Mathematical note

Personal note

FPT University

Compiled by

Nguyễn Đức Trọng

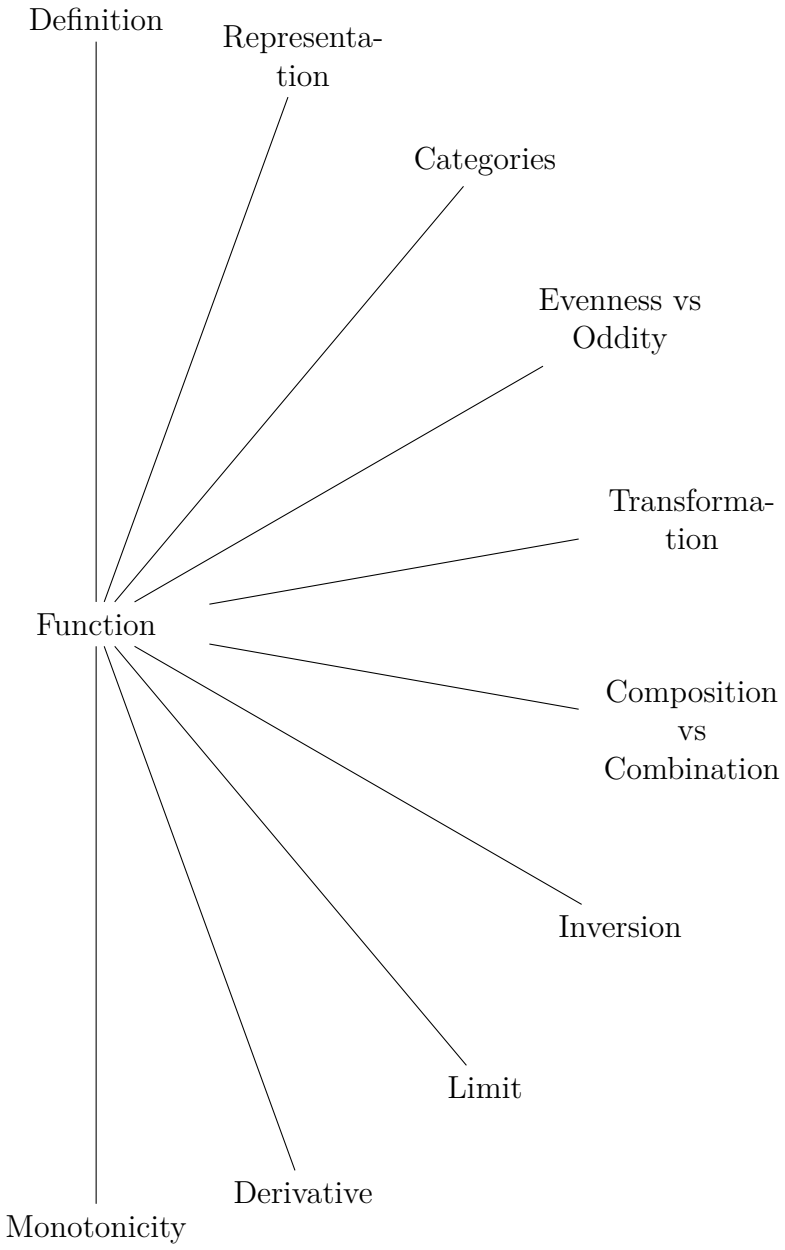
November 7, 2022

Signature

Contents

A Univariate Calculus

I Overview



II Definition

Formal definition: A function relates each element of a set with exactly one element of another set.

Informal definition: A specific calculation rule will yield what we want for inputs.

Notation:
f: X $\mapsto$ Y

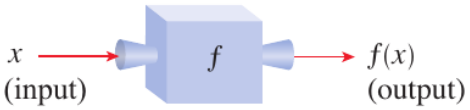


Figure 1: Machine diagram for a function f

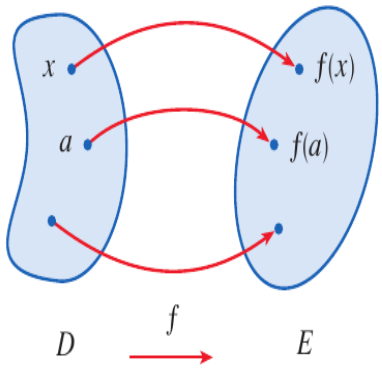


Figure 2: A mapping f from set D to set E

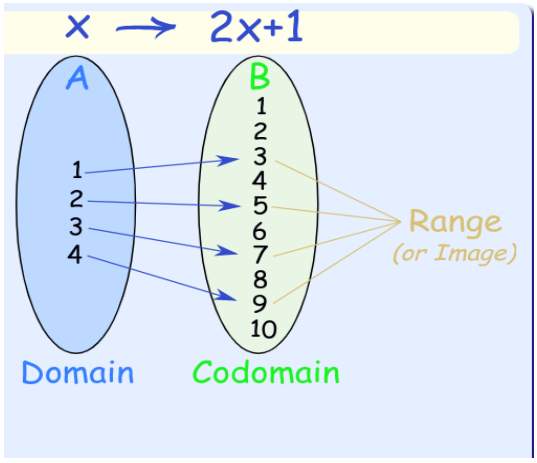


Figure 3: Some relevant terms

Notes:
+ Function domain (tập xác định): What goes into the function.
+ Function range/ image (tập miền, ảnh): What is returned by the function.
+ Codomain: What may possibly come out from the function.

Examples:
a/ $y = f(x) = 5x + 7$
b/ $y = f(x) = 99x^2 + 100x + 9999 + 7\log_2 5$

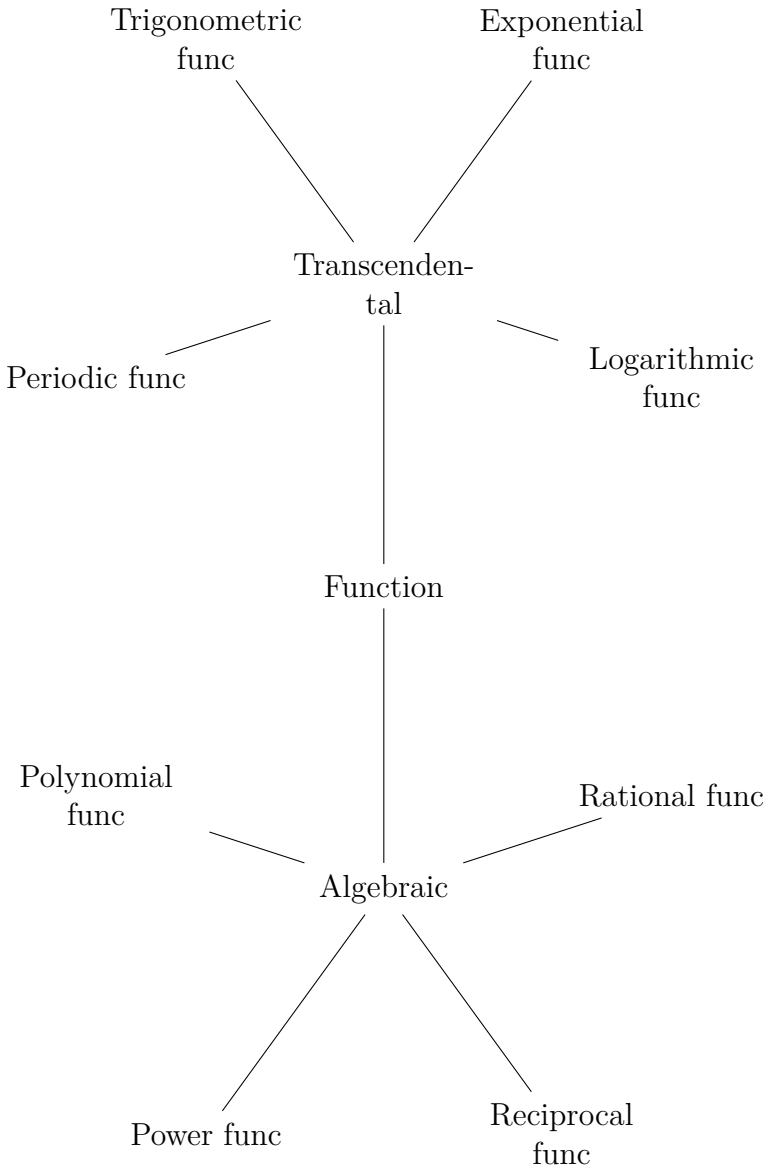
In practice,
+ The right hand side (RHS) is an independent variable or inputs.
+ The left hand side (LHS) is an dependent variable or outputs.

III Representations

1/ Verbally (description)
2/ Numerically (table)
3/ Visually (graph)
4/ Algebraically (formula)

IV Categories

There are several rudimentary functions we should bear in mind their properties.



1 Algebraic functions

Polynomial

A function f(x) is called a polynomial if

$$f(x) = a_n x^n + a_{n-1} x^{n-1} + a_{n-2} x^{n-2} + ... + a_1 x^1 + a_0 x^0$$

where:
+ n (>= 0): polynomial degree.

+ a_0, a_1, a_2, ...: coefficients.
+ a_n: leading coefficient.

Examples:
a/ $P(x) = 2^6 - x^4 + \frac{2}{5}x^3 + \sqrt{2}$
b/ $H(x) = x^4 + x^2 + 10$

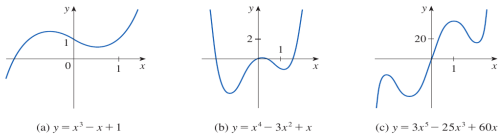


Figure 4: Some graphs of polynomial function]

In polynomial function at a fundamental level, we pay more attention to **quadractic function** and **cubic function**.

a/ Quadratic function
General form:

$$f(x) = ax^2 + bx + c$$

Solution formula:

$$x = \frac{-b \pm \sqrt{\Delta}}{2a}, \text{ where } \Delta = b^2 - 4ac = 0$$

Δ: Discriminant (định thức)
Discriminant properties:
+ $\Delta < 0$: Function has no solution (vô nghiệm)
+ $\Delta = 0$: Function has double root (nghiệm kép)
+ $\Delta > 0$: Function has two distinct solutions (hai nghiệm p/b)

Viète's formula for quadratic function
Sum of roots: $\Sigma = S = r_1 + r_2 = -\frac{b}{a}$

Product of roots: $\Pi = P = r_1 r_2 = \frac{c}{a}$

From Σ and Π we can establish the function via the Viète's formula:

$$f(x) = x^2 - Sx + P \Rightarrow x = \frac{S \pm \sqrt{S^2 - 4P}}{2}$$

b/ Cubic function
General form:

$$f(x) = ax^3 + bx^2 + cx + d = 0$$

Solution formula: See at Wiki How

Viète's formula for cubic function
Sum of roots: $\Sigma = S = r_1 + r_2 + r_3 = -\frac{b}{a}$

Product of roots: $\Pi = P = r_1 r_2 r_3 = -\frac{d}{a}$

Pairwise sum of product:$r_1 r_2 + r_1 r_3 + r_2 r_3 = \frac{c}{a}$

Then applying substitution for find solutions.

Power function

General form:

$$f(x) = x^a$$

where:
+ a = n or $\frac{1}{n}$ with n > 0
+ a = n or $\frac{1}{n}$ with n < 0

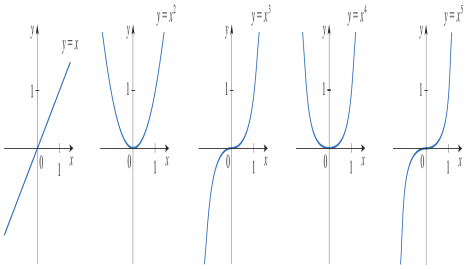
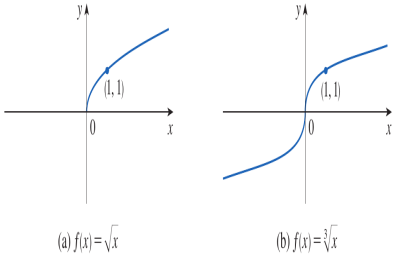


Figure 5: a = n with n > 0



(a) $f(x) = \sqrt{x}$ (b) $f(x) = \sqrt[3]{x}$

Figure 6: a = $\frac{1}{n}$ with n > 0

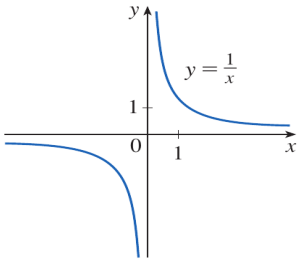


Figure 7: a = -1

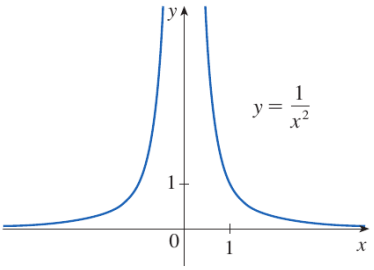


Figure 8: a = -2

It is not finished yet.

2 Transcendental functions

Trigonometric Functions

In calculus, **radian** will be used in place of **degree**. Plus, trigonometric is one of the typical **periodic functions** that will be discussed in section.

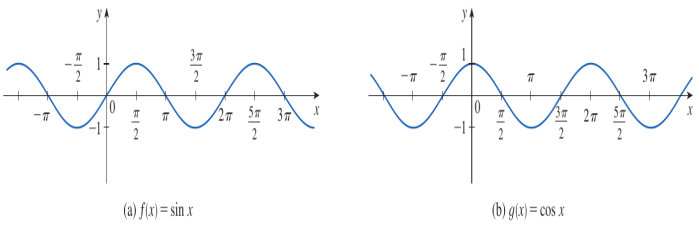


Figure 9: Graph of sin(x) and cos(x)

For more info, watch video from 79 → 108 in Precalculus course of Professor Leonard

Periodic Functions

Function that repeats itself in **regular intervals** or **periods**. Conversely, it is an aperiodic function.

$$f(x) = f(x + T), \text{ T is known beforehand}$$

Terminology:
Consider in the form that: $f(x) = Acos(\omega t + \varphi)$
1/ Period - T: The length between two successive peaks
SI unit: s

2/ Frequency - f: How many peaks in one unit of time
SI unit: Hz

3/ Angular velocity - ω: Time rate at which an object rotates, or revolves
SI unit: rad/s^2
Conversion formula: $\omega = 2\pi f = \frac{2\pi}{T}$

4/ Initial phase - φ: The initial function position on Ox
SI unit: radian

5/ Amplitude - A: The high of peak from the coordinate origin

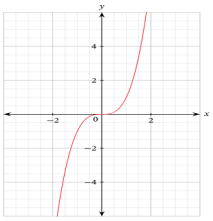
Periodic functions have an intimate relationship with Fourier serier, Fourier transform, or Harmonic oscillation.

V Evenness vs Oddity

Odd function

A function such that $f(-x) = -f(x)$ over the domain of f.
Geometrically, $f(-x)$ is symmetry with respect to the origin (O)
Ex: x, x^3, and $sin(x)$

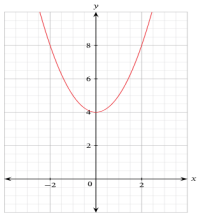
Figure 10: Graph of odd function



Even function

A function such that $f(-x) = f(x)$ over the domain of f.
Geometrically, $f(-x)$ is symmetry with respect to y axis
Ex: $\frac{1}{x^2}$, $|x|$, x^3, and $cos(x)$

Figure 11: Graph of even function



Properties

1/ $f(x) = 0$ satisfies both odd and even condition.

Additivity
2/ Even + Even = Even
3/ Odd + Odd = Odd

Multiplicity
4/ Even * Even = Even
5/ Odd * Odd = Even
6/ Odd * Even = Odd

Divisibility
7/ $\frac{Even}{Even} = Even$
8/ $\frac{Odd}{Odd} = Even$
9/ $\frac{Odd}{Even} = Odd\ g(x) = x^2)$

Derivative
10/ $\frac{\mathrm{d}}{\mathrm{d}x}$(Even) = Odd
11/ $\frac{\mathrm{d}}{\mathrm{d}x}$(Odd) = Even

Composition
12/ $(Odd \circ Even)(x)$ = Even

VI Transformation

VII Composition vs Combination

VIII Inversion

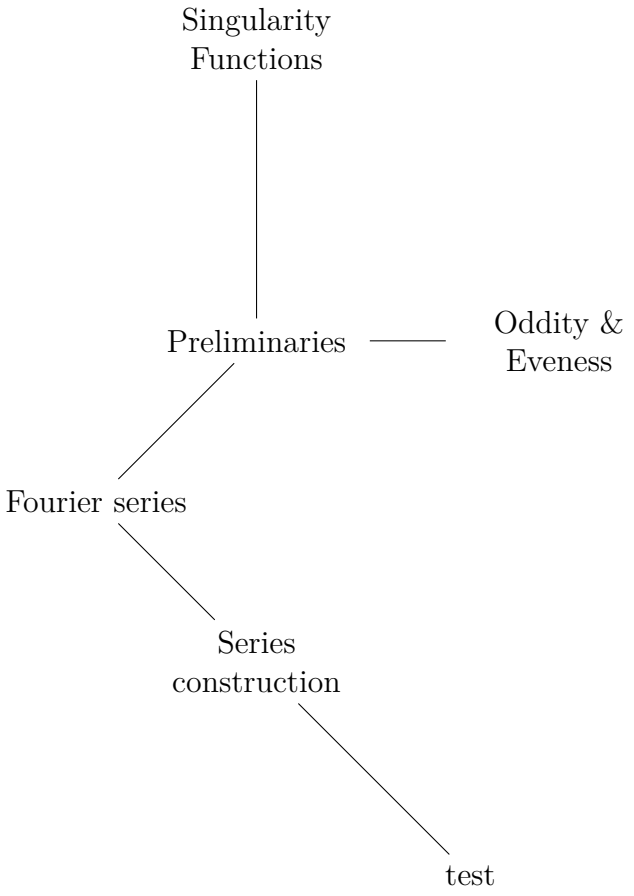
IX Limit

X Derivative

XI Monotonicity

B Fourier series

I Overview



II Preliminaries

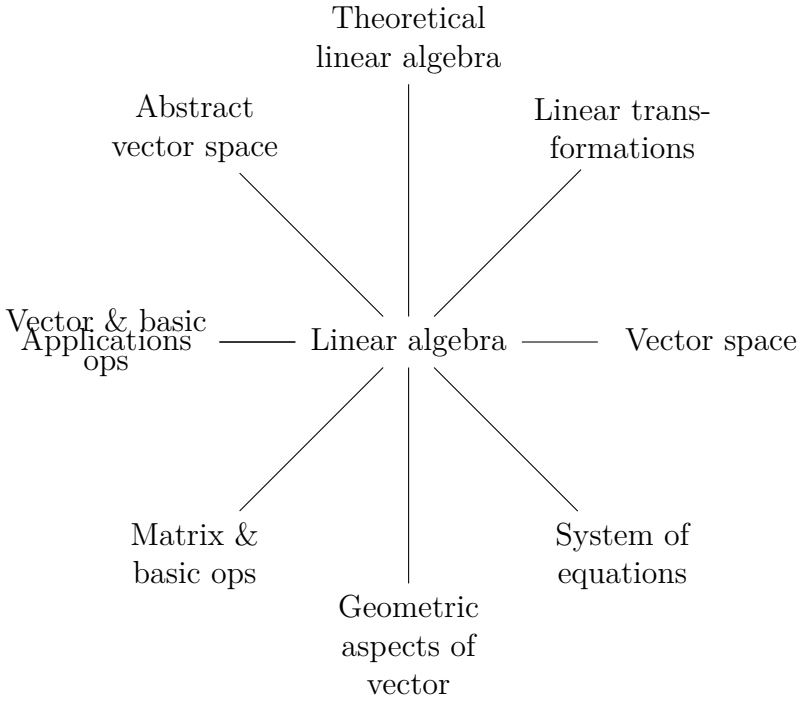
1 Function Oddity & Eveness

See at Evenness vs Oddity

C Linear Algebra

I Overview

Linear algebra is the way to organize and deal with a large set of number - matrix. Linear algebra is widely applied in almost AI fields from rudimentary to advanced topics. In addition, The combination between Linear algebra and Calculus leads to matrix calculus which will be discussed later.



II Vector & basic operations

Definition: A vector $\vec{v} \in \mathbb{R}^n$ is an n-tuple of real numbers.

Notation
a/ $\begin{bmatrix} x_1 & x_2 & \dots & x_n \end{bmatrix}$ for row vector

b/ $\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$ for col vector

$x_1, x_2, x_3, ...$ are called components or coordinates of vector.
To derive a uniform for further formula, col vector is chosen as a representation for vector.

Example
a/ $\vec{v} = \begin{bmatrix} 1 \\ 2 \\ \vdots \\ n \end{bmatrix}$ is a column vector in $\mathbb{R}^n$

Basic vector operations

+ Addition
+ Scaling
+ Subtraction (represented via addition and scaling)

Example

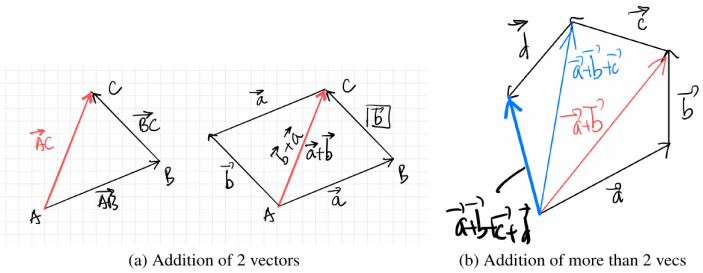
a/ **Addition**

Algeraic aspect

$$\vec{v_1} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}, \vec{v_2} = \begin{bmatrix} 4 \\ 5 \\ 6 \end{bmatrix}$$

$$\vec{v_3} = \vec{v_1} + \vec{v_2} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} + \begin{bmatrix} 4 \\ 5 \\ 6 \end{bmatrix} = \begin{bmatrix} 5 \\ 7 \\ 9 \end{bmatrix}$$

Geometric aspect



(a) Addition of 2 vectors (b) Addition of more than 2 vecs

Figure 12: Addition of two vectors: the parallelogram rule

b/ **Subtraction**

Algeraic aspect

$$\vec{v_1} = \begin{bmatrix} 4 \\ 5 \\ 6 \end{bmatrix}, \vec{v_2} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$$

$$\vec{v_3} = \vec{v_1} - \vec{v_2} = \begin{bmatrix} 4 \\ 5 \\ 6 \end{bmatrix} - \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} = \begin{bmatrix} 3 \\ 3 \\ 3 \end{bmatrix}$$

Geometric aspect

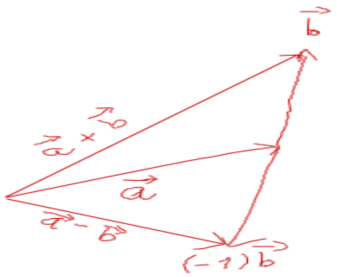


Figure 13: Subtraction of two vectors

c/ **Scailing**

Algeraic aspect

$$\vec{v_1} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$$

$$\vec{v_3} = 3\vec{v_1} = 3\begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} = \begin{bmatrix} 3 \\ 6 \\ 9 \end{bmatrix}$$

III Matrix & basic ops

Matrix can be interpreted in two different ways.

1^{st} way: A set of column vectors is put in adjacent order.

2^{nd} way: A way to represent coefficients of system of equations.

First interpretation

Assume that we have a contrived dataset as the following:

Price	Type	Area	Location
1B$	Semi-detached	400	10km to centre
1.5B$	Single family	700	20km to centre
2B$	Multi-family	1000	50km to centre

Table 1: Contrived dataset for housing price

from this dataset, the matrix will be

$$\begin{bmatrix} 10^9 & 0 & 400 & 10 \\ 1.5 * 10^9 & 1 & 700 & 20 \\ 2 * 10^9 & 2 & 1000 & 50 \end{bmatrix}$$

Second intepretation

Assume that we have a system of equations as the following:

$$\begin{cases} x + 2y + 3z = 5 \\ 2x + 9y + 10z = 50 \\ 3x + 12y + 15z = 90 \end{cases}$$

then the matrix will be

$$\left[\begin{array}{ccc|c} 1 & 2 & 3 & 5 \\ 2 & 9 & 10 & 50 \\ 3 & 12 & 15 & 90 \end{array}\right]$$

Note:

+ Uppercase for matrix name
+ Lowercase for matrix entry/ component

Basic matrix operations

+ Addition: $\mathbb{R}^{m\times n} + \mathbb{R}^{m\times n} = \mathbb{R}^{m\times n}$
+ Subtraction: $\mathbb{R}^{m\times n} - \mathbb{R}^{m\times n} = \mathbb{R}^{m\times n}$
+ Multiplication: $\mathbb{R}^{m\times l} \times \mathbb{R}^{l\times n} = \mathbb{R}^{m\times n}$

Addition and subtraction are performed as follows:
$C = A \pm B \iff c_{ij} = a_{ij} \pm b_{ij}, \forall i \in [1, ..., m], \forall j \in [1, ..., n]$

Matrix multiplication are performed as follows:
$C = AB \iff c_{ij} = \sum_{k=1}^{l} a_{ik} b_{kj}, \forall i \in [1, ..., m], \forall j \in [1, ..., n]$

Example
Suppose we have

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} \; B = \begin{bmatrix} 2 & 3 & 4 \\ 5 & 6 & 7 \end{bmatrix} \; C = \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix}$$

a/Addition

$$D = A + B = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} + \begin{bmatrix} 2 & 3 & 4 \\ 5 & 6 & 7 \end{bmatrix} = \begin{bmatrix} 1+2 & 2+3 & 3+4 \\ 4+5 & 5+6 & 6+7 \end{bmatrix} = \begin{bmatrix} 3 & 5 & 7 \\ 9 & 11 & 13 \end{bmatrix}$$

b/Multiplication

$$E = AC = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix} = \begin{bmatrix} 1.1 + 2.3 + 3.5 & 1.2 + 2.4 + 3.6 \\ 4.1 + 5.3 + 6.5 & 4.2 + 5.4 + 6.6 \end{bmatrix} = \begin{bmatrix} 25 & 28 \\ 49 & 64 \end{bmatrix}$$

Matrix taxonomy
1/ Square matrix
Matrix has size/ dimension n by n 2/ Retangular matrix
Matrix has size/ dimension n by n 3/ Unit/ Identity matrix
All entries are 1

4/ Zero/ Null matrix
All entries are 0

5/ Diagonal matrix
Off-diagonal entries equals 0 6/ Upper triangular matrix
7/ Lower triangular matrix

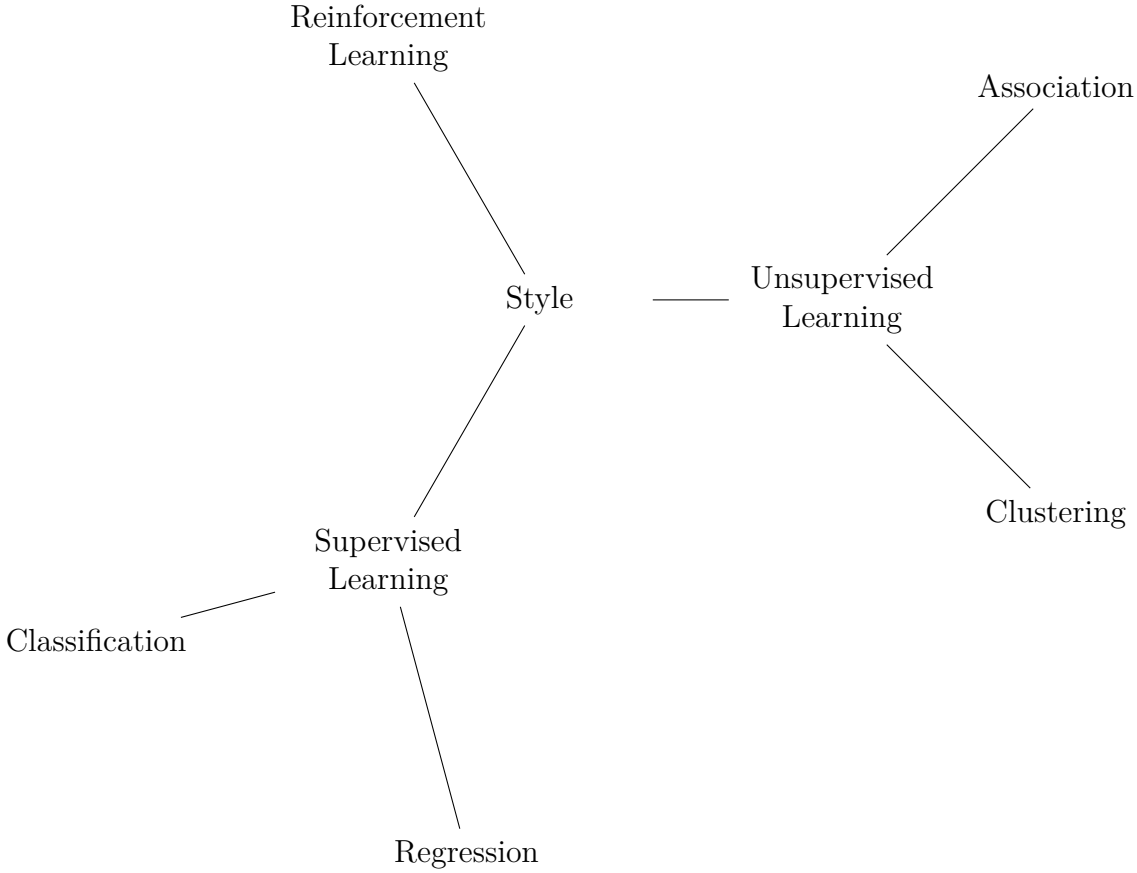
D Machine learning

I Overview

Machine learning is commonly split according to:

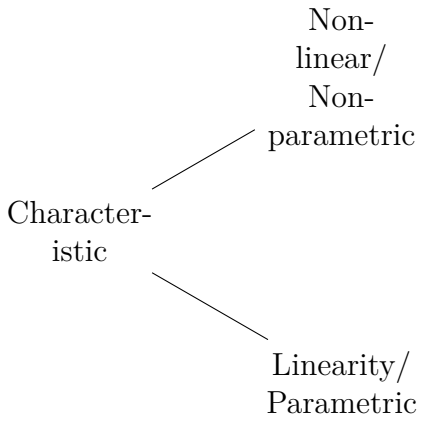
+ Learning style
+ Funtion-based algorithm
+ Charactered-based algorithm

According to the taxonomy in Artificial Intelligence A Modern Approach chapter 5, when mentioning about learning style, we have as following:



In function-based algorithm, we will have:

With the last one, we will have:
In function-based algorithm, we will have:



II Learning style

1 Supervised learning

Supervised learning (SVL) is applied when we want to make a prediction for a whole dataset and a set of desired labels/ patterns. We can think it as the following flow,
(Label, Data) — SVL model $\rightarrowtail$ Trained model — Predict with new data $\rightarrowtail$ What label of a new data is.

In mathematical lingo, we have:
Training dataset $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ and label dataset $\mathcal{Y} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$, where $\mathcal{X}\&\mathcal{Y}$ are two attributes of a collected dataset. Then we want to make a mapping such that:

$$\mathbf{y}_i \approx f(\mathbf{x}_i), \quad \forall i = 1, 2, \ldots, N$$

To take Iris Species (see at Iris Species) as an illustration, this dataset has 6 columns/ features including:
+ Id
+ SepalLengthCm
+ SepalWidthCm
+ PetalLengthCm
+PetalWidthCm
+Species

We can build a supervised learning model with classification algorithms with training and label dataset as the follows:
+ $\mathcal{X} = (SepalLengthCm)$
+ $\mathcal{Y} = (Species)$

This is the simplest model with one feature as an variable. If we want to build a more complex model, we can take into account all of features (SepalLengthCm, SepalWidthCm, PetalLengthCm, PetalWidthCm). In that scenario, we will have:
+ $\mathcal{X} = (SepalLengthCm, SepalWidthCm, PetalLengthCm, PetalWidthCm)$
+ $\mathcal{Y} = (Species)$

Supervised learning applications
+ Predictive analytics (Regression)
+ Image/ Object recognition (Classification)
+ Customer segmentation (Classification)
+ Spam detection (Classification)

2 Unsupervised Learning

Unsupervised learning (USVL) is applied when we want to do a job irrelevant to pattern in dataset such as clustering or dimensionality reduction.

In mathematical lingo, we have:
Training dataset $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ and no label dataset $\mathcal{Y} = \emptyset$

3 Reinforcement Learning

III Function-based Algorithm

1 Regression

a Linear Regression

Other names: Ordinary Least square (OLS) method.
Before diving into Linear Regress we need have a quick review some statiscal formulae including mean, standard deviations, correlations, covariance, and coeffecient of determination.

Note: All following formula is applied in population (stats aspect) or sample (machine learning aspect).

Mean - Expected Value
Def: Display the central tendency of data.
Notation: E(X), $\bar{x}$, μ
Discrete var

$$E(X) = \mu = \frac{1}{N}\sum_{i=1}^{N} x_i = \sum_{i=1}^{N} x_i p(x_i)$$

Continuous var
write later

Variance
Def: Mean square of deviation of data from mean..
Notation: V(X), σ^2
Derivative: Standard deviation
Notation: std, σ
Discrete variable

$$V(X) = E(X-\mu)^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)(x_i-\mu)^T \tag{1}$$

$$V(X) = E(X^2) - E^2(X) = \left(\frac{1}{N}\sum_{i=1}^{N} x_i^2\right) - \mu^2 = \sum_{i=1}^{N} x_i^2 p(x_i) - \mu^2 \tag{2}$$

(1) formula for computer
(2) formula for us

Continuous variable
write later

Mathematical operations:
+ Var(X+a) = Var(X)
+ Var(aX) = a^2Var(X)
+ Var(aX ± bY) = a^2Var(X) ± Cov(X,Y) + b^2Var(Y) (similar to $(a+b)^2$)

Meaning:
+ The smaller Variance is, the better data we have.

Covariance

Def: The measure of joint variability of two random variables
Notation: Cov
Discrete variable

$$Cov(X,Y) = E\left[(X-\mu_X)(Y-\mu_Y)\right] = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu_X)(y_i-\mu_Y) \tag{3}$$

$$\begin{aligned} Cov(X,Y) = E(XY) - E(X)E(Y) &= \left(\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}x_i y_j\right) - \mu_x\mu_y \\ &= \sum_{i=1}^{N}\sum_{j=1}^{N}x_i y_j p(x_i,y_j) - \mu_x\mu_y \end{aligned} \tag{4}$$

Continuous variable
write later

Mathematical operations:
+ Cov(X,X) = μ_X
+ Cov(X,Y) = Cov(Y,X)
+ Cov(aX,bY) = abCov(Y,X)
+ Cov(X+a,Y+b) = Cov(X,Y)
+ Cov(aX+bY, cW+dZ) = acCov(X,W) + adCov(X,Z) + bcCov(Y,W) + bdCov(Y,Z)

Meaning:
+ X, Y are ind rand vars → Cov(X,Y) = 0. But the reverse is not always true

Correlation Coefficient

Def: A normalized version of covariance
Synonym: Pearson's correlation coefficient
Notation: ρ, corr

$$corr(X,Y) = \rho_{X,Y} = \frac{Cov(X,Y)}{\sigma_x\sigma_y}$$

Meaning:
+ X, Y are ind rand vars → $\rho_{X,Y} = 0$ (X, Y are uncorrelated). But the reverse is not true.
+ -1 <= corr <= 1.
+ corr < 0, we say that X, Y are **negatively correlated**.
+ corr = 0, we say that X, Y are **uncorrelated**.
+ corr > 0, we say that X, Y are **positively correlated**.

Coefficient of Determination

Def: A measure of the goodness of fit of a regression model. Notation: R^2 or r^2

$$R^2 = 1 - \frac{SS_{reg}}{SS_{total}} = 1 - \frac{Explained\ variation}{Total\ variation} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y}_i)^2}$$

Meaning
$R^2 = 0$: The model does not predict the outcome.
(0,1): Between 0 and 1 The model partially predicts the outcome.
1: The model perfectly predicts the outcome.

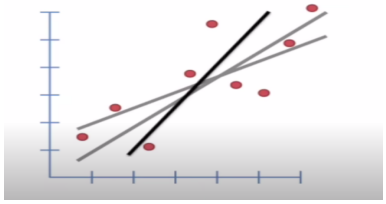
Let's dive into linear regression. Linear regression is a statistical model used for making a prediction between **dependent variable(s)** and **independent variable(s)**. For example, we want to predict house price (dependent variable) via other metrics (independent variable) such as area, number of rooms, downtown proximity, etc.

To start with 2D case, supposing that we have some contrived data as follows:

Table 2: Housing data

Price	Area
30	25
50	30
100	35
⋮	⋮

Our mission is to figure the line (black line in the picture below) such that the residuals reach optimum value. Residual is the difference between actual value versus predicted value



We want to model it as:
Price $\approx$ w_0 + w_1Area + ε

In 3D case, we would seek a plane having minimal residuals.
In 4D or higher dim, we would seek a hyperplane having minimal residuals.

In general, we have $Y = Xw + \varepsilon$

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1m} \\ 1 & x_{21} & x_{22} & \dots & x_{2m} \\ 1 & x_{31} & x_{32} & \dots & x_{3m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & w_{nm} \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix} + \begin{bmatrix} \varepsilon_0 \\ \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Note:
+ All of them are column vectors.
+ This form has m features and n observations.

Terminology for X, Y, w
+ Y: independent var, regressand, measured var, exogenous var
+ X: dependent var, regressor, explanatory var, endogenous var
+ w: slope, weight
+ ε: residual

As we know that residual is the distance between actual value to predicted value, we therefore minimize

$$R = \frac{1}{N}\sum_{i=1}^{N} |\varepsilon_i| = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}_i|$$

Mathematically this work is quite troublesome and daunting when taking derivative, because its derivative is not continuous at 0, so mathematicians took square of it. Ultimately, we need to minimize this guy,

$$R = \frac{1}{N}\sum_{i=1}^{N} |\varepsilon_i|^2 = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}_i|^2$$

The only method for finding is to take the derivative of it with respect to w. Mathematically, we have loss function $\mathcal{L}(\mathbf{w})$ as follow:

$$\begin{aligned}
\mathcal{L}(\mathbf{w}) &= \arg\min_{w} \left(Y - \hat{Y}\right)^2 \\
&= \arg\min_{w} \left(Y - Xw\right)^2 \\
&= \arg\min_{w} \left(Y - Xw\right)^T \left(Y - Xw\right) \\
&= \arg\min_{w} \left(Y^T - X^T w\right) \left(Y - w^T X\right) \\
&= \arg\min_{w} \left(Y^T Y - Y^T X w - w^T X^T Y + w^T X^T X w\right)
\end{aligned}$$

Now we take partial derivative w.r.t w

$$\frac{\partial \mathcal{L}(\mathbf{w})}{\partial \mathbf{w}} = -2X^T Y + 2w^T X^T X w$$

Setting the gradient to zero for finding min

$$\begin{aligned}
\frac{\partial \mathcal{L}(\mathbf{w})}{\partial \mathbf{w}} &= 0 \\
X^T X w &= X^T Y \\
w &= \left(X^T X\right)^{-1} X^T Y
\end{aligned}$$

Drawbacks of Linear Regression:
+ Very sensitive to noise, thereby requiring data preprocessing
+ Can not represent complex function (Function with a combination of polynomial and trigonometry)

E Bias & Variance

What does Bias mean ?

Bias is an error calculated by the mean of the difference b/w **predicted** ($\hat{y}$) and **observed/actual/ ground true** value (y)

What does Variance means

Variance is the difference in in fits b/t datasets (e.g. training vs testing sets)

The trade-off of Bias-Variance

Bias-variance tradeoff is one of the vital aspects in building a good model.

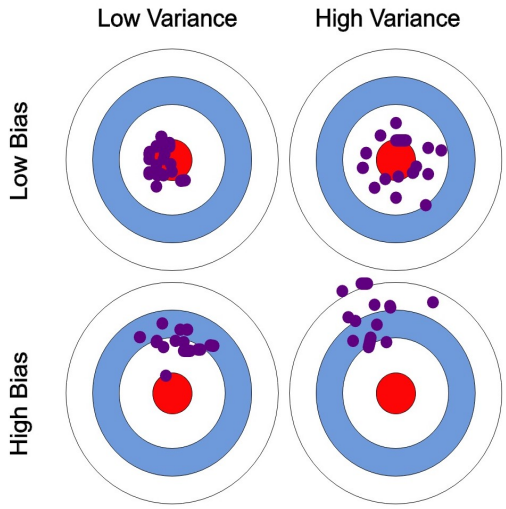


Figure 14: The correlation b/t Bias & Variance
Red circle for y. Purple circle for $\hat{y}$

Three leading phenomena: Overfitting, Well-fitted, Underfitting

What is overfitting

Overfitting occurs when model is well-performed on **training set/ seen data** but performs poorly on **testing set/ unseen data**. In technical jargon, we can say this model has low bias and high variance as well.

What is well-fitting

Well-fitting is when model well performance on both training & testing set. We can say that this model has a 'sweet' point for bias & variance

What is underfitting

Underfitting happens when model performs poorly on both traning and testing set. In technical jargon, we can say this model has low bias and low variance as well

What can we do to balance bias & variance in machine learning context ?

a/ Regularization
b/ Early stopping
c/ Data augmentation
d/ Pruning

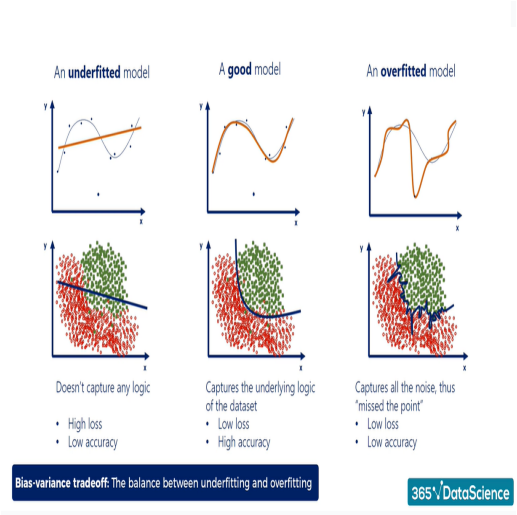


Figure 15: Example 1

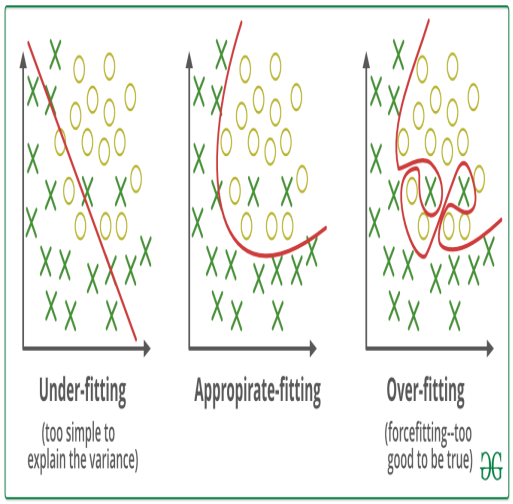


Figure 16: Example 2

e/ Ensembling

What can we do to balance bias & variance in deep learning context ?

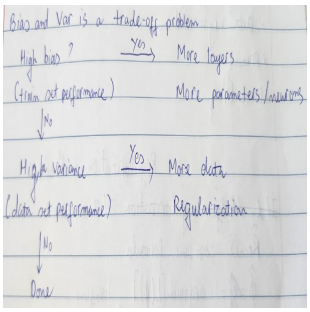


Figure 17: Solution for balancing bias-variance in deep learning context

F (Explicit) Regularization

Recall cost function formula:

$$Cost = \arg\min_{\phi}[Loss[\phi]] = \arg\min_{\phi}\left[\sum_{i=1}^{N} l_i[x_i, y_i]\right] \tag{5}$$

where,

a/ ϕ: arg set includes both weights and biases

b/ each $l_i[x_i, y_i]$: mismatch/ difference b/t the network predictions $y = f[x_i, \phi]$ and output targets y_i for each training pair

To bias this minimization, we include an additional term into the model.

$$Cost = \arg\min_{\phi}[Loss[\phi]] = \arg\min_{\phi}\left[\sum_{i=1}^{N} l_i[x_i, y_i] + \lambda.g[\phi]\right] \tag{6}$$

where,

a/λ: positive scalar ranging from 0 to 1 controls an impact of the regularization term

b/$g(\phi)$: regularization function (e.g L1, L2, L1+L2, etc.)

L1 regularization

Syn: LASSO regularization

$$Cost = \arg\min_{\phi}[Loss[\phi]] = \arg\min_{\phi}\left[\sum_{i=1}^{N} l_i[x_i, y_i] + \lambda \sum_{j} |\phi_j|\right]$$

L2 Regularization

Syn: Ridge regularization, Tikhonov regularization

$$Cost = \arg\min_{\phi}[Loss[\phi]] = \arg\min_{\phi}\left[\sum_{i=1}^{N} l_i[x_i, y_i] + \lambda \sum_{j} \phi_j^2\right]$$

Elastic Net regularization

Syn: L1 + L2

$$Cost = \arg\min_{\phi}[Loss[\phi]] = \arg\min_{\phi}\left[\sum_{i=1}^{N} l_i[x_i, y_i] + \lambda_1 \sum_{j} |\phi_j| + \lambda_2 \sum_{j} \phi_j^2\right]$$

*Math proof will be given later

Remark:

+ **LASSO Regression (L1)** can exclude useless vars from model, so it is a litter better than **Rigde Regression (L2)**

sdaasdsa