 Machine Learning

Lecture 2 Maths background review

Lei Fang

Spring 2021

1 Introduction

So why this review session ?

Maths is useful

- rigorous and concise way of communicating results
- help us understand why *and* why not models work
- be able to derive your own model and algorithms!

Refresher on essential concepts

- only a refresher; we expect you have learnt them
  - don't expect to know everything after this
- not complete and not rigorous

A chance for self-assessment

- identify rusty area
- do self studies afterwards
- maths learning should be never-ending :-)

Mathematics for machine learning

Linear algebra

- leap forward from 1-d to N-dimensional
- number line axis to a plane and hyper-space

Calculus

- continuous (real-valued) functions by approximation, (say *polynomial*)
  - $y = sin(x)$ is approximated by $y = x$ when $x \approx 0$
- useful for optimisation

Probability theory and statistics

- study of uncertainty: uncertainty is the norm
  - e.g. rain tomorrow? blood pressure measurement (reading error)?
- how to generalise your results
  - from one sample to the universe: training error $\rightarrow$ testing?

Useful textbook and references (read the italic entries!)

Linear algebra

- *Learning from Data Supplementary Mathematics (Vector and Linear Algebra) by David Barber*; https://api.semanticscholar.org/CorpusID:18857001
- *Chapter 2 of Deep Learning by Ian Goodfellow, Yoshua Bengio and Aaron Courville* https://www.deeplearningbook.org/contents/linear_algebra.html
- Introduction to Linear Algebra by Gilbert Strang; http://math.mit.edu/~gs/linearalgebra/
- The Matrix Cookbook by Kaare Brandt Petersen, Michael Syskind Pedersen; https://www2.imm.dtu.dk/pubdb/pubs/3274-full.html
  - useful as a reference manual

Probability theory

- *Chapter 2.1-2.3 Information Theory, Inference, and Learning Algorithms by David J.C. MacKay* http://www.inference.org.uk/itprnn/book.pdf
- *Chapter 3.1-3.9 of Deep Learning by Ian Goodfellow, Yoshua Bengio and Aaron Courville* https://www.deeplearningbook.org/contents/prob.html
- Introduction to Probability Models by Sheldon Ross
  - chapter 1; chapter 2.1-2.5, 2.8; chapter 3.1-3.5

Calculus

- Use your book of choice; read multivariate calculus part as well
- Appendix of Bayesian Reasoning and Machine Learning by David Barber http://web4.cs.ucl.ac.uk/staff/

2 Linear Algebra

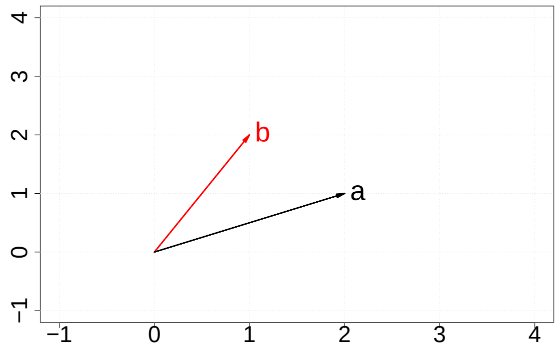
Linear algebra: basic concepts

- vectors
- norms and projection
- linear independence, span, subspace
- matrices, linear transformation
- rank, determinant, trace

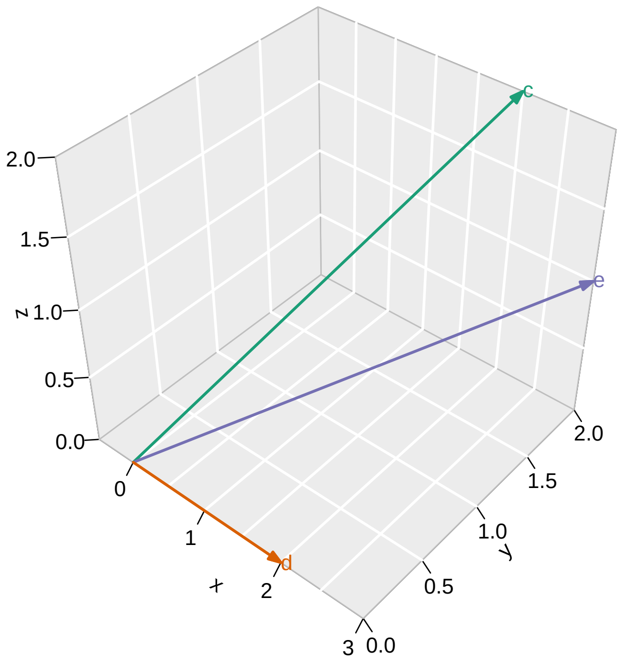
Vector

A vector is a collection of n salars

- $\boldsymbol{a} \in R^n$, default option is column vector i.e. $n \times 1$
- represents a **displacement** in R^n
- e.g. $\boldsymbol{a} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}, \boldsymbol{b} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$ (or $\boldsymbol{a} = [2, 1]^T$ to save space)



Some 3-d vectors $\boldsymbol{c} = \begin{bmatrix} 2 \\ 2 \\ 2 \end{bmatrix}, \boldsymbol{d} = \begin{bmatrix} 2 \\ 0 \\ 0 \end{bmatrix}, \boldsymbol{e} = \begin{bmatrix} 3 \\ 2 \\ 1 \end{bmatrix}$

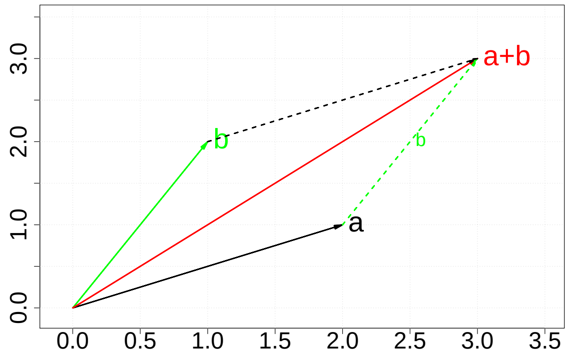


Vector addition

$$\boldsymbol{a} + \boldsymbol{b} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_d \end{bmatrix} + \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_d \end{bmatrix} = \begin{bmatrix} a_1 + b_1 \\ a_2 + b_2 \\ \vdots \\ a_d + b_d \end{bmatrix}$$

- generalisation from scalar addition; remember 2+1 on a number axis ?

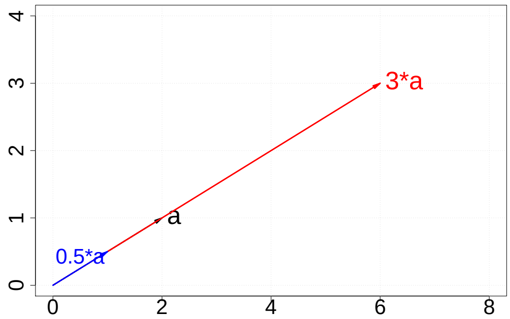
- parallelogram rule



Vector scaling/multiplication

$$k \cdot \boldsymbol{a} = k \cdot \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_d \end{bmatrix} = \begin{bmatrix} k \times a_1 \\ k \times a_2 \\ \vdots \\ k \times a_d \end{bmatrix}, k \in R \text{ or a scalar}$$

- geometrically, scaling means shrinking or streching a vector
  - the direction does not change but length changes
- arithmetically, $n \cdot \boldsymbol{a} = \boldsymbol{a} + \ldots + \boldsymbol{a} = \sum_n \boldsymbol{a}$
- $0 \cdot \boldsymbol{a} = \boldsymbol{0}$



Inner product

$$\boldsymbol{a}^T \boldsymbol{b} = [a_1, a_2 \ldots, a_d] \cdot \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_d \end{bmatrix} = \sum_{i=1}^{d} a_i \times b_i$$

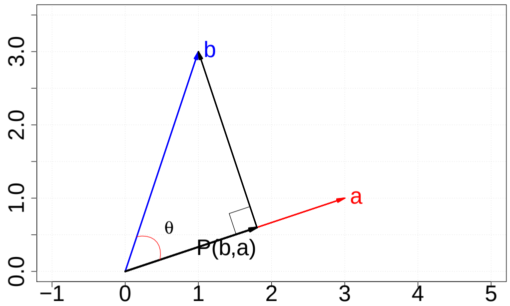
- $\boldsymbol{a}^T \boldsymbol{b} = \boldsymbol{b}^T \boldsymbol{a}$ and the result is a scalar
- $\boldsymbol{a}^T (\boldsymbol{b} + \boldsymbol{c}) = \boldsymbol{a}^T \boldsymbol{b} + \boldsymbol{a}^T \boldsymbol{c}$
- $(k\boldsymbol{a})^T \boldsymbol{b} = \boldsymbol{a}^T (k\boldsymbol{b}) = k(\boldsymbol{a}^T \boldsymbol{b})$
- $\boldsymbol{a}^T \boldsymbol{a} = \sum_{i=1}^{d} a_i^2$ is squared Euclidean distance between $\boldsymbol{a}$ and $\boldsymbol{0}$

- $\boldsymbol{a}^T\boldsymbol{a} \geq 0$ and $\boldsymbol{a} = \boldsymbol{0}$ if and only if $\boldsymbol{a}^T\boldsymbol{a} = 0$

Inner product and projection

Another interpretation:

$$\boldsymbol{a}^T\boldsymbol{b} = \|\boldsymbol{a}\|\|\boldsymbol{b}\|cos\theta$$



- θ is the angle between $\boldsymbol{a}, \boldsymbol{b}$
  - $\boldsymbol{a}^T\boldsymbol{b} = 0$ if and only if $\boldsymbol{a} \perp \boldsymbol{b}$
- $\|\boldsymbol{b}\|cos\theta = \|P(\boldsymbol{b}, \boldsymbol{a})\|$ is the projected length of $\boldsymbol{b}$ on $\boldsymbol{a}$
- $P(\boldsymbol{b}, \boldsymbol{a})$ denotes the projected vector of $\boldsymbol{b}$ to $\boldsymbol{a}$

$$P(\boldsymbol{b}, \boldsymbol{a}) = \|\boldsymbol{b}\|cos\theta * \frac{\boldsymbol{a}}{\|\boldsymbol{a}\|} = \frac{\boldsymbol{a}^T\boldsymbol{b}}{\boldsymbol{a}^T\boldsymbol{a}}\boldsymbol{a}$$

Matrix

A rectangular array of real numbers $A \in R^{m \times n}$

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix} = \begin{bmatrix} | & | & & | \\ \boldsymbol{a}_1 & \boldsymbol{a}_2 & \dots & \boldsymbol{a}_n \\ | & | & & | \end{bmatrix} = \begin{bmatrix} — & \tilde{\boldsymbol{a}}_1 & — \\ \vdots & \vdots & \vdots \\ — & \tilde{\boldsymbol{a}}_m & — \end{bmatrix}$$

- can be viewed as a collection of n column vectors $\boldsymbol{A} = [\boldsymbol{a}_1, \boldsymbol{a}_2, \dots, \boldsymbol{a}_n]$;
- or row vectors $\boldsymbol{A} = [\tilde{\boldsymbol{a}}_1^T, \tilde{\boldsymbol{a}}_2^T, \dots, \tilde{\boldsymbol{a}}_m^T]^T$
- sometimes written as $\boldsymbol{A} = (a_{ij})$ $i = 1, \dots m, j = 1, \dots, n$

Matrix operations

- addition: $\boldsymbol{A} + \boldsymbol{B} = \boldsymbol{C} = (c_{ij})$ where $c_{ij} = a_{ij} + b_{ij}$
- scaling: $k\boldsymbol{A} = \boldsymbol{C}$ where $c_{ij} = k * a_{ij}$
- transpose: $\boldsymbol{A}^T = \boldsymbol{C}$ where $c_{ij} = a_{ji}$
- multiplication: Let $\boldsymbol{A} \in R^{m \times s}, \boldsymbol{B} \in R^{s \times n}$

  $$\boldsymbol{AB} = \boldsymbol{C}, \boldsymbol{C} \in R^{m \times n}$$

  where

  $$c_{ij} = \sum_{k=1}^{s} a_{ik}b_{jk}$$

  or $c_{ij} = \tilde{\boldsymbol{a}}_i^T\boldsymbol{b}_j$

- $\boldsymbol{A}(\boldsymbol{B}\boldsymbol{C}) = (\boldsymbol{A}\boldsymbol{B})\boldsymbol{C}$
  - $\boldsymbol{A}\boldsymbol{B} \neq \boldsymbol{B}\boldsymbol{A}$
  - $(\boldsymbol{A}\boldsymbol{B})^T = \boldsymbol{B}^T\boldsymbol{A}^T$
  - $\boldsymbol{I}$ identity matrix: $\boldsymbol{I}\boldsymbol{A} = \boldsymbol{A}$ or $\boldsymbol{A}\boldsymbol{I} = \boldsymbol{A}$
- inverse (only square matrix): $\boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{I}$

Examples

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

it is not allowed as the dimensions do not match

$$\begin{bmatrix} 2 & 3 \\ 6 & 4 \\ 1 & 1 \end{bmatrix}^T = \begin{bmatrix} 2 & 6 & 1 \\ 3 & 4 & 1 \end{bmatrix}$$

Example

$$\begin{bmatrix} 2 & 3 \\ 6 & 4 \\ 1 & 1 \end{bmatrix} \times \begin{bmatrix} 5 & 2 & 2 \\ 1 & 1 & 2 \end{bmatrix} = \begin{bmatrix} \boxed{2 \times 5 + 3 \times 1} & c_{12} \\ c_{21} & c_{22} \\ c_{31} & c_{32} \end{bmatrix}$$

$$\begin{bmatrix} 2 & 3 \\ 6 & 4 \\ 1 & 1 \end{bmatrix} \times \begin{bmatrix} 5 & 2 & 2 \\ 1 & 1 & 2 \end{bmatrix} = \begin{bmatrix} 13 & 7 & 10 \\ 34 & 16 & 20 \\ 6 & 3 & 4 \end{bmatrix}$$

$$\begin{bmatrix} 5 & 2 & 2 \\ 1 & 1 & 2 \end{bmatrix} \times \begin{bmatrix} 2 & 3 \\ 6 & 4 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 24 & 25 \\ 10 & 9 \end{bmatrix}$$

Example

The inverse of $\boldsymbol{I} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ is itself $\boldsymbol{I}^{-1} = \boldsymbol{I}$

The inverse of $\boldsymbol{A} = \begin{bmatrix} 3 & 0 \\ 0 & 5 \end{bmatrix}$ is $\boldsymbol{A}^{-1} = \begin{bmatrix} 1/3 & 0 \\ 0 & 1/5 \end{bmatrix}$ as $\boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{I}$

The inverse of $\boldsymbol{B} = \begin{bmatrix} 1 & 2 & 0 \\ 1 & 2 & 1 \\ 1 & 1 & 0 \end{bmatrix}$ is $\boldsymbol{B}^{-1} = \begin{bmatrix} -1 & 0 & 2 \\ 1 & 0 & -1 \\ -1 & 1 & 0 \end{bmatrix}$

$\boldsymbol{C} = \begin{bmatrix} 2 & 0 & 1 \\ 1 & 1 & 1 \\ 0 & 0 & 0 \end{bmatrix}$ is not invertible: $\boldsymbol{C}^{-1}$ does not exist

Span, linear independence

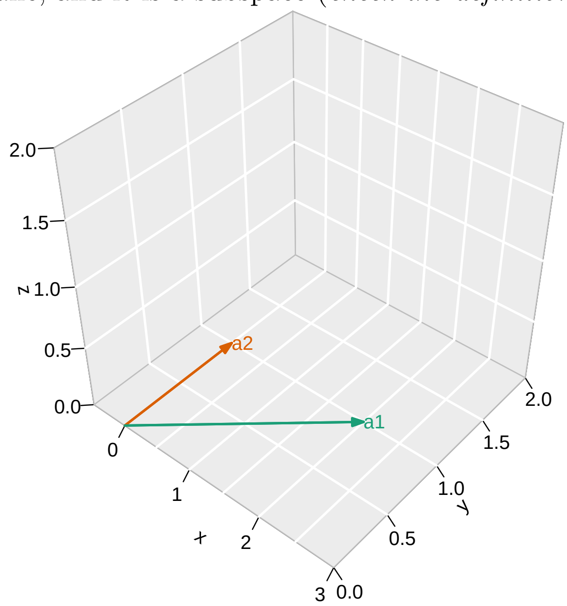
- **linear combination** is just sum of some scaled vectors
  - $\lambda_1 \cdot \boldsymbol{a}_1 + \lambda_2 \cdot \boldsymbol{a}_2 + \ldots + \lambda_n \boldsymbol{a}_n$, $\boldsymbol{a}_i \in R^m$ for $i = 1, \ldots, n$
  - $\boldsymbol{a}_i$ are vectors (of the same length) and λ_i are the scalars
- **span** is the set of all possible linear combination

$$\text{Span}(\{\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n\}) = \left\{ \sum_{i=1}^{n} \lambda_i \boldsymbol{a}_i | \lambda_i \in R, i = 1, \ldots, n \right\}$$

  - what is the span of $\{[1, 0]^T, [0, 1]^T\}$?
  - how about $\text{Span}(\{[2, 1]^T, [0, 1]^T\})$?
  - how about $\text{Span}(\{[2, 1]^T, [4, 2]^T\})$?
  - how about $\text{Span}(\{[2, 1, 0]^T, [0, 1, 0]^T\})$?
    * it is a **subspace** (xy plane) in R^3

$$\boldsymbol{a}_1 = [2, 1, 0]^T, \boldsymbol{a}_2 = [0, 1, 0]^T$$

the span is the xy plane, and it is a subspace (*check the definition of subspace!*)



- **linear independence**: $\{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n\}$ is linear independent if there exist no $\lambda_1, \ldots, \lambda_n$ (except all being 0) such that

$$\lambda_1 \cdot \boldsymbol{a}_1 + \lambda_2 \cdot \boldsymbol{a}_2 + \ldots + \lambda_n \boldsymbol{a}_n = \mathbf{0}$$

  - how about $\{[2,3]^T, [4,6]^T\}$?
  - are $\{[1,0]^T, [0,1]^T\}$ LI?
  - essentially a way to tell whether there is any *redundant* vectors in the set
- **rank** of $\boldsymbol{A} = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n]$: the maximum number of linearly independent column vectors
  - the span of the column vectors is called *column space*
  - *rank* also called the *dimension* of the column space
  - *dimension*: the minimum # of the vectors needed to span a space
    - what is the *dimension* for R^3?

Example

The column vectors of the matrix

$$[\boldsymbol{a}_1, \boldsymbol{a}_2, \boldsymbol{a}_3] = \begin{bmatrix} 2 & 0 & 1 \\ 1 & 1 & 1 \\ 0 & 0 & 0 \end{bmatrix}$$

are not linearly independent, as

$$\lambda_1 \boldsymbol{a}_1 + \lambda_2 \boldsymbol{a}_2 + \lambda_3 \boldsymbol{a}_3 = \mathbf{0}$$

holds for $\lambda_1 = \lambda_2 = 1, \lambda_3 = -2$. In other words, one of them is redundant. And $\text{rank}(\boldsymbol{A}) = 2$

Matrix vector multiplication

$$\boldsymbol{A}\boldsymbol{x} = \begin{bmatrix} | & | & & | \\ \boldsymbol{a}_1 & \boldsymbol{a}_2 & \ldots & \boldsymbol{a}_n \\ | & | & & | \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = x_1 \begin{bmatrix} | \\ \boldsymbol{a}_1 \\ | \end{bmatrix} + x_2 \begin{bmatrix} | \\ \boldsymbol{a}_2 \\ | \end{bmatrix} + \ldots + x_n \begin{bmatrix} | \\ \boldsymbol{a}_n \\ | \end{bmatrix}$$

- another view of the multiplication
- *linear combination* of the column vectors of $\boldsymbol{A}$
  - $x_1 \cdot \boldsymbol{a}_1 + x_2 \cdot \boldsymbol{a}_2 + \ldots + x_n \boldsymbol{a}_n$
  - $\boldsymbol{a}_i$ are the column vectors and x are the scalars
- so ... $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}$ essentially solves for what ?
  - $\boldsymbol{y}$ is in the column space of $\boldsymbol{A}$ or not ...
  - if not, then there is no solution
  - if yes, there will be some solution(s)? unique solution or ?

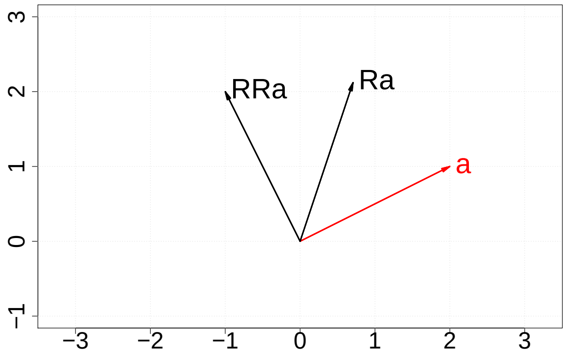
Matrix vector multiplication (some interpretations)

So $\boldsymbol{Ax}$ is a linear transformation: $R^n \to R^m$

- rotation: rotate $\boldsymbol{x}$ anti-clockwise by θ

$$R\boldsymbol{x} = \begin{bmatrix} cos\theta & -sin\theta \\ sin\theta & cos\theta \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0.71 & -0.71 \\ 0.71 & 0.71 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

say $\theta = \pi/4$



- $\boldsymbol{R}$ is a rotation (orthogonal) matrix if $\boldsymbol{R}^T = \boldsymbol{R}^{-1}$ (what does it imply?)
- $\boldsymbol{R}^T\boldsymbol{R} = \boldsymbol{R}\boldsymbol{R}^T = \boldsymbol{I}$
  - preserves length $(\boldsymbol{Rx})^T(\boldsymbol{Rx}) = \boldsymbol{x}^T\boldsymbol{R}^T\boldsymbol{Rx} = \boldsymbol{x}^T\boldsymbol{x}$
- rotation transform is alwasy invertible: inverse means rotating back!
- mathematics is always so reasonable :-)

- projection (an example): project $\boldsymbol{x} \in R^n$ to $\boldsymbol{a} \in R^n$

$$P(\boldsymbol{x}, \boldsymbol{a}) = \|\boldsymbol{x}\| cos\theta * \frac{\boldsymbol{a}}{\|\boldsymbol{a}\|} = \frac{\boldsymbol{a}^T\boldsymbol{x}}{\boldsymbol{a}^T\boldsymbol{a}}\boldsymbol{a}$$
$$= \frac{\boldsymbol{a} \cdot \boldsymbol{a}^T\boldsymbol{x}}{\boldsymbol{a}^T\boldsymbol{a}} = \frac{\boldsymbol{a}\boldsymbol{a}^T}{\boldsymbol{a}^T\boldsymbol{a}}\boldsymbol{x} = \boldsymbol{Px}$$

$$\boldsymbol{P} = \frac{\boldsymbol{a}\boldsymbol{a}^T}{\boldsymbol{a}^T\boldsymbol{a}}$$

- $\boldsymbol{P}$ is n by n square matrix
- $\boldsymbol{P}$ is a projection matrix: $\boldsymbol{Px}$ returns the projection
- what if we project it again (and again ...) ? $\boldsymbol{P}\ldots\boldsymbol{PPx}$
  - it remains unchanged, $\boldsymbol{P}\ldots\boldsymbol{PPx} = \boldsymbol{Px}$
  - or $\boldsymbol{PP} = \boldsymbol{P}$

$$\frac{\boldsymbol{a}\boldsymbol{a}^T}{\boldsymbol{a}^T\boldsymbol{a}}\frac{\boldsymbol{a}\boldsymbol{a}^T}{\boldsymbol{a}^T\boldsymbol{a}} = \frac{\boldsymbol{a}\boldsymbol{a}^T\boldsymbol{a}\boldsymbol{a}^T}{(\boldsymbol{a}^T\boldsymbol{a})^2} = \frac{\boldsymbol{a}\boldsymbol{a}^T}{\boldsymbol{a}^T\boldsymbol{a}}$$

- some projection cannot be reversed: some $\boldsymbol{P}$ is not invertible !
- mathematics always makes sense :-)

3 Calculus

Our revision on calculus will be brief. We will only review the concept of partial derivative, gradient, and hessian matrix. You should try to convince yourself all the results listed here if you can.

Calculus

- univariate derivative
- partial derivative
- gradient
- hessian matrix
- quadratic form and its derivative
- Lagrange multiplier

Derivative

Given function $f(x) : R \to R$, the derivative is defined as

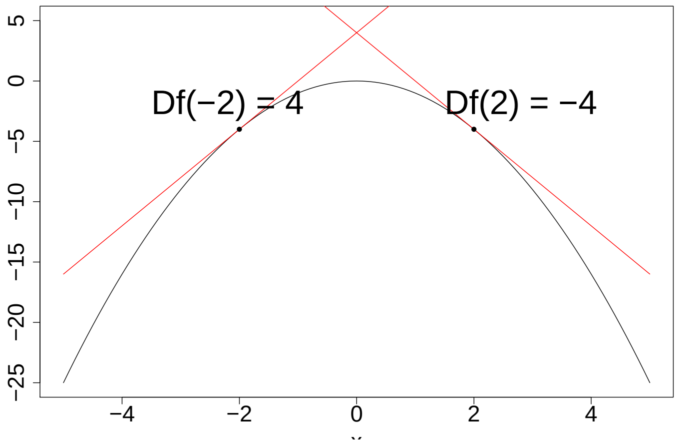
$$\frac{df}{dx} \equiv Df(x) \equiv \lim_{\Delta x \to 0} \frac{f(x + \Delta x)}{\Delta x}$$

- the growth rate at or near x
- the derivative points in the direction of steepest ascent of f
- when $Df(x) = 0$, f doesn't grow or decline near x, or *stationary*
  - either maximum (negative second derivative) or minimum (positive second derivative)
  - saddle point
  - for optimisation, just find $Df(x) = 0$

Example

$f(x) = -x^2$ what are the derivatives at -2 and 2?

- $Df(x) = -2x$;
- at $x = -2$, $Df(-2) = 4 > 0$ points to the right (ascent direction)
- at $x = 2$, $Df(2) = -4 < 0$ points to the left (ascent direction)
- at $x = 0$, $Df(0) = 0$, it's maximum as $D^2 f(0) = -2 < 0$



Partial derivative and gradient

Given function $f(\boldsymbol{x}) : R^n \to R$, partial derivative w.r.t. x_i is

$$\frac{\partial f}{\partial x_i} = \lim_{h \to 0} \frac{f(x_1, \ldots, x_i + h, \ldots, x_n)}{h}$$

The vector

$$\nabla_{\boldsymbol{x}} f = \text{grad} f = \frac{df}{d\boldsymbol{x}} = \left[\frac{\partial f(\boldsymbol{x})}{\partial x_1}, \frac{f(\boldsymbol{x})}{\partial x_2}, \ldots, \frac{f(\boldsymbol{x})}{\partial x_n} \right]$$

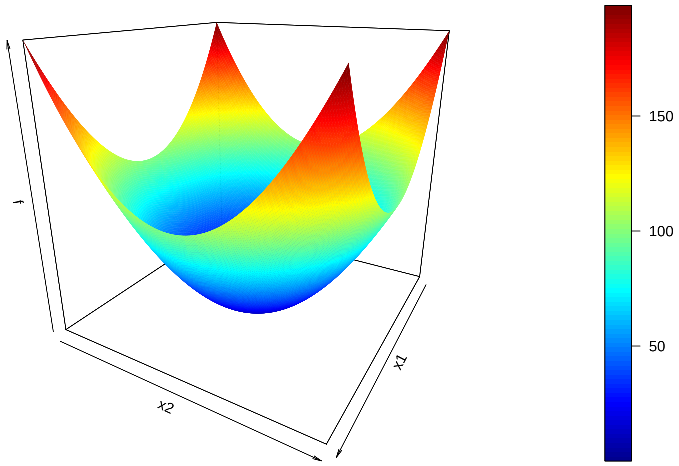
- partial derivative measure the growth rate at axis aligned directions
- similarly, the gradient points to the greatest ascent direction

Example

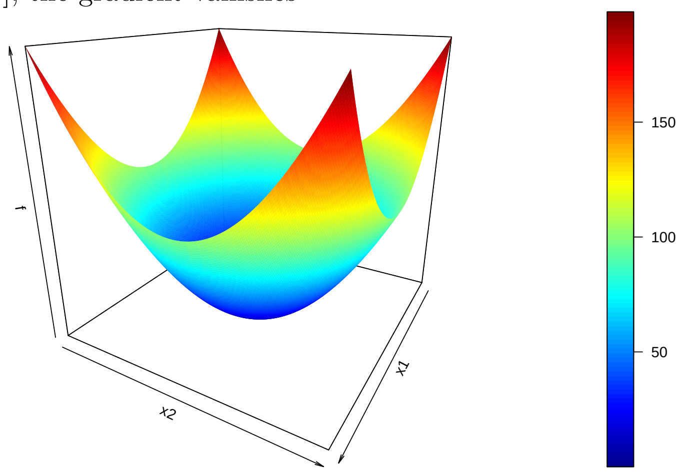
Find the gradient of $f(\boldsymbol{x}) = f(x_1, x_2) = x_1^2 + x_2^2$

$$\nabla_{\boldsymbol{x}} f = \left[\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2} \right] = [2x_1, 2x_2]$$

- gradient is a vector field: at any input point $[x_1, x_2]$, it gives a direction vector $[2x_1, 2x_2]$



- at $\boldsymbol{x} = [1, 1]$, the gradient vector is $[2, 2]$, pointing to steepest ascent direction
- at $\boldsymbol{x} = [1, 0]$, the gradient vector is $[2, 0]$
- at $\boldsymbol{x} = [0, 0]$, the gradient vanishes



Quadratic form xAx

Another view: $f(\boldsymbol{x}) = x_1^2 + x_2^2 = \boldsymbol{x}^T \boldsymbol{A}\boldsymbol{x}$, where $\boldsymbol{A}$ is $\boldsymbol{I}$ (*verify this!*)

- $\boldsymbol{x}\boldsymbol{A}\boldsymbol{x}$ is called quadratic form
- generalisation of quadratic function $f(x) = ax^2 = xax$
- the gradient is $\nabla_{\boldsymbol{x}} f = 2\boldsymbol{A}\boldsymbol{x}$ (when A is symmetric)
- the hessian matrix is $\nabla_{\boldsymbol{x}}^2 f = 2A$
  - f has a maximum if $\boldsymbol{A}$ is negative definite (a is negative)
  - f has a minimum if $\boldsymbol{A}$ is positive definite (a is positive)
- positive definite (P.D.) matrix $\boldsymbol{A}$: $\boldsymbol{x}\boldsymbol{A}\boldsymbol{x} > 0$ for all $x \in R^m$
  - $\boldsymbol{I}$ is P.D., why ?
  - f has a minimum
- negative definite (N.D.) matrix $\boldsymbol{A}$: $\boldsymbol{x}\boldsymbol{A}\boldsymbol{x} < 0$ for all $x \in R^m$

Example: connecting the dots

Given a bunch of numbers $\boldsymbol{x} = [3, 4, 3.5, 5, 6, 5.5]$, *to summarise the data set, sample mean*

$$\mu(\boldsymbol{x}) = \frac{1}{N}\sum_{i=1}^{N} x_i$$

is often used; why ?

- $(\mu - x_i)^2$ meansures the distance between x_i and μ
- $\sum_{i=1}^{N}(\mu - x_i)^2$ measures the total distance or cost
- note that $\mu \begin{bmatrix}1\\ \vdots \\ 1\end{bmatrix} - \begin{bmatrix}x_1\\ \vdots \\ x_n\end{bmatrix} = \begin{bmatrix}\mu - x_1\\ \vdots \\ \mu - x_n\end{bmatrix} \equiv \boldsymbol{e}$

$$\text{Let } f(\mu) = \sum_{i=1}^{N}(\mu - x_i)^2 = (\mu\mathbf{1} - \boldsymbol{x})^T(\mu\mathbf{1} - \boldsymbol{x}) = \boldsymbol{e}^T\boldsymbol{e} = \boldsymbol{e}^T\boldsymbol{I}\boldsymbol{e}$$

$$\frac{df}{d\mu} = \frac{\partial f}{\partial \boldsymbol{e}}\frac{d\boldsymbol{e}}{d\mu} = 2\boldsymbol{e}^T\mathbf{1} = 2(\mu\mathbf{1} - \boldsymbol{x})^T\mathbf{1}$$

set the derivative to 0

$$\mu\mathbf{1}^T\mathbf{1} - \boldsymbol{x}^T\mathbf{1} = 0 \Rightarrow \mu = \frac{\mathbf{1}^T\boldsymbol{x}}{\mathbf{1}^T\mathbf{1}} = \frac{\sum_i^N x_i}{N}$$

- can you tell it is actually a projection ?

- note that the projection of $\boldsymbol{x}$ on $\mathbf{1}$ is

$$P(\boldsymbol{x}, \mathbf{1}) = \frac{\mathbf{1}^T\boldsymbol{x}}{\mathbf{1}^T\mathbf{1}}\mathbf{1} = \mu\mathbf{1}$$

- this is a ML model actually (specific case)
- for this simple example, we have used
  - linear algebra
  - vector calculus
- how about probability theory ?

4 Probability theory

Probability theory

- Random variable

- Probability distribution
- Probability mass function and density function
- Probability rules
- Expectation, variance, covariance
- Conditional expectation

Random variable and probability distribution

Random variable X associates with a **probability distribution** $P(X)$

- formally, a r.v. is a mapping from sample space Ω to a target space $\mathcal{T}$
- e.g. toss a fair coin twice, r.v. X is the number of heads turned up
  - the sample space is $\Omega = \{HH, TT, HT, TH\}$
  - target space is $T = \{0, 1, 2\}$
  - the probability distribution is

$$P(X) = \begin{cases} 0.25 & X = 0 \\ 0.5 & X = 1 \\ 0.25 & X = 2 \end{cases}$$

- the distribution P must satisfy

$$P(X = x) > 0, \text{ and } \sum_{x \in T} P(X = x) = 1$$

Random variable - discrete r.v.

If r.v. X's target space $\mathcal{T}$ is discrete

- X is called **discrete random variable**
- the probability distribution P is called **probability mass function** (p.m.f.)
- and

$$0 \leq P(X = x) \leq 1, \text{ and } \sum_{x \in T} P(X = x) = 1$$

Example - discrete r.v.

Bernoulli distribution Tossing a coin , $\mathcal{T} = 1, 0$ (1 is H, 0 is T),

$$P(X = 1) = p, P(X = 0) = 1 - p, 0 \leq p \leq 1$$

Binomial distribution Tossing a coin N times, the r.v. X is the number of head shows up

$$P(X = k) = \binom{N}{k} \cdot p^k (1 - p)^{N-k}$$

- what's the relationship between Binomial and Bernoulli?

Multinoulli distribution Throw a fair 6-facet die, $\mathcal{T} = 1, 2, \ldots, 6$, the distribution is

$$P(X = i) = 1/6$$

Verify the above Ps satisfy the conditions of p.m.f.

Random variable - continuous r.v.

If r.v. X's target space $\mathcal{T}$ is continuous

- X is called **continuous random variable**
- the probability distribution p is called **probability density function** (p.d.f.)
- and satisfies

$$p(x) \geq 0, \text{ and } \int_{x \in T} p(x)dx = 1$$

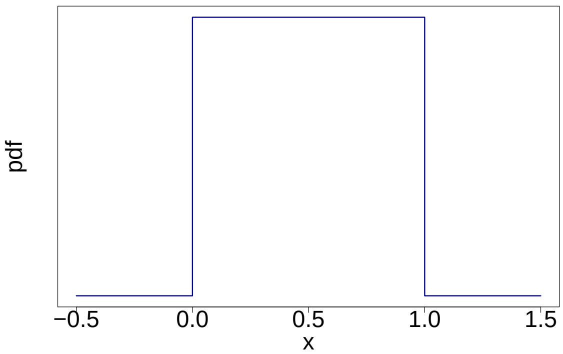
- pdf is not probability as $p(x)$ can be greater 1;
- for $\forall x \; P(X = x) = 0$
- calculate probability over an interval: e.g.

$$P(X \in [a, b]) = \int_a^b p(x)dx$$

Example - continuous r.v.

Uniform distribution $\mathcal{T} = [0, 1]$, X has equal chance to take any value between 0 and 1; the pdf is

$$p(x) = \begin{cases} 1 & x \in [0, 1] \\ 0 & \text{otherwise} \end{cases}$$



Easy to verify $\int_0^1 p(x)dx = \int_0^1 dx = 1$

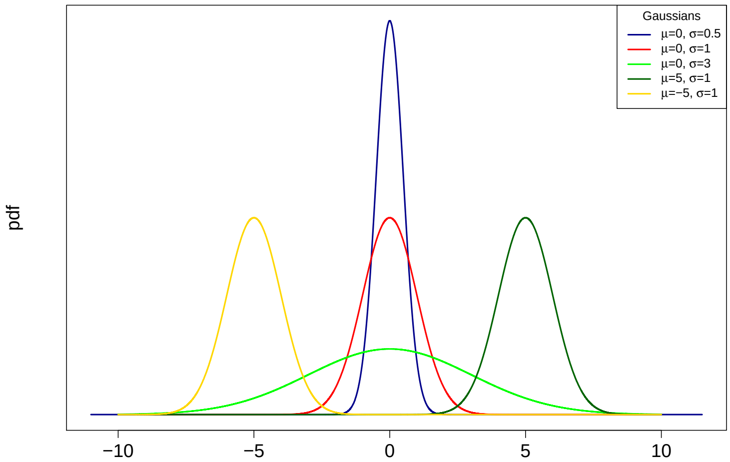
What's the probability that $0 < X < 0.5$?

Example - continuous r.v.

Gaussian distribution $\mathcal{T} = R$, or $X \in R$ the pdf is

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

$(\frac{x-\mu}{\sigma})^2$ is a distance measure: how far x is away from μ (measured by σ as a unit)



Question
Calculate quickly:

$$\int_{-\infty}^{\infty} e^{-\frac{1}{2}x^2} dx = ?$$

For $X \sim \mathcal{N}(\mu, \sigma)$, what is $P(X < \mu) = ?$

Joint distribution

- r.v. $\boldsymbol{X} = [X_1, X_2, \ldots, X_n]^T$ can be multidimensional (each X_i is r.v.)
  - essentially a *random vector*

- Still satisfies the same requirements

$$\forall \boldsymbol{x}, 0 < P(\boldsymbol{X} = \boldsymbol{x}) < 1, \sum_{x_1} \sum_{x_2} \ldots \sum_{x_n} P(\boldsymbol{X} = [x_1, x_2, \ldots, x_n]) = 1$$

  or

$$\forall \boldsymbol{x}, p(\boldsymbol{X} = \boldsymbol{x}) > 0, \int \int \ldots \int p(\boldsymbol{X} = \boldsymbol{x}) dx_1 dx_2 \ldots dx_n = 1$$

- for bivariate case, i.e. $n = 2$, X_1, X_2 are **independent** if $P(\boldsymbol{X}) = P(X_1)P(X_2)$ (e.g. rolling two dice independently)

Example: discrete joint distribution
The joint distribution of X snow or not, $Y \in \{\text{spring, summer, autumn, winter}\}$ represents the season that x belongs to :

It is easy to verify that

$$\sum_x \sum_y p(x, y) = 1$$

	$y =$ Spring	$y =$ Summer	$y =$ Autumn	$y =$ winter
$x = F$	0.05	0.25	0.075	0
$x = T$	0.2	0	0.175	0.25

	$y =$ Spring	$y =$ Summer	$y =$ Autumn	$y =$ winter
$x = F$	0.05	0.25	0.075	0
$x = T$	0.2	0	0.175	0.25

Example: continuous joint distribution

If X, Y's joint p.d.f is

$$p(x, y) = \frac{1}{2\pi\sigma_x\sigma_y} e^{-\frac{1}{2}[(\frac{x-\mu_x}{\sigma_x})^2 + (\frac{y-\mu_y}{\sigma_y})^2]}$$

X, Y are bivariate Gaussian distributed

X,Y are *independent* for this case, why ?

Probability rules

There are only two probability rules (integration for continuous r.v.):

1. product rule:
$$p(x, y) = p(y|x)p(x) = p(x|y)p(y)$$
2. sum rule (marginalisation):
$$p(x) = \sum_y p(x, y),\ p(y) = \sum_x p(x, y)$$

Conditional probability

Conditional probability distribution (by product rule):

$$p(x|y) = \frac{p(x, y)}{p(y)}$$

- probability distribution of x conditional on the value of y

- $P(Y = \text{Spring})$? use sum rule $P(Y = \text{Spring}) = \sum_{x=\{T,F\}} P(X = T, Y = \text{Spring}) = 0.05 + 0.2 = 0.25$
- $P(X = T|Y = \text{Spring})$? $P(x = T|y = \text{Spring}) = \frac{P(x=T,y=\text{Spring})}{P(y=\text{Spring})} = \frac{0.2}{0.25} = 0.8$

Expectation and variance

Expection of a r.v. is defined as

$$\mathbb{E}[g(X)] = \sum_x g(x)P(x) \text{ or } \mathbb{E}[g(X)] = \int g(x)P(x)dx$$

- $\mathbb{E}[a] = a$, a is a constant (not r.v.)
- $\mathbb{E}[\mathbb{E}[X]] = \mathbb{E}[X]$
- $\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$: linearity

Variance of a r.v. is defined as

$$\mathrm{Var}[g(X)] = \mathbb{E}[(g(X) - \mathbb{E}[g(X)])^2]$$

- $\mathrm{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$
- $\mathrm{Var}[aX] = a^2\mathrm{Var}[X]$

Prove them or convince yourself ! A very useful identity that links expectation and variance together is

$$\mathrm{Var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2,$$

which can be proved as follows:

$$\mathrm{Var}[x] = \mathbb{E}[x^2 - 2\mathbb{E}[x]x + \mathbb{E}[x]^2] = \mathbb{E}[x^2] - 2\mathbb{E}[x]^2 + \mathbb{E}[x]^2 = \mathbb{E}[x^2] - \mathbb{E}[x]^2;$$

the second equality holds because $\mathbb{E}[x] = \mu$ is a constant. Note that $\mathbb{E}[x^2]$ is called the *second moment* of r.v. x.

Example

X is a Bernoulli r.v. with parameter $p = 0.5$; what is $\mathbb{E}[X]$?

- $\mathbb{E}[X] = 1 \times P(X = 1) + 0 \times P(X = 0) = p = 0.5$;

Y is a Binomial r.v. with $N = 10, p = 0.5$, what is $\mathbb{E}[Y]$?

- $Y = \sum_{i=1}^{N} X = N \times X$
- $\mathbb{E}[Y] = \mathbb{E}[N \times X] = N \times \mathbb{E}[X] = N \times p = 5$
- interpretation: you expect to see 5 successes out of 10 (on average the result is 5 if you repeat the experiment a lot of times)

General advice on reading maths

> "The fish trap exists because of the fish. Once you've gotten the fish you can forget the trap. The rabbit snare exists because of the rabbit. Once you've gotten the rabbit, you can forget the snare. Words exist because of meaning. Once you've gotten the meaning, you can forget the words. Where can I find a man who has forgotten words so I can talk with him?" — Zhuang Zhou

- maths intuition is more important than equations

- think first then verify your beliefs by reading (different) text books or code it up and check

- stand higher and look at a greater picture
  - accepting intermediate results without fully understanding is totally cool