QBUS6850

Lecture 1

Machine Learning Introduction & Linear Algebra Review

BUSINESS SCHOOL

QBUS6850 Team

THE UNIVERSITY OF SYDNEY

- Topics covered
  - Why studying machine learning
  - Types of learning
  - Linear algebra and matrix computation review
- References
  - Chapter 1 (Alpaydin, 2014)

THE UNIVERSITY OF SYDNEY

Learning Objectives

- Be able to distinguish two major types of learning (supervised and unsupervised)
- Understand the notations of linear algebra
- Understand the basic operations of vectors/matrices, such as the inner product of two vectors, the norm of a vector, transpose of matrices, the matrix product, the product of matrices, matrix rank and determinant
- Understand concepts such as the vector norm or length, orthogonality of vectors and projecting a vector
- Be familiar with linear equations systems and matrix inverse

Why Study

Machine Learning?

InfoWorld FROM IDG

INSIDER Sign In | Register

IDG CONTRIBUTOR NETWORK Want to Join?

CHANGING DATA SCIENCE

By Vivian Zhang and Chris Neimeth, InfoWorld | MAR 5, 2018

Opinions expressed by ICN authors are their own.

Why data science and machine learning are the fastest growing jobs in the US

The US could have as many as 250,000 open data science jobs by 2024, and the data science skills gap will find companies scrambling to train or hire talent in the coming years

All is about Data and Analytics

- Every minute, Americans use 2,657,700GB of data.
- Every minute, Instagram users post 46,750 photos.
- Every minute, 15,220,700 texts are sent.
- Every minute, Google conducts 3,607,080 searches.

Linked in Official Blog Recent Posts Topics

The Fastest-Growing Jobs in the U.S. Based on LinkedIn Data

Rachel Bowley December 7, 2017

Share Tweet Like 301 G+

LinkedIn.com Email Subscription

The New Year is almost here and you might be exploring the idea of a new role that's

Guess please!

The top 10 emerging positions are:

1. Machine Learning Engineer (9.8X growth)
2. Data Scientist (6.5X)
3. Sales Development Representative (5.7X)
4. Customer Success Manager (5.6X)
5. Big Data Developer (5.5X)

6. Full Stack Engineer (5.5X)

7. Unity Developer (5.1X)
8. Director of Data Science (4.9X)

9. Brand Partner (4.5X)
10. Full Stack Developer (4.5X)

Our study also took a look at the most common skills among the top 20 emerging jobs. While it's key to have some technical chops for some of these, several soft skills that make the list as well.

LinkedIn.com

The full top 10 list includes:

Email Subscription

https://blog.linkedin.com/2017/december/7/the-fastest-growing-jobs-in-the-u-s-based-on-linkedin-data

Machine Learning for Business

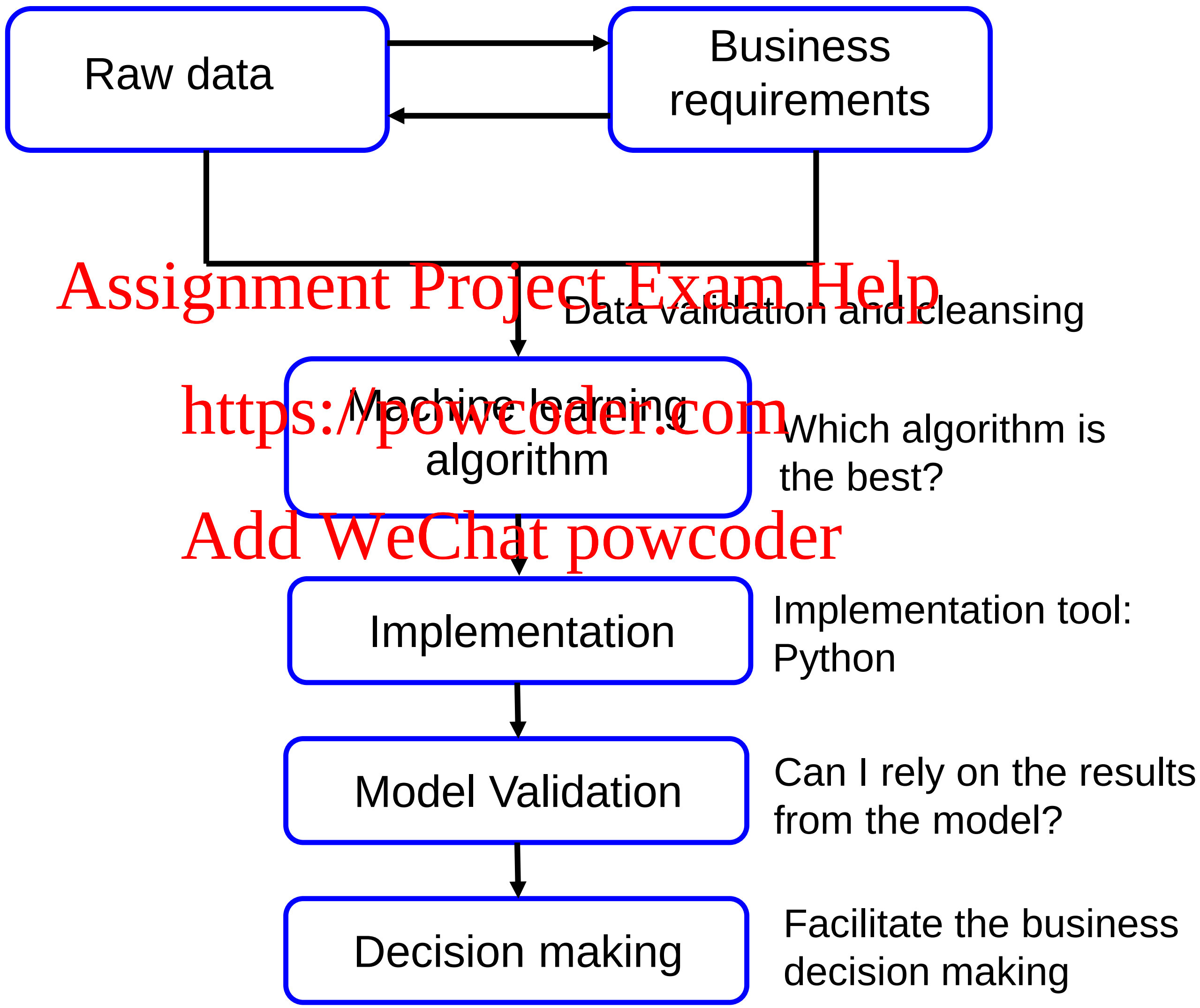
- Data-driven decisions are more profitable
  - Tradition: Decision-making relying on CEOs
  - New trend: Data-driven, allowing for nonpersonal decision-making

- Machine Learning is changing how we do business
  - The advanced algorithms saving time and resources
  - Mitigating risks with better decisions

- Machine Learning provides better forecasting
  - ML makes it possible to find hidden insights in the data
  - ML makes it possible to extract patterns from vast amounts of data

https://www.infoworld.com/article/3259891/data-science/why-data-science-and-machine-learning-are-the-fastest-growing-jobs-in-the-us.html

Machine Learning for Business



Applications

- Prediction: market demand prediction, price prediction
- Pattern Recognition
- Data Compression
- Outlier detection
- Recommendations

https://techcrunch.com/2016/03/19/how-real-businesses-are-using-machine-learning/

https://www.bloomberg.com/news/articles/2016-02-03/google-search-chief-singhal-to-retire-replaced-by-ai-manager

THE UNIVERSITY OF SYDNEY

Machine Learning Introduction

What is Machine Learning?

Arthur Samuel described **machine learning** it as: "the field of study that gives computers the ability to learn without being explicitly programmed."

Machine Learning aims to build algorithms that can learn from and make predictions on data, and is evolved study of pattern recognition and computational learning theory in artificial intelligence.

- Machine learning is using computers to analyse data.
- Wants the computers think like human.
- What is "learning"? Often, we do not want just to describe the data we have, but be able to predict (yet) unseen data.
- With new data, the "learning" can be refreshed/updated

What is Machine Learning?

Traditional Statistics: You have **a specific question** about population.
E.g. What's the Average Height of Australians?

- Expensive or impossible to collect data for entire population
- Collect a sample and use inference to say things about the feature of population you want
- Parameter = unknown feature of population of interest
- Estimator = Sample based estimate of a parameter

Current situation

Lots of data collected with **no specific questions** in mind.

Often, it would be quite easy to make a model that would describe already known data. It is more difficult to predict unseen data (generalization).

Types of Learning

Two major types of Learning

Supervised Learning:

- In supervised learning, an imaginary "supervisor" tells us in the training phase what is the correct response/target variable (t), given the feature (x).
- Dependent or outcome variable is given.
- t is distinguished from inputs x.
- Two major techniques here:
- Regression: t is quantitative, or continuous variable
- Classification: t is discrete or category variable
- Goal: prediction of t

Supervised Learning & Unsupervised Learning

Two major types of Learning

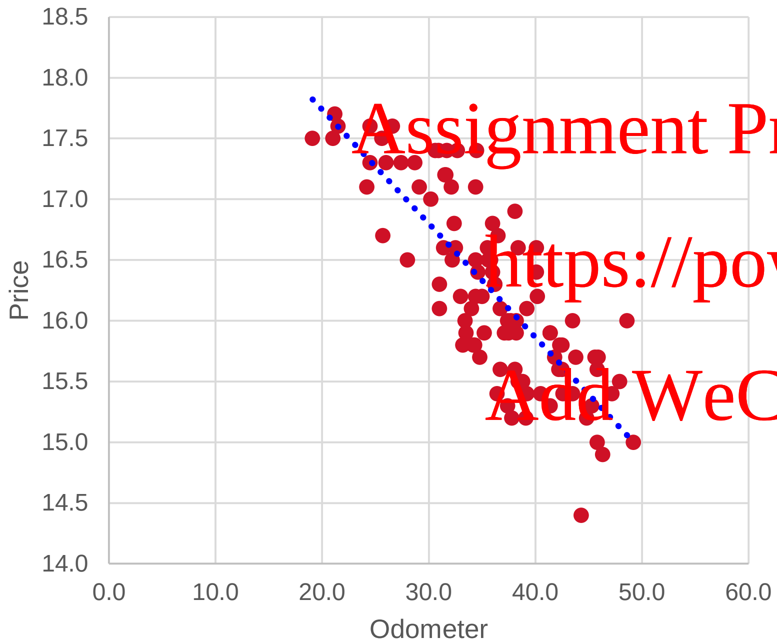
Unsupervised Learning:

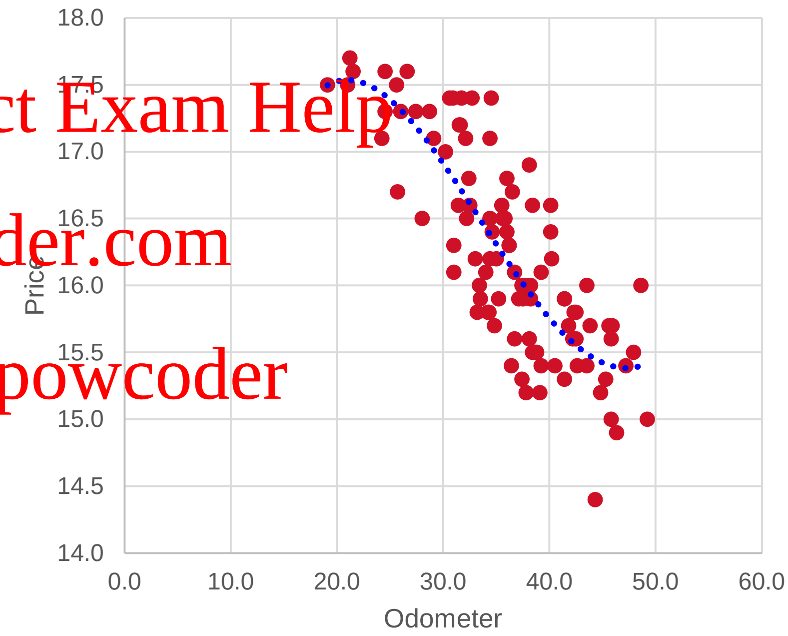
- In unsupervised learning we do not make the distinction between the response/target variable (t) and the feature (x)

- "Unlabelled" or "Unclassified" data
- Used to uncover hidden patterns, clusters, relationships or distribution, e.g. k-means clustering
- Goal: hypothesis f() generation, e.g. clustering rule, then to be tested in supervised learning

Regression

Is this a **supervised** or **unsupervised** learning example?





Response variable t is a continuous variable.

Example: price of second hand cars.

t: car price; x: odometer reading. $t = f(\mathbf{x}|\boldsymbol{\beta})$; $\mathbf{f}(\cdot)$: a model; $\boldsymbol{\beta}$: model parameters.

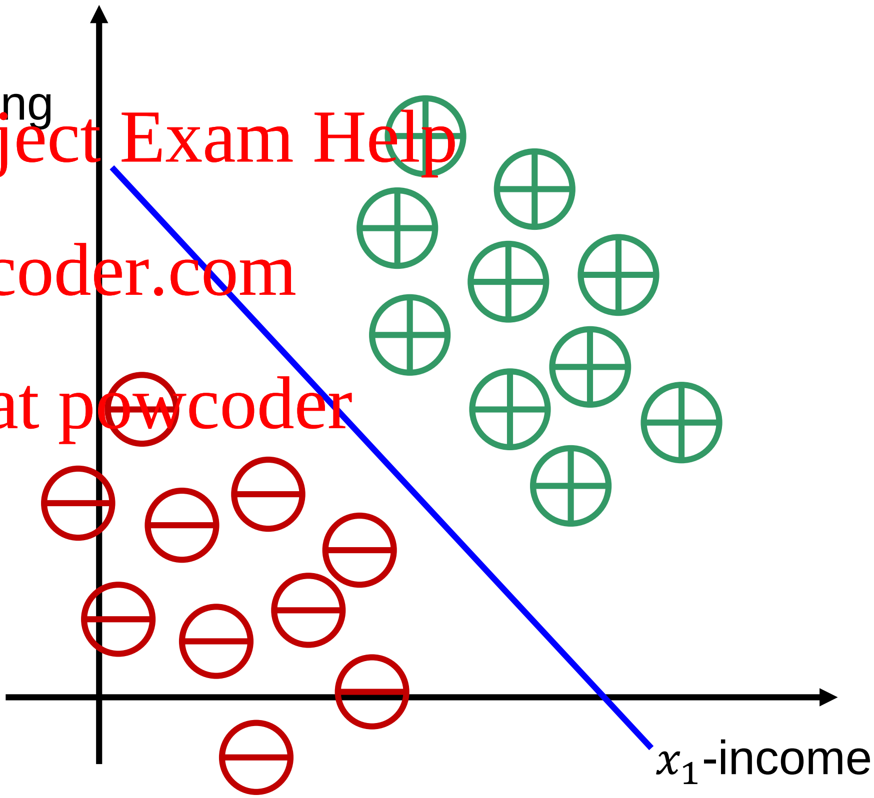
Classification

Is this a **supervised** or **unsupervised** learning example?

Example:
Data on credit card applicants

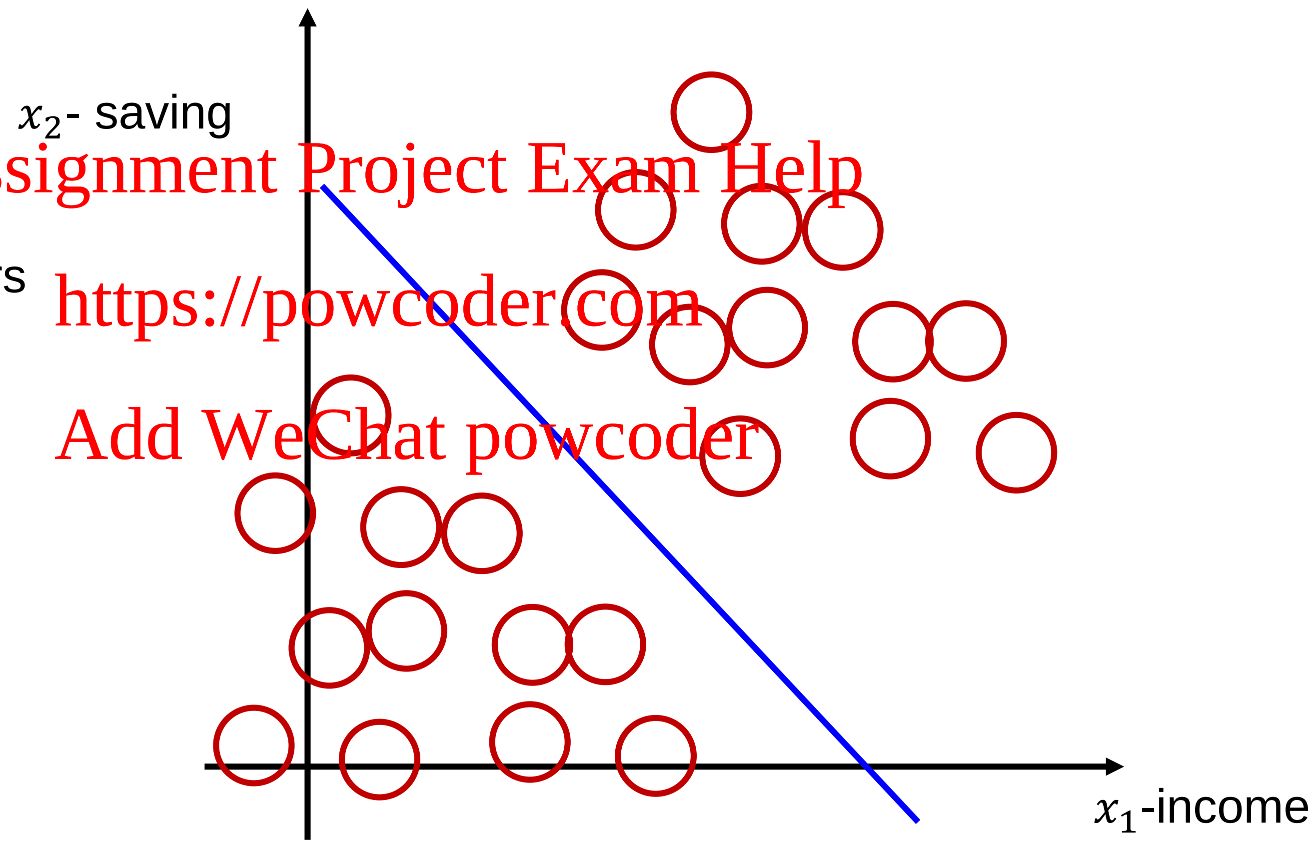
Question:
Should a new application be granted a credit card?

Data:
Low-risk (+) and high-risk (-) customers are labelled based their income and savings.



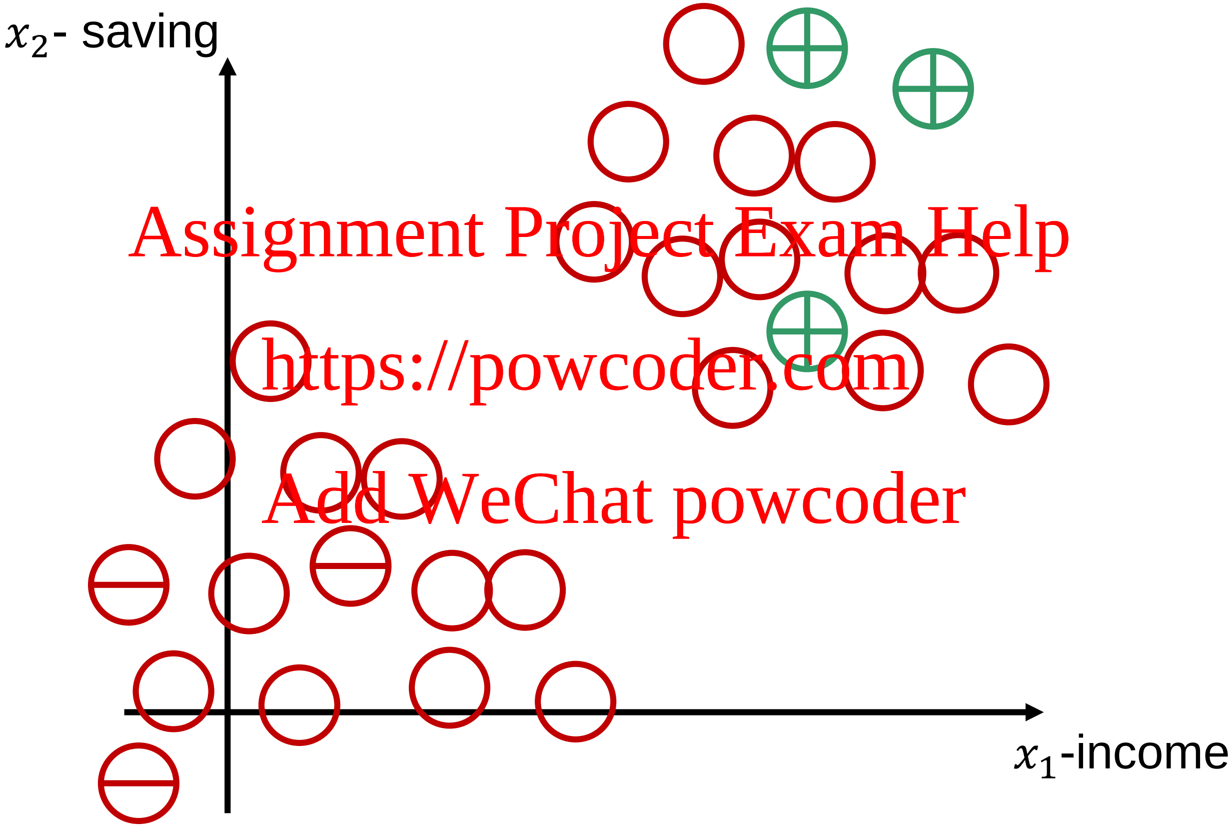
Clustering

Is this a **supervised** or **unsupervised** learning example?



Question:
Segment the customers risk levels based on income and saving

Data:
Labelled or not?



Semi-supervised learning

Datasets Sources

Often, finding a good data set is one of the most difficult tasks in developing machine learning methods.

Useful Links:

UCI Repository: http://www.ics.uci.edu/~mlearn/MLRepository.html

UCI KDD Archive: http://kdd.ics.uci.edu/summary.data.application.html

Delve: http://www.cs.utoronto.ca/~delve/

Linear Algebra

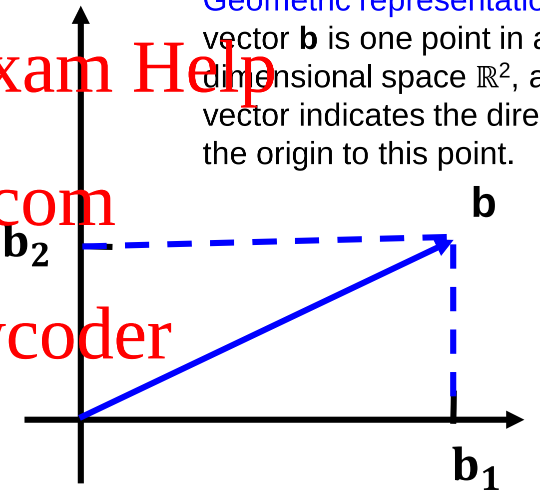
(please read it through at home)

Vector

A vector is a collection of numbers (scalars) ordered by column (or row). We assume vectors are of columns.

Vector **a** is one point in a n-dimensional space $\mathbb{R}^n$, with coordinates provided by the elements a_i

$$\mathbf{a} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_{n-1} \\ a_n \end{bmatrix} \qquad \mathbf{b} = \begin{bmatrix} 5 \\ 2 \end{bmatrix}$$



Geometric representation of **b:** vector **b** is one point in a 2-dimensional space $\mathbb{R}^2$, and the vector indicates the direction from the origin to this point.

The symbol $\mathbf{a}^T$ (**a**'s transpose) will be a row vector:

$$\mathbf{a}^T = [a_1, a_2, \ldots, a_{n-1}, a_n] \qquad \mathbf{b}^T = [5,2]$$

Understanding Vector

Special cases

Zeros vector: $\mathbf{0}_n = [0, 0, 0, \dots, 0]^T$;

- All n components are 0's

Why it called unit vector?

Unit vector: $\mathbf{e}_i = [0, 0, 1, \dots, 0]^T$;

- All components zeros except for the one at i_{th} position (=1)

Ones vector: $\mathbf{1}_n = [1, 1, 1, \dots, 1]^T$;

- All n components are 1's

Basic Operations of Vector

Equality of vectors:

a = b $\Longleftrightarrow a_i = b_i$ for **all** i = 1, 2, ..., n;

Multiplication by scalars:

let ρ denote a scalar; ρ**a** is the vector with elements $\{\rho a_i\}$. E.g., Let $\mathbf{a} =$ $[5, 2, 3]^T$, then 0.5**a** $= [0.5 * 5, 0.5 * 2, 0.5 * 3]^T = [2.5, 1, 1.5]^T$

Sum of two vectors:

Let **a** and **b** be two vectors with **the same size n**; their sum **x = a + b** is the vector with elements $c_i = a_i + b_i$, e.g. Let $\mathbf{a} = [5, 2, 3]^T$ and $\mathbf{b} = [1, -11, 2]^T$, then $\mathbf{c} = \mathbf{a} + \mathbf{b} = [5, 2, 3]^T + [1, -11, 2]^T = [6, -9, 5]^T$

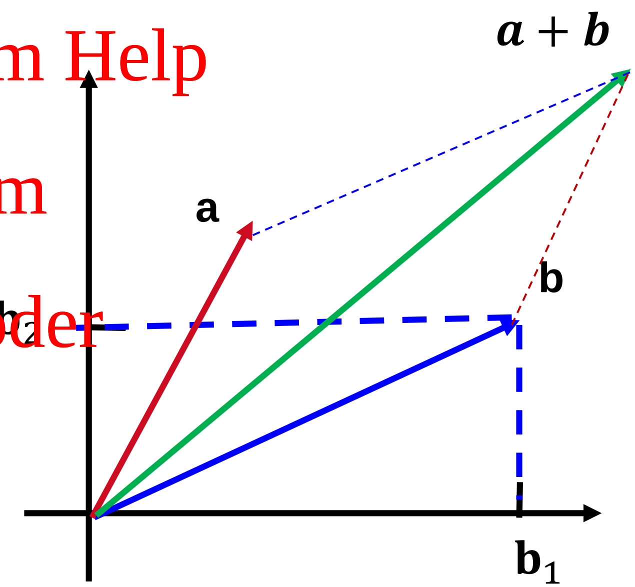
Basic Operations of Vector

Linear combination:

Let $\mathbf{a} = [1, 2]^T$ and $\mathbf{b} = [3, 1]^T$, and let ρ_1 and ρ_2 denote be two coefficients (scalars),

$$\rho_1 \mathbf{a} + \rho_2 \mathbf{b}$$

is their linear combination.



If $\rho_1 = 1$ and $\rho_2 = 1$, what is $\rho_1 \mathbf{a} + \rho_2 \mathbf{b}$?

If $\rho_1 = 3$ and $\rho_2 = 7$, what is $\rho_1 \mathbf{a} + \rho_2 \mathbf{b}$?

Geometric representation of sum of two vectors (parallelogram rule) and linear combination

https://en.wikipedia.org/wiki/Parallelogram

Vector Inner Product

The inner product between two n-dimensional vector **a** and **b** is defined as:

$$< \mathbf{a}, \mathbf{b} > = \mathbf{a}^T \mathbf{b} = \sum_{i=1}^{n} a_i b_i$$

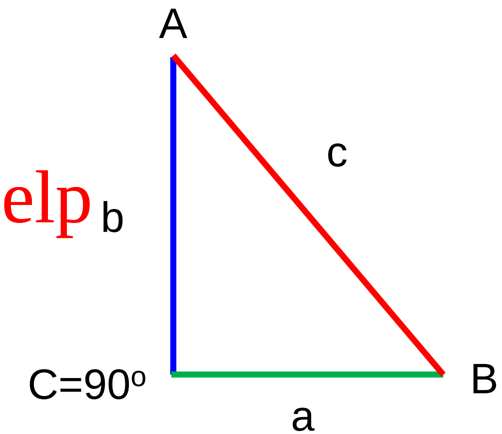
If $\mathbf{a} = [5, 2, 3]^T$ and $\mathbf{b} = [1, -11, 2]^T$, $< \mathbf{a}, \mathbf{b} > =?$

Properties

1. $(\rho \mathbf{a})^T \mathbf{b} = \rho(\mathbf{a}^T \mathbf{b})$
2. $\mathbf{a}^T \mathbf{b} = \mathbf{b}^T \mathbf{a}$
3. $\mathbf{a}^T (\mathbf{b} + \mathbf{c}) = \mathbf{a}^T \mathbf{b} + \mathbf{a}^T \mathbf{c}$

Vector Norm or Length

By Pythagoras theorem, the norm of vector **a** is the square root of the inner product of **a** with itself:



$$\| \mathbf{a} \| = \sqrt{\mathbf{a}^T \mathbf{a}} = \left(\sum_{i=1}^{n} a_i^2 \right)^{1/2} \geq 0$$

$$a^2 + b^2 = ?$$

This is the distance from the origin to the point a or the **length** of the vector. The (normalized) vector $\boldsymbol{a}/||\boldsymbol{a}||$ has **unit** length.

https://en.wikipedia.org/wiki/Pythagorean_theorem

Euclidean Distance and Orthogonality

The distance between the vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ is the norm of the difference vector $\mathbf{x}_i - \mathbf{x}_j$:

$$d_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\| = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^T(\mathbf{x}_i - \mathbf{x}_j)} = \left(\sum_{k=1}^{d}(x_{ik} - x_{jk})^2\right)^{1/2} \geq 0$$

Orthogonality: two vectors are orthogonal, $\mathbf{a} \perp \mathbf{b}$, if and only if their inner product is zero, $\mathbf{a}^{\mathrm{T}}\mathbf{b} = \mathbf{0}$.

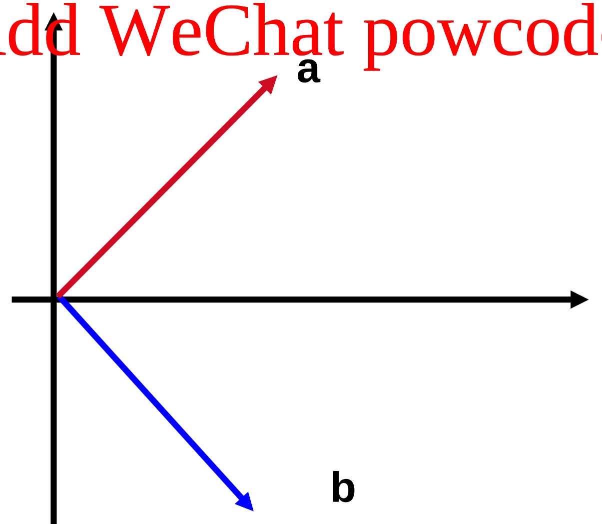
Example?

Geometric Representation

$$\mathbf{a} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \qquad \mathbf{b} = \begin{bmatrix} 1 \\ -1 \end{bmatrix} \qquad ||\mathbf{a}|| = ||\mathbf{b}|| = \sqrt{2} \qquad \mathbf{a}^{\mathrm{T}}\mathbf{b} = \mathbf{0}$$

$$\mathbf{c} = \begin{bmatrix} 1 \\ 3 \end{bmatrix} \qquad \mathbf{d} = \begin{bmatrix} 0.6 \\ -0.2 \end{bmatrix} \qquad ||\boldsymbol{c}|| = ? \quad ||\mathbf{d}|| = ? \qquad \mathbf{c}^{\mathrm{T}}\mathbf{d} = ?$$

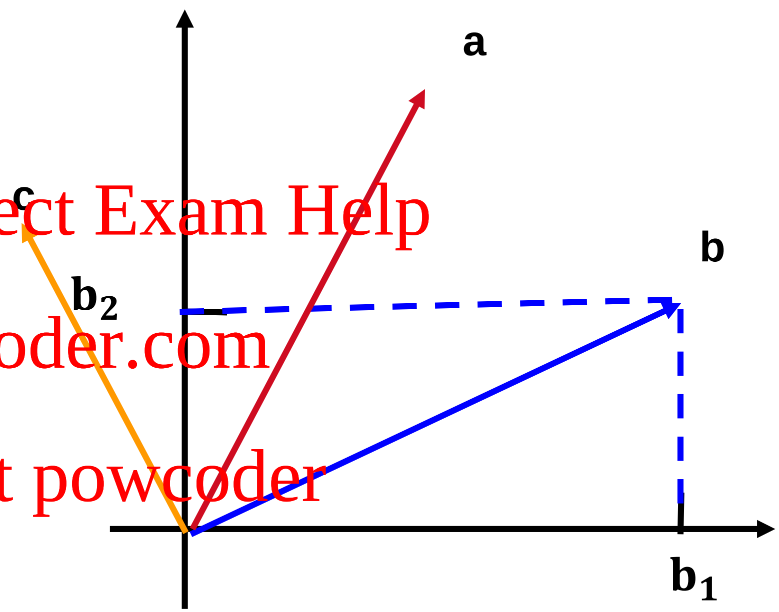
Orthogonality?



Inner Product Geometric Interpretation

$$\|\mathbf{a}\| = \sqrt{\mathbf{a}^T\mathbf{a}} = \left(\sum_{i=1}^{n} a_i^2\right)^{1/2} \geq 0$$

$$\mathbf{a} = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} \qquad \mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$$



Suppose we have a vector **c** that is orthogonal to **b,**

by the parallelogram law, $\mathbf{a} = \mathbf{c} + \rho\mathbf{b}$.

ρ here is a scala.

Inner Product Geometric Interpretation

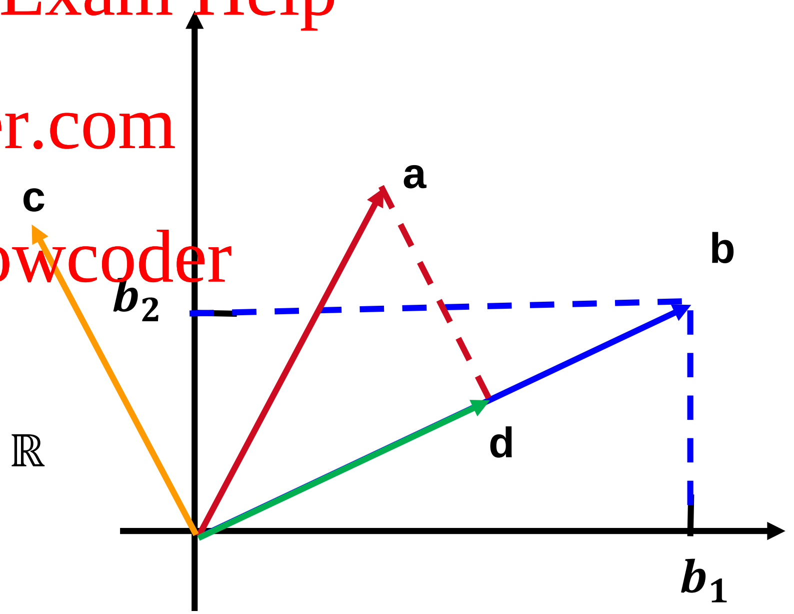
Vector $\mathbf{d} = \rho \boldsymbol{b}$ is called the orthogonal projection of **a** onto **b.**

Connecting to inner product

$$\mathbf{a} = \mathbf{c} + \rho\mathbf{b}$$

$$\mathbf{a}^{\boldsymbol{T}} = \mathbf{c}^{\boldsymbol{T}} + \rho\mathbf{b}^{T}$$

$$\mathbf{a}^{\boldsymbol{T}}\mathbf{b} = \mathbf{c}^{\boldsymbol{T}}\mathbf{b} + \rho\mathbf{b}^{T}\mathbf{b} = \mathbf{0} + \rho||\boldsymbol{b}||^2 = \rho||\boldsymbol{b}||^2 \in \mathbb{R}$$



Example

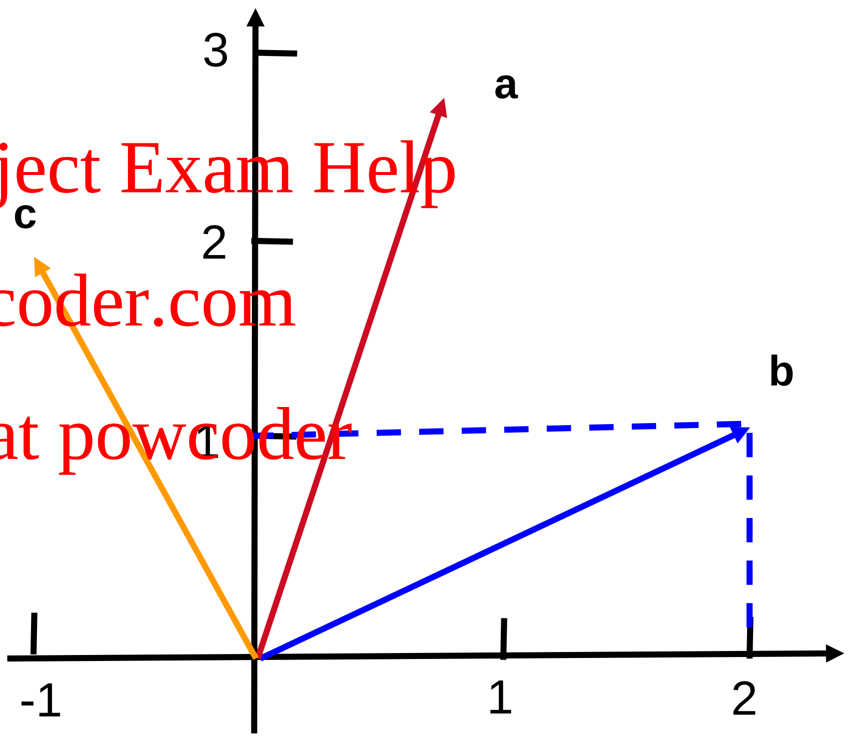
$$\mathbf{a} = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} \quad \mathbf{b} = \begin{bmatrix} 2 \\ 1 \end{bmatrix} \quad \mathbf{c} = \begin{bmatrix} -1 \\ 2 \end{bmatrix}$$

Vector **c** is orthogonal to **b**.
How to check orthogonality?

Suppose $\rho = 0.7$;

$$\mathbf{a} = \mathbf{c} + \rho\mathbf{b}$$

$$\mathbf{a} = \begin{bmatrix} -1 + 0.7 * 2 \\ 2 + 0.7 * 1 \end{bmatrix} = \begin{bmatrix} 0.4 \\ 2.7 \end{bmatrix}$$



Vector $\mathbf{d} = \rho\mathbf{b}$ is called the orthogonal projection of **a** onto **b.**

$\rho = 0.7$ $\quad \mathbf{b} = \begin{bmatrix} 2 \\ 1 \end{bmatrix} \Rightarrow \mathbf{d} = \begin{bmatrix} 1.4 \\ 0.7 \end{bmatrix}$ $\quad ||\mathbf{d}|| = \sqrt{1.4^2 + 0.7^2} = 1.565$

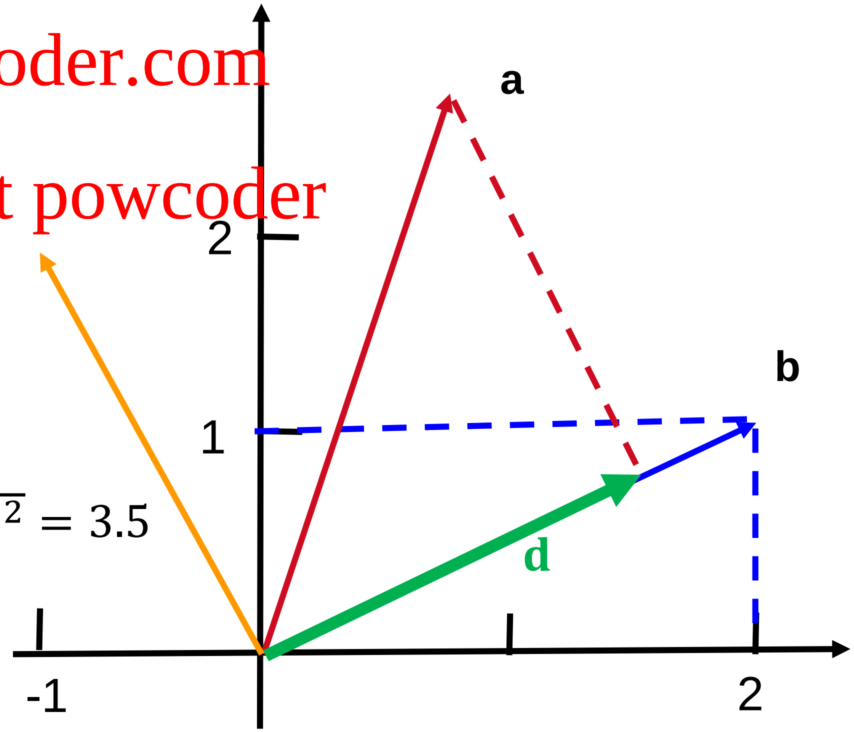
$||\mathbf{d}|| = \rho||\boldsymbol{b}||$

Let's test $\mathbf{a}^T\mathbf{b} = \rho||\boldsymbol{b}||^2$

$\mathbf{a}^T\mathbf{b} = 0.4 * 2 + 2.7 * 1 = 3.5$

$\mathbf{a}^T\mathbf{b} = \rho||\boldsymbol{b}||^2 = 0.7 * \sqrt{2^2 + 1^2}^2 = 3.5$

$\mathbf{a}^T\mathbf{b} = ||\mathbf{d}||* \, ||\mathbf{b}|| = \sqrt{1.4^2 + 0.7^2} * \sqrt{2^2 + 1^2} = 3.5$



Will be used in SVM later

Basic Calculation of Matrix

A matrix is a rectangular array of numbers (scalars) for which operations such as addition and multiplication are defined. It is a rectangular ($N \times d$) or two-dimensional array of scalars (numbers), represented as:

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1j} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2j} & \cdots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ x_{i1} & x_{i2} & \cdots & x_{ij} & \cdots & x_{id} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{Nj} & \cdots & x_{Nd} \end{bmatrix}$$

$\mathbb{R}^{N \times d}$

Can also write:

$$\mathbf{X} = [x_{ij}]$$

In a typical data matrix, the index $i = 1, 2, \ldots, N$ refers to the statistical units/training examples, and the index $j = 1, 2, \ldots, d$ to the variables or **features**.

Matrix

Square matrix is a matrix with the same number of row and column numbers.

$$\mathbf{a} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{22} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix} \qquad \mathbb{R}^{4*4}$$

We can represent **X** as a partitioned matrix whose generic block is the $1 \times d$ row vector

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_i^T \\ \vdots \\ \mathbf{x}_N^T \end{bmatrix}$$

Each row is a $1 \times d$ row vector

originally $\mathbf{x_i}$ is a column vector

Or we can partition as below, where each column is a $N \times 1$ column vector

$$\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_j, \dots, \mathbf{x}_d]$$

A column vector of size N can be represented as $N \times 1$ matrix

A row vector of size d can be represented as $1 \times d$ matrix.

Matrix Transpose

Matrix transpose: transposition yields the $N \times d$ matrix with rows and columns interchanged

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1j} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2j} & \cdots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ x_{i1} & x_{i2} & \cdots & x_{ij} & \cdots & x_{id} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{Nj} & \cdots & x_{Nd} \end{bmatrix}$$

$$\mathbf{X}^T = \begin{bmatrix} x_{11} & x_{21} & \cdots & x_{i1} & \cdots & x_{N1} \\ x_{12} & x_{22} & \cdots & x_{i2} & \cdots & x_{N2} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ x_{1j} & x_{2j} & \cdots & x_{ij} & \cdots & x_{Nj} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{1d} & x_{2d} & \cdots & x_{id} & \cdots & x_{Nd} \end{bmatrix}$$

THE UNIVERSITY OF SYDNEY

Matrix Product

Let **A** be an $N \times p$ matrix whose i_{th} row is the $1 \times p$ vector $\mathbf{a}_i^T$,
Let **B** be an $p \times d$ matrix whose j_{th} column is the $p \times 1$ vector $\mathbf{b}_j$, so that

$$A = \begin{bmatrix} \mathbf{a}_1^T \\ \mathbf{a}_2^T \\ \vdots \\ \mathbf{a}_i^T \\ \vdots \\ \mathbf{a}_N^T \end{bmatrix}$$

$$\mathbf{B} = [\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_j, \dots, \mathbf{b}_d]$$

The matrix product **C = AB**, where **A** pre-multiplies **B**, is the $N \times d$ matrix with elements

$$c_{ij} = \mathbf{a}_i^T \mathbf{b}_j = \sum_{k=1}^{p} a_{ik} b_{kj}, \qquad i = 1, \dots, N; j = 1, \dots, d$$

Matrix Product

How matrix product works?

$$\mathbf{A} = \begin{bmatrix} 3 & 1 \\ 7 & 0 \\ 1 & 2 \\ 5 & 3 \end{bmatrix} \qquad \mathbf{B} = \begin{bmatrix} \boxed{\begin{matrix}1\\5\end{matrix}} & \boxed{\begin{matrix}3\\2\end{matrix}} & \boxed{\begin{matrix}-9\\6\end{matrix}} \end{bmatrix}$$

The number of columns of **A** is 2 and the number of rows of **B** is 2. We can do product **C** = **AB** which has 4 rows (= **A**'s row number) and 3 columns (= **B**'s column number):

$$\mathbf{C} = \begin{bmatrix} \boxed{\begin{matrix}c_{11}\\c_{21}\\c_{31}\\c_{41}\end{matrix}} & \boxed{\begin{matrix}c_{12}\\c_{22}\\c_{32}\\c_{42}\end{matrix}} & \boxed{\begin{matrix}c_{13}\\c_{23}\\c_{33}\\c_{43}\end{matrix}} \end{bmatrix}$$

C_{11} is the inner product of first row of **A** and first column of **B**

$$\mathbf{a}_1 = \begin{bmatrix} 3 \\ 1 \end{bmatrix} \qquad \mathbf{b} = \begin{bmatrix} 7 \\ 0 \end{bmatrix}$$

$$c_{11} = \mathbf{a}_1^T \mathbf{b}_1 = 3\times 1 + 1\times 5 = 8$$

C_{32} =?

Properties of Matrix

Not any two matrices have a product. You must make sure that the number of columns of the first matrix is **EQUAL** to the number of rows of the second matrix. Hence in general Matrix product is not commutative: **AB** is unequal to **BA**;

So in previous example: **BA** is not defined.

We do have **AB**=**BA** for some appropriate matrices **A** and **B**.

$$\left(\mathbf{A}^T\right)^T = \mathbf{A}; \quad \left(\mathbf{A}+\mathbf{B}\right)^T = \mathbf{A}^T + \mathbf{B}^T;$$

$$\left(\mathbf{AB}\right)^T = \mathbf{B}^T\mathbf{A}^T; \quad \left(\mathbf{AB}\right)\mathbf{C} = \mathbf{A}\left(\mathbf{BC}\right);$$

$$\mathbf{A}\left(\mathbf{B}+\mathbf{C}\right) = \mathbf{AB}+\mathbf{AC}; \quad \left(\mathbf{A}+\mathbf{B}\right)\mathbf{C} = \mathbf{AC}+\mathbf{BC}.$$

If **A** is an $\boldsymbol{m} \times \boldsymbol{p}$ matrix, notice the difference between $\mathbf{A}^{\mathbf{T}}\mathbf{A}$ ($\boldsymbol{p} \times \boldsymbol{p}$ matrix of crossproducts) and $\mathbf{A}\mathbf{A}^{\mathbf{T}}$ (size $\boldsymbol{m} \times \boldsymbol{m}$).

Matrix Special Cases

A square matrix has row number equals to the column number: $\boldsymbol{N} = \boldsymbol{d}$

A square matrix **A** is **symmetric** if $\mathbf{A}^{\mathrm{T}} = \mathbf{A}$

Diagonal matrix: a **square matrix** with all zeros on the nondiagonal positions

Can you have Diagonal matrix if N is unequal to d?

$$D = \begin{bmatrix} d_1 & 0 & \dots & 0 & 0 \\ 0 & d_2 & \ddots & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & d_{N-1} & 0 \\ 0 & 0 & \dots & 0 & d_N \end{bmatrix} = \mathrm{diag}(d_1, d_2, \dots, d_{\mathrm{N}-1}, d_N)$$

Identity matrix $(\mathbf{I}_N)$ of order N is a diagonal matrix with all $d_i = 1$

Properties of Matrix

If **A** is $N \times d$, then $$\mathbf{I}_N\mathbf{A} = \mathbf{A} \text{ and } \mathbf{A}\,\mathbf{I}_d = \mathbf{A}$$

Scalar matrix: $\rho\mathbf{I}_d$

Quadratic form: Let **A** be an d dimensional **symmetric square** matrix and **x** be an $d \times 1$ vector. Below scalar is called a quadratic form.

$$\mathbf{x}^T\mathbf{A}\mathbf{x}$$

$\mathbf{x}^T\mathbf{A}\mathbf{x} \geq 0 \Longrightarrow$ Semi-positive (nonnegative) definite

$\mathbf{x}^T\mathbf{A}\mathbf{x} > 0 \Longrightarrow$ Positive definite

Examples?

Outer product: If **x** is an $N \times 1$ vector and **y** is an $d \times 1$ vector, the outer product $\mathbf{xy}^\mathbf{T}$ is an $N \times d$ matrix.

Matrix Mean Vector

Let $\boldsymbol{X} \in \mathbb{R}^{N \times d}$ be a matrix of size $N \times d$ and $\boldsymbol{i}_N = [1, 1, \dots, 1]^T$ be $N \times 1$ vector of all ones.

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1j} & \dots & x_{1d} \\ x_{21} & x_{22} & \dots & x_{2j} & \dots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ x_{i1} & x_{i2} & \dots & x_{ij} & \dots & x_{id} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \dots & x_{Nj} & \dots & x_{Nd} \end{bmatrix}$$

d features, each has
N training examples

$$\mathbf{X}^T = \begin{bmatrix} x_{11} & x_{21} & \dots & x_{i1} & \dots & x_{N1} \\ x_{12} & x_{22} & \dots & x_{i2} & \dots & x_{N2} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ x_{1j} & x_{2j} & \dots & x_{ij} & \dots & x_{Nj} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{1d} & x_{2d} & \dots & x_{id} & \dots & x_{Nd} \end{bmatrix}$$

$d \times 1$

$d \times N$

$$\bar{\mathbf{X}} = \frac{1}{N} \mathbf{X}^T \mathbf{i}_N = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_j \\ \vdots \\ \bar{x}_d \end{bmatrix}$$

$N \times 1$

Mean of feature 1

Mean of feature 2

$1 \times d$

$$\bar{\mathbf{X}}^T = \frac{1}{N} \mathbf{i}_N^T \mathbf{X} = [\bar{x}_1, \bar{x}_2, \dots . \bar{x}_j, \dots, \bar{x}_d]$$

Sample Variance-Covariance Matrix

Suppose **X** is a matrix of size $N \times d$: N training examples and d features

$$\mathbf{S} = \begin{bmatrix} s_1^2 & s_{12} & \cdots & s_{1j} & \cdots & s_{1d} \\ s_{21} & s_2^2 & \cdots & s_{2j} & \cdots & s_{2d} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ s_{j1} & s_{j2} & \cdots & s_j^2 & \cdots & s_{jd} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ s_{d1} & s_{d2} & \cdots & s_{dj} & \cdots & s_d^2 \end{bmatrix}$$

Is **S** symmetric?

$S_{12} =?$

Mean of feature j

$$s_j^2 = \frac{1}{N-1} \sum_{i=1}^{N} \left(x_{ij} - \boxed{\bar{x}_j}\right)^2,$$

Feature j variance

$$s_{jk} = \frac{1}{N-1} \sum_{i=1}^{N} \left(x_{ij} - \bar{x}_j\right)\left(x_{ik} - \bar{x}_k\right)$$

Sample covariance between feature j and k

Matrix Rank

In linear algebra the rank of a matrix **A** is the dimension of the vector space generated (or spanned) by its columns. This is the same as the dimension of the space spanned by its rows.

The column and row rank are coincident and so we can define **the rank of the matrix as the maximum number of linearly independent** vectors (those forming either the rows or the columns) and denote it by r(**A**). Obviously **r(A) ≤ min(N, d)**.

$$\mathbf{A} = \begin{bmatrix} 1 & 4 & 7 \\ 2 & 5 & 8 \\ 3 & 6 & 9 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 4 & 7 \\ 2 & 5 & 8 \\ 3 & 6 & 9 \end{bmatrix} \rightarrow -3r_1 + r_3 \rightarrow \begin{bmatrix} 1 & 4 & 7 \\ 2 & 5 & 8 \\ 0 & -6 & -12 \end{bmatrix} \rightarrow -2r_1 + r_2 \rightarrow \begin{bmatrix} 1 & 4 & 7 \\ 0 & -3 & -6 \\ 0 & -6 & -12 \end{bmatrix} \rightarrow -2r_2 + r_3 \rightarrow \begin{bmatrix} 1 & 4 & 7 \\ 0 & -3 & -6 \\ 0 & 0 & 0 \end{bmatrix}$$

Determinant

If **A** is $N \times N$, its determinant, det(**A**) or |**A**|, is a **scalar**, whose absolute value measures the volume of the parallelogram delimited in d-dimensional space by the columns of **A**.

For the identity matrix $$|\mathbf{I}_N| = 1$$

For the diagonal matrix $$|\mathbf{D}| = d_1 d_2 \ldots d_N = \prod_{n=1}^{N} d_n$$

Moreover, if ρ is a scalar $$|\rho\mathbf{D}| = (\rho d_1)(\rho d_2)\ldots(\rho d_N) = \rho^N |\mathbf{D}|$$

- If the columns (rows) of A are linearly dependent, so that rank(A) < N, then |**A**| = 0;
- |**AB**| = |**A**| · |**B**|;
- $|\mathbf{A}^T| = |\mathbf{A}|$.

THE UNIVERSITY OF SYDNEY

Matrix Determinant

The general expression for the determinant is the following Laplace (cofactor) expansion

$$|\mathbf{A}| = \sum_{j=1}^{N} a_{ij}(-1)^{i+j}|\mathbf{A}_{ij}|, \qquad \text{for any fixed } i = 1, 2, \dots, N$$

where $\mathbf{A}_{\mathrm{ij}}$ is the submatrix obtained from $\mathbf{A}$ by removing the i_{th} row and the j_{th} column; $|\mathbf{A}_{\mathrm{ij}}|$ is called a minor of $\mathbf{A}$ and $(-1)^{i+j}|\mathbf{A}_{\mathrm{ij}}|$ is called cofactor.

$$\mathbf{A} = \begin{bmatrix} 1 & 7 \\ -5 & 2 \end{bmatrix}$$

$$\begin{aligned} |\mathbf{A}| &= 1 \times (-1)^{1+1} \times |2| + 7 \times (-1)^{1+2} \times |-5| \\ &= 1 \times 2 + 7 \times (-1) \times (-5) = 2 + 35 = 37 \end{aligned}$$

Matrix (3×3) Determinant

$$\mathbf{a} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

$$|\mathbf{a}| = \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = a_{11}\begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} - a_{12}\begin{vmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{vmatrix} + a_{13}\begin{vmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{vmatrix}$$

The **determinant** is only defined for **square matrices.** For non-square matrices, there's no determinant value.

Matrix Trace

The trace of a square matrix is the sum of its **diagonal** elements. If **A** is $N \times N$,

$$\mathrm{tr}(\mathbf{A}) := \sum_{i=1}^{N} a_{ii}$$

$$\mathbf{tr}(\rho\mathbf{A}) = \rho\,\mathbf{tr}(\mathbf{A})$$

$$\mathbf{tr}(\mathbf{A}+\mathbf{B}) = \mathbf{tr}(\mathbf{A}) + \mathbf{tr}(\mathbf{B})$$

$$\mathbf{tr}(\mathbf{A}^{T}) = \mathbf{tr}(\mathbf{A})$$

$$\mathbf{tr}(\mathbf{AB}) = \mathbf{tr}(\mathbf{BA})$$

Linear Equations Systems

Consider the system of n linear equations in n unknown, where **A** is a known $\boldsymbol{n} \times \boldsymbol{n}$ coefficients matrix and b a known $\boldsymbol{n} \times \mathbf{1}$ vector:

$$\mathbf{Ax} = \mathbf{b}$$

A non homogeneous system admits a unique solution if and only if |**A**| is unequal to 0, or equivalently rank(A) = n. In such case, the solution can be written as

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$$

3 unknowns; 3 equations

$$\begin{cases} -x_1 + 2x_2 + 4x_3 = 1 \\ 2x_1 - 2x_2 + 3x_3 = -0.5 \\ 3x_1 + 0.7x_2 - 5x_4 = 1.3 \end{cases} \Longrightarrow \begin{bmatrix} -1 & 2 & 4 \\ 2 & -2 & 3 \\ 3 & 0.7 & -5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ -0.5 \\ 1.3 \end{bmatrix}$$

Matrix Inverse

Let **A** be a square matrix of dimension n with full rank: rank(**A**) = n.

The inverse matrix is the matrix **X** which when pre-multiplied or post-multiplied by **A** returns the identity matrix

$$\mathbf{XA} = \mathbf{I}_n, \qquad \mathbf{AX} = \mathbf{I}_n \qquad\qquad \mathbf{A} = \mathbf{X}^{-1}, \qquad \mathbf{X} = \mathbf{A}^{-1}$$

If such **X** exists, then it is unique. We can write $\mathbf{X} = \mathbf{A}^{-1}$, called the inverse of **A**.

For a diagonal matrix, the computation of the inverse is immediate

$$D^{-1} = \begin{bmatrix} 1/d_1 & 0 & \dots & \cdots & 0 & 0 \\ 0 & 1/d_2 & \cdots & \dots & 0 & 0 \\ 0 & 0 & \ddots & \cdots & \vdots & \vdots \\ \vdots & \vdots & \cdots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \cdots & 0 & 1/d_{n-1} & 0 \\ 0 & 0 & \cdots & \cdots & 0 & 1/d_n \end{bmatrix}$$

Matrix Inverse

Example

We now illustrate the 2 × 2 case. From the definition of an inverse, $\mathbf{AX} = \mathbf{I_2}$, it follows

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

This yields a system of 4 equations in 4 unknowns:

$$a_{11}x_{11} + a_{12}x_{21} = 1$$

$$a_{11}x_{12} + a_{12}x_{22} = 0$$

$$a_{21}x_{11} + a_{22}x_{21} = 0$$

$$a_{21}x_{12} + a_{22}x_{22} = 1$$

Matrix Inverse

Recall matrix determinant

$$\mathbf{X} = \mathbf{A}^{-1} = \frac{1}{|\mathbf{A}|}\mathbf{A}^{*} = \frac{1}{a_{11}a_{22} - a_{12}a_{21}} \begin{bmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{bmatrix}$$

where $\mathbf{A}^{*}$ is known as the adjoint matrix of $\mathbf{A}$, with elements given by the cofactors of $\mathbf{A}$, e.g.,

$$a_{ji}^{*} = (-1)^{i+j}\,|\mathbf{A}_{ij}|$$

If |**A**|=0 or rank(A) < n, then its inverse does not exist.