MATHEMATICAL FOUNDATIONS FOR DATA SCIENCE

FUNCTIONS

Set — unordered collection, unique items

$x \in S$, $x \notin S$, $\{x \in S : P(x)\}$, $[a,b]$, (a,b)

$P(x)$ → cond^n; $[a,b]$ → closed interval; (a,b) → open interval

$$\sum_{x=a} f(x) = f(a) + f(a+1) + \cdots f(b-1) + f(b)$$

$$\prod_{x=a}^{b} f(x) = f(a) \times f(a+1) \times \cdots - f(b-1) * f(b)$$

$f : X \to Y$

X → domain; Y → co-domain → all possible outputs

$\{f(x) : x \in X\}$ → image (set of outputs) → for given range of inputs

Zeroes of $f(x) = \{x \in X : f(x) = 0\}$ → roots

Inverse : f^{-1} $\lceil f : X \to Y$

$f^{-1}: Y \rightarrow X$

$f^{-1}(f(x)) = x$ $\qquad$ $f^{-1}(x) \neq \frac{1}{x}$

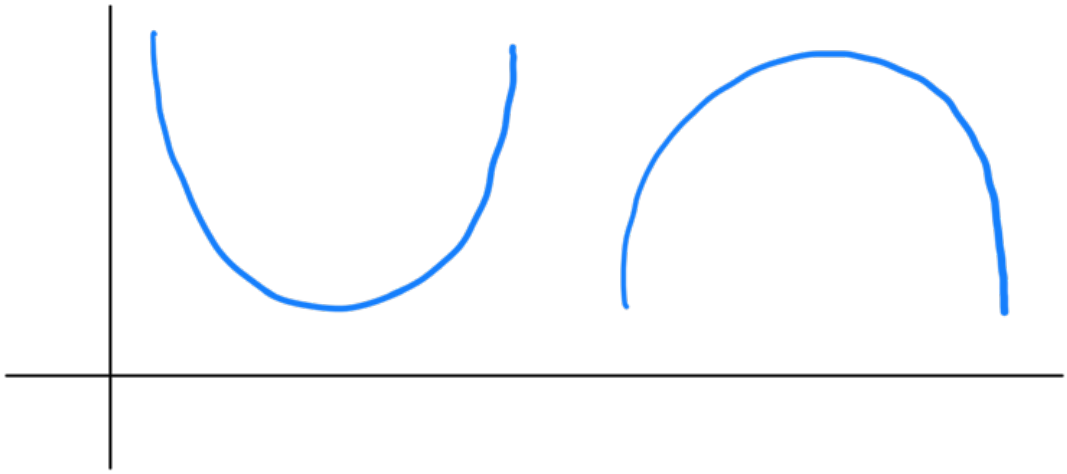
TYPES OF FUNCTION

(A) Injective – One to One (1)

Surjective – for every $f^{-1}(Y)$ there exists a $x \in X$ (2)

Bijective – BOTH (1) and (2)

(B) Convex $\qquad$ Concave



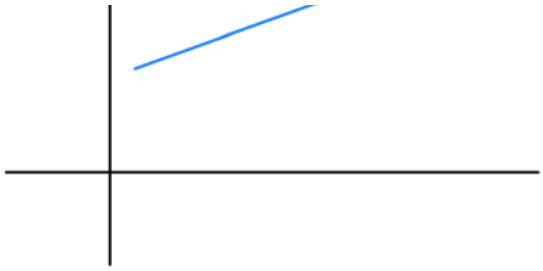
(C) Linear

$f(x) = mx + b$

Only 1 zero at $x = -b/m$

Bijective if $m \neq 0$

both convex and Concave



Polynomial

$$f(x) = \sum_{i=0}^{n} a_i x_i$$

$$= a_0 + a_1 x + a_2 x^2 + \ldots a_n x^n$$

$$\downarrow$$

$$x^0 = 1$$

Exponential

$$f(x) = b \cdot a^x$$

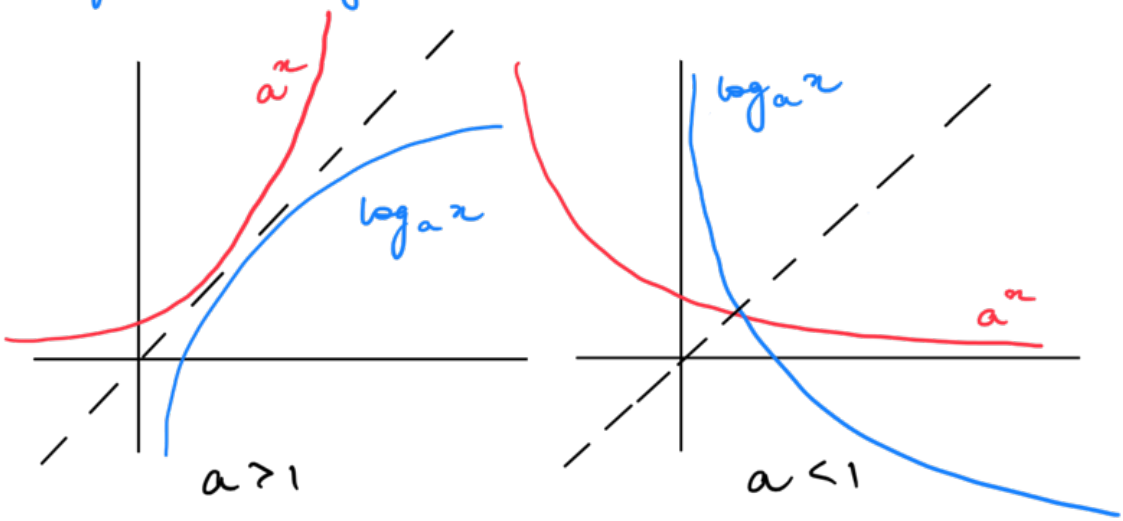
Logarithmic

$$f(x) = \log_a(x)$$

inverse of each other

if $f(x) = a^x$

$f^{-1}(x) = \log_a(x)$



$$\log_a(mn) = \log_a m + \log_a n$$

$$\log_a(m/n) = \log_a m - \log_a n$$

$$\log_a x^n = n \log_a x$$

$\log_a x$ $\qquad\qquad$ $\dfrac{\log_b x}{\log_b a}$

$\log_a x^m = m \log_a x \longrightarrow \log_a(a^x) = x$

$a^{\log_a x} = x$

Power law

$f(x) = b \cdot x^{-a}$

exp $f(x)$ decay faster than power laws

$\downarrow$

$b \cdot a^x$ $\qquad$ log $\longrightarrow$ power law $\longrightarrow$ st. line

DIFFERENTIATION

LINEAR TRANSFORMATION

* log $\qquad \log_a(x)$

* log - log $\qquad$ power

data $(x_1, y_1) - - - (x_n, y_n)$

plot $(\ln x_1, \ln y_1) - - - (\ln x_n, \ln y_n)$

$f(x) = b\, x^{-a} \qquad [f(x) = y]$

$\ln(y) = \ln b + (-a) \ln x$

$\tilde{y} = \ln b - a\, \tilde{x}$

$m = (-a)$ $\quad$ and $\quad b = \ln(b)$

* Semi - log

log lin lin log

expp log

log lin

$(x_1, \ln(y_1) \text{ ---- } x_n, \ln(y_n))$

$f(x) = b \cdot a^x$

$\ln(y) = \ln b + x \ln a$

$\hat{y} = (\ln a)\, x + \ln b$

$m = \ln a$ and $b = \ln(b)$

lin log

$(\ln(x_1), y_1 \text{ ---- } \ln(x_n), y_n)$

$f(x) = b \cdot \log_a x$

$= b \cdot \frac{\ln x}{\ln a}$

$y = \frac{b}{\ln a} (\ln x)$

$\tilde{y} = \frac{b}{\ln a} \cdot \tilde{x}$ $m = b/\ln(a)$

DERIVATIVE

$f(x) \longrightarrow f'(x)$: slope of tangent

c	0
x^n	$n x^{n-1}$
a^x	$\ln a \cdot a^x$
$\log_a x$	$\frac{1}{\ln a \cdot x}$

$e^x \to e^x$

$\ln x \to \frac{1}{x}$

$f(x) + g(x)$ $f'(x) + g'(x)$

$f(x) - g(x)$	$f'(x) - g'(x)$
$f(x)\,g(x)$ product rule	$f'(x)\cdot g(x) + f(x)\cdot g'(x)$
$f(g(x))$ chain rule	$g'(x)\cdot f'(g(x))$

$|x|$ is not differentiable

$f(x) = x^3 \quad f'(x) = 3x^2 \quad f''(x) = 6x \quad f'''(x) = 6$

OPTIMISING FUNCTION

INCREASING FUNCTION

$x \uparrow$ and $f(x) \uparrow$

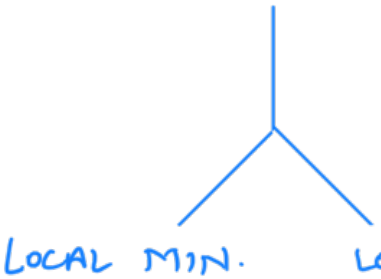
$f'(x) > 0$

DECREASING FUNCTION

$x \downarrow$ and $f(x) \downarrow$

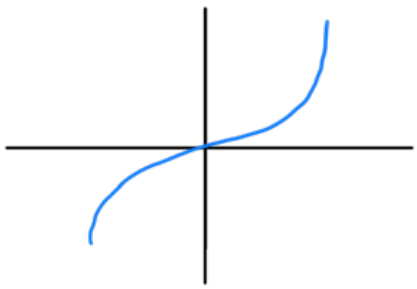
$f'(x) < 0$

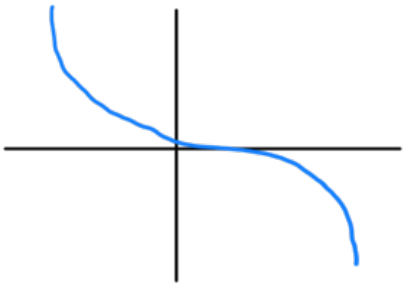
x where $f'(x) = 0$: STATIONARY POINT



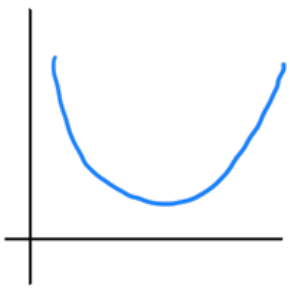
$f'(x)$ changes from -ve to +ve / +ve to -ve

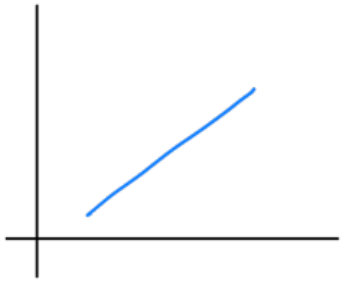
Inflection point – st. pt. but neither local max nor local min

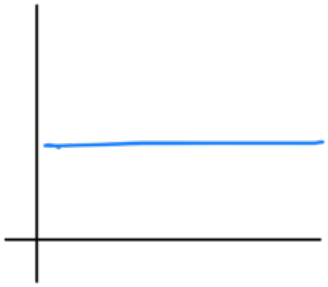




ex–







$f(x)$ $f'(x)$ $f''(x)$

GLOBAL EXTREMUM

$f(a) \leq f(x)$ $f(a) \geq f(x)$ $\forall x \in X$

for a $f(x)$ with $x \in [a,b]$, we may get global extrema at :

end pt. 1. a or b

st. pt. 2. c where $f'(c) = 0$

critical pt. 3. c where $f'(c)$ does not exist

i.e. $f''(x) > 0$ Concavity

$f'(x) < 0$ concavity

∵ if a $f(x)$ is completely concave or completely convex, then its local extrema is also the global extrema.

↳ if NOT, break into sub domains

RESIDUAL SUM OF SQUARES

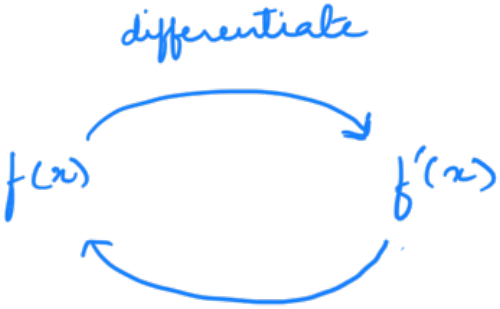
for data (x_1, y_1) ---- (x_n, y_n)

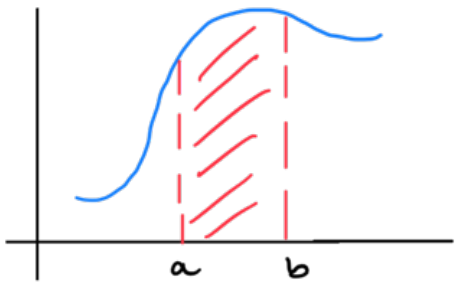
$$RSS = \sum_{i=1}^{n} (y_i - f(x_i))^2$$

Squaring to measure deviations ?

∵ it allows finding derivatives and penalise large deviations heavily

INTEGRATION





- area under the curve

↓

area below x-axis is -ve

anti-derivate

$$\int_a^b f(x) \cdot dx = F(b) - F(a)$$

x^a $\quad\quad$ $\frac{1}{a+1} x^{a+1}$

$\frac{1}{x}$ $\qquad$ $\ln(x)$

e^{ax} $\qquad$ $\frac{1}{a} e^{ax}$

LINEAR ALGEBRA

VECTORS

$$R^d = \left\{ \begin{matrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{matrix} : x_1, x_2 \dots x_n \in R^d \right.$$

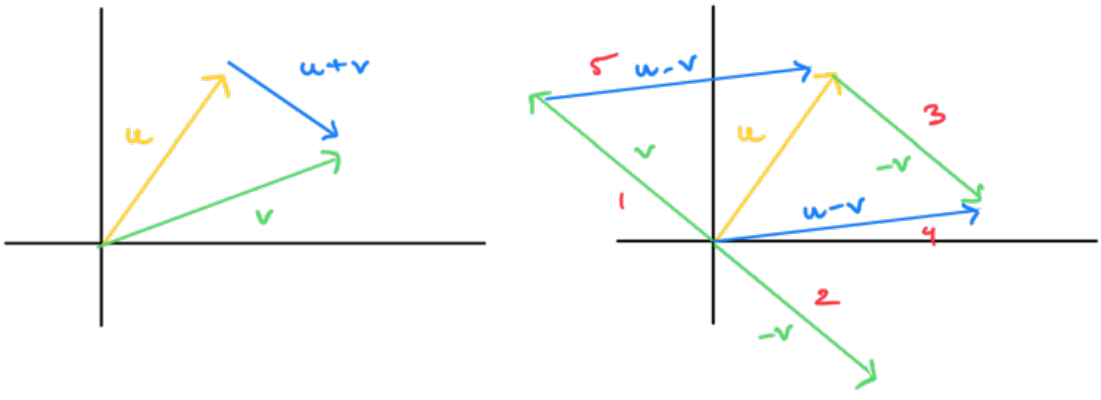
R^d ↓ set of d tuples

ex- R^3 contains $\begin{pmatrix} 1 \\ 0 \end{pmatrix}$, $\begin{pmatrix} 2.5 \\ 4 \end{pmatrix}$ and $\begin{pmatrix} c \\ x \end{pmatrix}$

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} + \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_1+y_1 \\ x_2+y_2 \\ \vdots \\ x_n+y_n \end{pmatrix}$$

$$c \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} cx_1 \\ cx_2 \\ \vdots \\ cx_n \end{pmatrix}$$

a vector $(v, \vec{v}, \tilde{v})$ has length and dirⁿ but no posⁿ



A line joining points u and v

contains u contains the points corresponding to $\alpha u + (1-\alpha)v$ where $\alpha \in [0,1]$

$w = a_1 v_1 + a_2 v_2 \text{ ---- } + a_n v_n$

↓

linear comb^n of $v_1 \text{ ---} v_n$ ↴

→ Linear dependence leads to redundancy

If not, then linearly Independent

ex- $\begin{pmatrix}1\\0\end{pmatrix}$ & $\begin{pmatrix}0\\1\end{pmatrix}$ are linearly independent

but $\begin{pmatrix}1\\0\\0\end{pmatrix}$, $\begin{pmatrix}0\\1\\0\end{pmatrix}$ and $\begin{pmatrix}1\\1\\0\end{pmatrix}$ are not.

as $\begin{pmatrix}1\\1\\0\end{pmatrix} = \begin{pmatrix}1\\0\\0\end{pmatrix} + 1\begin{pmatrix}0\\1\\0\end{pmatrix}$

if $v = \begin{pmatrix}v_1\\\vdots\\v_n\end{pmatrix}$ $w = \begin{pmatrix}w_1\\\vdots\\w_n\end{pmatrix}$

$v \cdot w = (v, w)$

$= v_1 w_1 + v_2 w_2 \text{ ---- } v_n w_n$

if $v \cdot w = 0$ then v and w are orthogonal
i.e. perpendicular

EUCLIDEAN NORM

$\|v\| = \sqrt{v_1^2 + v_2^2 + v_3^2 + \text{---} v_n^2}$

$= \sqrt{v \cdot v}$

we can't use the term length in case of multiple dimensions therefore it is referred to as NORM

MATRICES

$(m \times n)$ matrix where m = row, n = column.

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

$$c \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} = \begin{pmatrix} ca_{11} & ca_{12} \\ ca_{21} & ca_{22} \end{pmatrix}$$

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \qquad B = \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix}$$

$$A \cdot B = \begin{pmatrix} a_{11} \cdot b_{11} + a_{12} \cdot b_{21} & a_{11} b_{12} + a_{11} \cdot b_{22} \\ a_{21} \cdot b_{11} + a_{22} \cdot b_{21} & a_{21} \cdot b_{12} + a_{22} \cdot b_{22} \end{pmatrix}$$

$$M_1 = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \qquad M_2 = \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix}$$

$\det M_1 = ad - bc$

$\det M_2 = a(ei - fh) - b(di - gf) + c(dh - ge)$

Identity Matrix $(I) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ or $\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$

Zero Matrix $= \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$

if $B \cdot A = I$ then $B = A^{-1}$ and $A = B^{-1} \rightarrow AB = I$

$A \cdot B \neq B \cdot A$

$\det(AB) = \det(A) \cdot \det(B)$

$\det(I) = 1$

$\det(A) \neq 0$ if A^{-1} exists

→ let $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$

then $A^{-1} = \frac{1}{\det(A)} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$

$Ax = b \implies x = A^{-1}b.$

GAUSSIAN ELIMINATION

Steps –

1. Swap 2 rows
2. Mul a row by non zero num.
3. Add a multiple of one row to another.

to achieve $\begin{pmatrix} * & * & * \\ 0 & * & * \\ 0 & 0 & * \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} * \\ * \\ * \end{pmatrix}$

for eg- $R_1 \longleftrightarrow R_2$

$R_2 \longrightarrow 5 * R_2$

$R_3 \longrightarrow 2R_3 - R_2$

Types of solution :

1. Exactly ONE solution
2. NO solution

$$\begin{pmatrix} * & * \\ 0 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} * \\ * \end{pmatrix}$$

$* \equiv$ non zero num. $\quad 0x + 0y = *$

$0 \neq *$

3. MULTIPLE Solutions

ex - $\begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & -2 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$

$x + z = 1$

$y - 2z = 0$

$\Rightarrow \quad x = 1 - z \quad$ and $\quad y = 2z$

put $z = t$

Sol^n $= (1-t, 2t, t)$

$\rightarrow$ free variable : INFINITE sol^n(s)

EIGEN VALUES, VECTOR

$Ax = \lambda x \rightarrow$ eigen vector (can't be zero)

$\lambda \rightarrow$ eigen value

$\Rightarrow (A - \lambda I)x = 0$ (1)

$\det(A - \lambda I) = 0$ Characteristic eq^n (2)

$n \times n$ matrix $\rightarrow n$ sol^n(s) $\rightarrow n$ eigen values

Eigen values aren't unique (multiples of 0)

Steps -

1. find λ from $\det(A - \lambda I) = 0$
2. use λ to find x from $(A - \lambda I)x = 0$

A is diagonalisable if $A = PDP^{-1}$ and

using $\begin{pmatrix} x & 0 & 0 \\ 0 & y & 0 \end{pmatrix}^n = \begin{pmatrix} x^n & 0 & 0 \\ 0 & y^n & 0 \end{pmatrix}$ we get.

$$A^n = P D^n P^{-1}$$

$n \times n$ matrix

$$\begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}$$ upto λ_n

$(v_1, v_2, \ldots v_n)$

BINARY RELATIONS

a binary rel^n relates 2 parameters

eg- points in xy plane $(1,2)$

every $f(x) \xrightarrow{\text{GIVES}}$ rel^n but NOT conversely

$\times$

eq^n of circle : $(x-u)^2 + (y-v)^2 = r^2$

ellipse : $\frac{(x-u)^2}{a^2} + \frac{(y-v)^2}{b^2} = r^2$

centre at (u, v)

MULTIVARIATE FUNCTION

$z = f(x, y)$

$f_x = \frac{\partial}{\partial x}(f)$ $\qquad$ $f_y = \frac{\partial}{\partial y}(f)$

$\Rightarrow$ PARTIAL DERIVATIVE

$\nabla f = \begin{bmatrix} f_x \\ f_y \end{bmatrix}$ $\rightarrow$ GRADIENT VECTOR

OPTIMISING MULT. FUNCTIONS

$$f(x+\Delta x) = f(x) + f'(x)\Delta x$$

$$f(x+\Delta x, y+\Delta y) \approx f(x,y) + f_x(x,y)\Delta x + f_y(x,y)\Delta y$$

STATIONARY POINT

$f(x,y)$ where $f_x = 0$ and $f_y = 0$

implying $\nabla f(x,y) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ and is either local extrema or saddle pt.

HESSIAN MATRIX

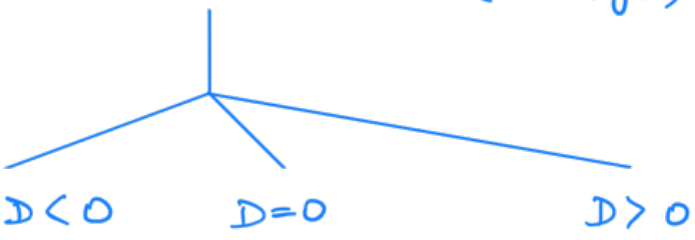
$$H(x,y) = \begin{bmatrix} f_{xx}(x,y) & f_{xy}(x,y) \\ f_{yx}(x,y) & f_{yy}(x,y) \end{bmatrix}$$

↓

Second partial derivative

if $f_{xy} = f_{yx}$ the nice funct^n

$$D = \det(H(x,y))$$



saddle pt. inconclusive local extremum

$f_{xx} > 0$ $f_{xx} < 0$

local min local max.

GLOBAL EXTREMA

possible candidates – 1. stationary pt.s

f_x & f_y are undefined ← 2. singular pt.s

3. boundary

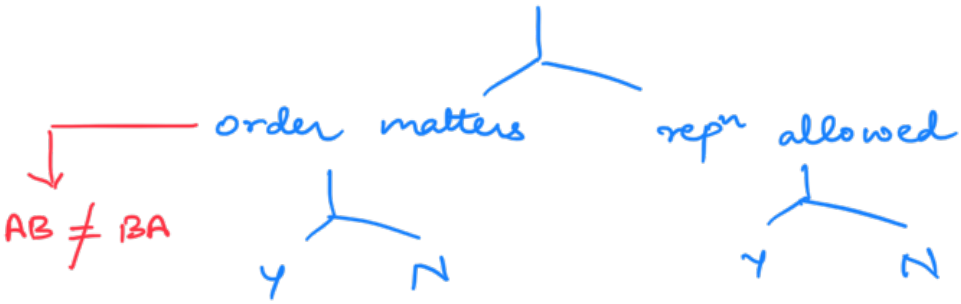
COMBINATORICS

Multiplication Rule $|S| = \prod_{j=1}^{n} k_j$ $k \times l$

Addition Rule $|S| = \sum_{j=1}^{n} s_j$ $k + l$

Complement Rule $S = S_g \cup S_b$ $S_g \cap S_b = 0$ } $1 - k$, $S_g = S - S_b$

TYPES OF SELECTION (4)



ordered sel^n w/o rep^n $\frac{n!}{(n-r)!}$

unordered sel^n w/o rep^n $\frac{n!}{r!(n-r)!}$

ordered sel^n with rep^n n^r

unordered set^n with rep^n $\frac{(n+r-1)!}{r!\,(n-1)!}$

PASCAL'S TRIANGLE

$$\binom{n}{r} = \binom{n}{n-r}$$

PIGEON HOLE PRINCIPAL

1. n items placed in m containers
2. $n > m$
3. at least one container has $\lceil \frac{n}{m} \rceil$ items

↓ rounding up.

PROBABILITY

Sample space : set of possible outcomes

$Pr: S \longrightarrow [0,1]$ prob. functn

Sum of prob.(s) of outcomes $= 1$

uniform prob. space : each event has equal prob.

Event - subset of Sample space.

ex - $Pr(A) = \frac{\text{no. of outcomes in } A}{\text{total no. of outcomes}}$

$= \frac{|A|}{|S|}$

$Pr(\bar{A}) = 1 - Pr(A)$

$Pr(A \cup B) = Pr(A) + Pr(B) - Pr(A \cap B)$

if $Pr(A \cap B) = 0 \longrightarrow$ mutually exclusive

i.e. A and B cannot occur together

Independent Events : $Pr(A \cap B) = Pr(A) Pr(B)$

if 1 true then all true:

1. A and B are independent
2. $\overline{A}$ and $\overline{B}$ are independent
3. A and $\overline{B}$ are independent
4. $\overline{A}$ and B are independent

INDEPENDENT REPEATED TRIALS

Sample space $= S_1 \times S_2$

$$Pr(A|B) = \frac{P(A \cap B)}{P(B)}$$

$$= \frac{P(B|A) \cdot P(A)}{P(B)}$$

if independent, $P(A|B) = P(A)$

BAYES THEOREM

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B|A) \cdot P(A) + P(B|\overline{A}) \cdot P(\overline{A})}$$

$$= \frac{P(B|A) \cdot P(A)}{P(B)}$$

ADVANCED PROBABILITY

RANDOM VARIABLE

$P(X = x)$ a funct^n from sample space to R

Random var·s are independent if $\forall x \& y$:

$P(X = x \text{ and } Y = y) = P(X = x) \cdot P(Y = y)$

→ Expected Value

$E[x] = p_1 x_1 + p_2 x_2 + \dots + p_n x_n$

↓ weights / probabilities

Laws of large no.(s) –

$\lim_{n \to \infty} \frac{1}{n} (x_1 + x_2 - \dots - x_n) = \mu$

if $E[x] = \mu$

$\text{Var}[x] = E[(x-\mu)^2] = E[x^2] - \mu^2$

↓ $\sigma^2 = \text{Var}[x]$ → std. dev^n

variance

If x and y are independent :

$E[x+y] = E[x] + E[y]$ → Linearity of Expect^n.

$E[x \cdot y] = E[x] \cdot E[y]$

$E[kx] = k E[x]$

$\text{Var}[kx] = k^2 \cdot \text{var}[x]$

$\text{Var}[x+y] = \text{Var}[x] + \text{var}[y]$

DISCRETE UNIFORM DISTRIBUTION

Set of cons. integers (equally likely)

$P(x=k) = \frac{1}{b-a+1}$ for $k \in \{a, a+1 \dots, b\}$

$E[x] = \frac{a+b}{2}$ $\quad \text{Var}[x] = \frac{(b-a+1)^2 - 1}{12}$

BERNOULLI DISTRIBUTION

prob. that a process succeeds or fails

$$P(X=k) = \begin{cases} p & k=1 \quad \text{success} \\ 1-p & k=0 \quad \text{fail} \end{cases} \qquad p \in [0,1]$$

$$E[X] = p \qquad \text{Var}[X] = p(1-p)$$

GEOMETRIC DISTRIBUTION

Seq of Bernolli trials — k failures before first success.

$$P(X=k) = p(1-p)^k \quad \text{for } k \in \mathbb{N}$$

$$E[X] = (1-p)/p \qquad \text{Var}[X] = (1-p)/p^2$$

BINOMIAL DISTRIBUTION

Seq of n independent Bernolli Trials with k sucess.

$n \in \mathbb{Z}^+$ and $p \in [0,1]$

$$P(X=k) = \binom{n}{k} p^k (1-p)^{n-k} \quad \text{for } k \in \{0, n\}$$

POISSON DISTRIBUTION

with an avg of λ events per unit time, it gives the prob. that k events occur in unit time.

$$P(X=k) = \frac{\lambda^k \cdot e^{-\lambda}}{k!} \quad \text{for } k \in \mathbb{N}$$

$$E[X] = \lambda \qquad \text{Var}[X] = \lambda$$

CONTINUOUS PROBABILITY

CONTINUOUS PROBABILITY

Probability Density function

→ PDF can not be negative

→ Integral of the PDF = 1 (over whole domain)

$$P(a \le X \le b) = \int_a^b f \cdot dx$$

$$E[x] = \int_{-\infty}^{\infty} x \cdot f \, dx$$ Var. stays same.

A continous uniform dist^n has following PDF

$$f(x) = \begin{cases} \frac{1}{b-a} & x \in [a,b] \\ 0 & \text{otherwise} \end{cases}$$

$$E[x] = (a+b)/2 \quad Var[x] = (b-a)^2/12$$

EXPONENTIAL DISTRIBUTION

Returns the time b/w events in a Poisson process.

→ randomly spread out with even density

PDF: $\lambda e^{-\lambda x}$ for $x \ge 0$

$$E[x] = 1/\lambda \quad Var[x] = 1/\lambda^2$$

NORMAL DISTRIBUTION

PDF: $$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

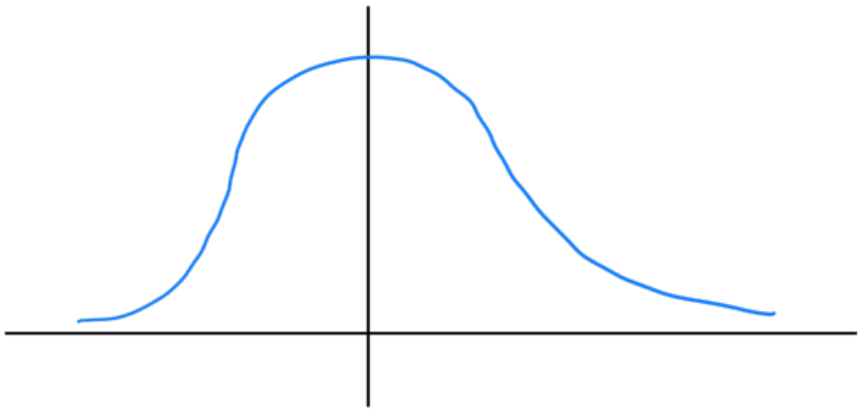
$$E[x] = \mu \quad Var = \sigma^2$$

many discrete dist^n(s) tend towards a normal dist^n

$z = (x - \mu)/\sigma$

how many std. deviations above or below the mean X is.

$$Pr(a < x < b) = Pr\left(\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}\right)$$



GRAPH THEORY

Graph – vertices + edges

Vertices – may or may not be labelled

vertices sharing an edge are labelled as ADJACENT

Edge – connect^n b/w 2 vertices

loop – vertex related to itself

Parallel Edges – multiple routes b/w same vertices

edges.

Multi-graph : loops & parallel allowed

Simple graph : NO loops / parallel edges

Digraph - directed graph (dir^n on edges)

Walks - Sequence of vertices (adjacent)

Length of walk - no. of steps.

Path - walk with distinct vertices

Connected Graph Disconnected Graph

Cycle - start vertex is the end ver.

Degree (of a vertex) - no. of edges that include vertex

Regular Graph - If every vertex has same degree k.

k-regular graph.

HANDSHAKING LEMMA

In any graph,

Sum of degrees = 2 × (no. of edges)

⇒ Even no. of vertices of ODD degree.

TREE

Connected graph with no cycles

Prop (s) – If one true, all true:
(i) T is a tree
(ii) Any 2 vertices – linked by unique path
(iii) Del^{n} of an edge – DISCONNECTED
(iv) Adding an edge – CYCLE

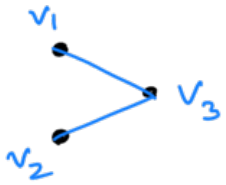
n vertices — (n-1) edges

SPANNING TREE

a tree contained in graph (includes all vertices)

Every connected graph contains a ST.

ADJACENCY MATRIX



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

$$a_{ij} = \begin{cases} 0 & \text{adjacent} \\ 1 & \text{Not adjacent} \end{cases}$$

EULER CIRCUIT

closed walk — uses every edge once

A connected graph is Eulerian iff every vertex has even degree

Euler Trail — start and end are diff

A connected graph has Euler Trail iff at most 2 vertices have odd degrees

STATISTICAL DATA MODELLING

model — rep^{n} of real world prob.
— allows exp^{n}, analysis
— probabilistic

Keywords — probability
expected value
variance
chebychev ineq.
central limit thm.
hypothesis
predictive modelling

frequentist approach — counting

median = $x_{(n+1)/2}$ or $\frac{1}{2}(x_{n/2} + x_{n/2+1})$

n is odd n is even

quartile (Q_k) = $x_p + \frac{q}{4}(x_{p+1} - x_p)$

p = floor $((k(n+1))/4)$

$q = (k(n+1)) \bmod 4$

variance $(s_x^2) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$

standard deviation = s_x

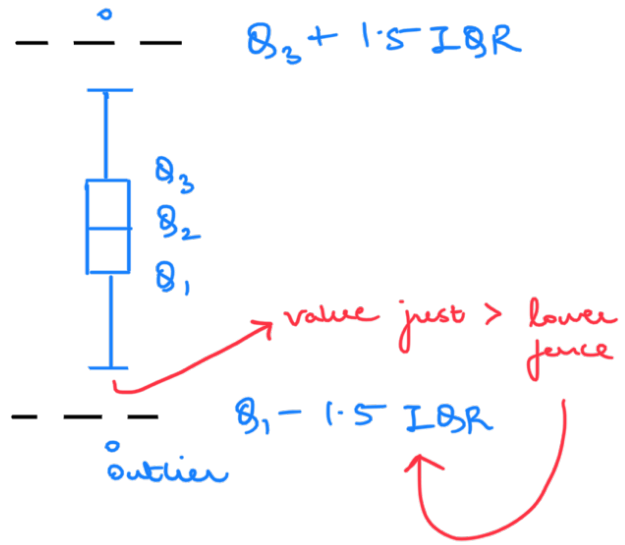
range = $\max(x_i) - \min(x_i)$

$IQR = Q_3 - Q_1$

covariance $(q_{xy}) = \frac{1}{n-1} \Sigma (x_i - \bar{x})(y_i - \bar{y})$

correlation coeff. $(r_{xy}) = \frac{q_{xy}}{s_x s_y}$

Boxplot.



PROBABILITY

random variables (experiment output)

randomness due to unmeasured factor
ex- reading error.

confinding var — hidden factor that contributes in computing output
(unsupervised)

majority of times unable to trace but if found it can be used to predict output → hypo. testing

$P(X = x)$, $x \in \mathcal{X}$

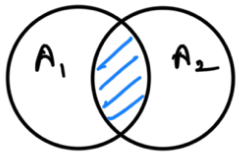
prob that a RV X takes on the value x from $\mathcal{X}$

also $0 \leq P(X = x) \leq [0,1] \quad \forall x \in X$

also ① $P(x = x) = [0,1] \quad \forall x \in X$

and $\sum_{x \in X} P(x = x) = 1$

② $P(X \in A_1 \cup A_2) = P(x \in A_1)$



$+ P(x \in A_2)$

$- P(x \in A_1 \cap A_2)$

Similarly for 2 RVs:

$P(X = x) = \sum_{y \in Y} P(X = x, Y = y)$

marginal pr. irrespective of Y — joint pr.

$\sum_{x \in X,\ y \in Y} P(X = x, Y = y) = 1$ ——— from ①

Conditional Probability

$P(X = x \mid Y = y) = \frac{P(x = x, Y = y)}{P(Y = y)}$

Bayes Theorem

$P(x \mid A) = \frac{P(A \mid x)\ P(x)}{P(A)}$

① if $P(X = x, Y = y) = P(X = x)\, P(Y = y)$

$\Rightarrow$ X and Y are independent

$\Rightarrow P(X = x \mid Y = y) = P(X = x)$

② a special case:

Independent & identically dist. (IID):

$X_1 \in X, \quad X_2 \in X$

$$P(X_1 = x) = P(X_2 = x) \quad \forall \quad x \in \mathcal{X}$$

Continuous RV

- prob. density funct^n (pdf)

$$P(X = x) = f(x)$$

$$f(x) > 0 \quad \forall \quad x \in \mathcal{X}$$

$$P(a < X < b) = \int_a^b f(x) \cdot dx$$

$$P(X \in A) = \int_A f(x) \cdot dx$$

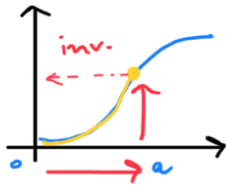
$$P(X = x) = 0 \qquad \text{(slide 39)}$$

$$\int_{\mathcal{X}} f(x)\, dx = 1$$

- cumm. density funct^n (cdf)

$$P(X \leq x) = \int_{-\infty}^{x} f(x) \cdot dx$$

$$F(x) = \int_0^a f(x)\, dx$$



- inverse cdf (quantile funct^n)

$$Q(x) = F^{-1}(x)$$

EXPECTATION

Expected value of a dist^n

↳ some char. known

$$E[X] = \sum_{x \in \mathcal{X}} x\, p(x) \qquad \text{discrete}$$

$$= \int x \cdot p(x)\, dx \qquad \text{cont.}$$

$E[X^n]$ — n^{th} moment

$E[(X - E[X^n]]$ — n^{th} central moment

Variance — variation around the mean.

↳ expected squared deviation

$$V[X] = E[(X - E[X])^2] = E[X^2] - E[X]^2$$

$$= \sum_{x \in \mathcal{X}} (x - E[X])^2 p(x)$$

(positive only; $E[X]$ ~ mean of x)

Standard Deviation

$$\sigma_x = \sqrt{V[X]}$$

Linearity of Expectation:

$$E[cX] = cE[X]$$

$$V[cX] = c^2 V[X]$$

$$E[c] = c$$

$$V[c] = 0$$

$$E[af(X) + bg(Y)] = aE[f(X)] + bE[g(Y)]$$

$$E[g(X)E[f(X)]] = E[f(X)]E[g(X)]$$

Covariance $(-\infty, \infty)$

$$Cov(X, Y) = E[(X - E[X])(Y - E[Y])]$$

$$= E[XY] - E[X]E[Y]$$

→ joint variability

0 1

Correlation $[-1, 1]$

$$\text{corr}(X,Y) = \frac{\text{cov}(X,Y)}{\sqrt{V[X]\,V[Y]}}$$

→ rel^{n} b/w the 2

$$Z_X = \frac{X - E[X]}{V[X]}$$

$$\Rightarrow \quad \text{corr}(X,Y) = \text{cov}(Z_X, Z_Y)$$

- strength of linearity
- +ve or -ve
- slope not relevant (due to standardising)

if X and Y independent then $\text{cov}(X,Y) = \text{corr}(X,Y) = 0$ } Converse NOT true

$$\Longrightarrow \quad E[XY] = E[X]\,E[Y]$$

$$V[X+Y] = V[X] + V[Y]$$

Chebyshev's Inequality

if X is a RV with mean(μ) and var(σ^2) then for any $k > 0$:
↳ no. of std. dev^{n} from μ

$$P\left(\frac{|X-\mu|}{\sigma} \geq k\right) \leq \frac{1}{k^2}$$

only mean & var. must be known
→ weak inference
general but not always correct (slide 36)

Weak law of large no.(s)

mean of a sample of RVs converges into expected value as sample size grows larger

$$P\left(\left|\frac{x_1 + x_2 - \cdots - x_n}{n} - \mu\right| > \varepsilon\right) \to 0 \quad \text{as } n \to \infty$$

where $E[X_i] = \mu$ & $\varepsilon > 0$

lot of data — no need to look for TRUE mean & variance.

DISTRIBUTIONS

prob. dist^n as models

$$P(X = x \mid \theta) = p(x \mid \theta)$$

(θ: parameters)

θ changes — dist^n changes

— Gaussian

$X = \mathbb{R}$

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

(μ: mean, σ^2: var.)

$X \sim N(\mu, \sigma^2)$

Std. Normal — $N(0, 1)$

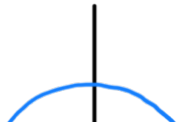
if $Z \sim N(0,1) \Rightarrow X = \mu + \sigma Z$

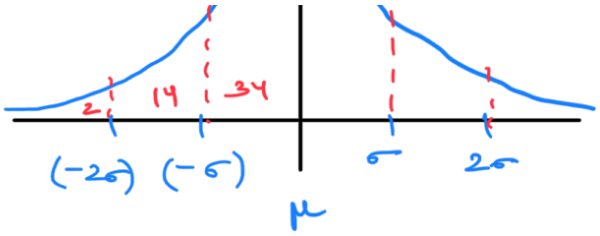
$E[X] = \mu = \int x \cdot p(x)\, dx$

$V[X] = \sigma^2$

mode = μ = median

almost all dist^n converge to gaussian as sample size increases!





- Bernoulli

$\mathcal{X} = \{0, 1\}$

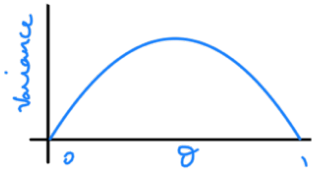
$p(x=1 \mid \theta) = \theta$ $\quad \theta \in [0,1]$

Success pr.

$p(x|\theta) = \theta^x (1-\theta)^{(1-x)}$

$E[x] = \theta$

$V[x] = \theta(1-\theta)$



Variance is highest when $\theta = 1/2$

- Binomial

n Bernoulli trials — m successes

$m \in \{0, 1, \ldots n\}$

$p(m \mid n, \theta) = \binom{n}{m} \theta^m (1-\theta)^{(n-m)}$

$E[x] = n\theta$

$V[x] = n\theta(1-\theta)$

conditions : n is finite
θ is constant
independent events

dist^n is skewed towards $n\theta$

- Uniform

when outcomes are equally likely

a) discrete

$$P(X=k \mid a,b) = \frac{1}{b-a+1}$$

$X \in \{a, \dots b\}$ $\quad b \geq a$

$$E[x] = \frac{a+b}{2}$$

$$V[x] = \frac{(b-a+1)^2 - 1}{12}$$

b) continous

$$p(x \mid a,b) = \begin{cases} 0 & x < a \\ \frac{1}{b-a} & a \leq x \leq b \\ 0 & x \geq b \end{cases}$$

$$E[x] = \frac{a+b}{2}$$

$$V[x] = \frac{(b-a)^2}{12}$$

— Poisson

$x \in Z_+$

$$p(x \mid \lambda) = \frac{\lambda^x e^{-\lambda}}{x!} \qquad \lambda = \text{rate}$$

$E[x] = \lambda$

$V[x] = \lambda$

conditions
- events are independent
- 2 events do not occur simultaneously
- λ is constant

STATISTICAL INFERENCE

while we may have idea about the shape of dist^n θ is generall unknown

so we estimate pop^n parameters (θ)

Sum of Squared Errors — SSE

$\hat{\mu} =$ arg min $\{\sum^n (\mu_i - \mu)^2\}$

estimator of μ [illegible]

$$SSE(\mu) = \sum_{i=1}^{n}(y_i - \mu)^2$$

$$\frac{d}{d\mu} SSE(\mu) = 0 \Rightarrow \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

sample mean

but no soln for σ

Max. Likelihood Est^n — MLE

$$\hat{\theta} = \arg\max \{ p(y|\theta) \}$$

① $\frac{\partial}{\partial \mu} l(y|\theta : \mu, \sigma) = 0$

$\Rightarrow \hat{\mu} = \frac{1}{n}\sum y_i$ (same as sample mean)

② $\frac{\partial}{\partial \sigma} l(y|\theta : \mu, \sigma) = 0$

$$\Rightarrow \hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \mu)^2}$$

DIFF from sample var.

while $\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \mu)^2}$

↓ unbiased

μ_{MLE} vs $\mu_{unbiased}$?

→ Comparing estimators

prop. accuracy

estimation is data dependent

∴ we do sampling — confidence interval
— hypo. testing
— compare estimators

assuming $Y_1, Y_2 \ldots Y_n$ follow a parametric dist^n $p(y|\theta)$ → population parameters

an estimator $\hat{\theta}(Y_1, \ldots Y_n)$ is a funct^n

if $Y_1, Y_2, \ldots, Y_n$ [illegible] $f(y|\theta)$

$\Rightarrow$ the estimator follows a dist^n determined by $f(y|\theta)$

for ex- if $Y \sim N(\mu, \sigma^2)$

then $\bar{Y} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$

↓

sample means.

Estimator Bias

$$B_\theta(\hat{\theta}) = E[\hat{\theta}(Y)] - \theta$$

if bias = 0 then estimator is unbiased

Variance of estimator

$$V_\theta(\hat{\theta}) = E\left[(\hat{\theta}(Y) - E[\theta(Y)])^2\right]$$

high bias low var vs high var low bias ?
net effect – MSE

Mean Squared Error

$$MSE_\theta(\hat{\theta}) = E\left[(\hat{\theta}(Y) - \theta)^2\right]$$

$$= b_\theta(\hat{\theta})^2 + V_\theta[\hat{\theta}]$$

an estimator is consistent if

$$\left.\begin{array}{l} b_\theta(\hat{\theta}) \to 0 \\ Var_\theta(\hat{\theta}) \to 0 \end{array}\right\} \text{ as } n \to \infty \ \forall \ \theta$$

if $Y_1, \ldots Y_n$ has $E[Y_i] = \mu$ & $V[X_i] = \sigma^2$

$b_\mu(\hat{\mu}_{ML}) = 0$ $\quad Var_\mu(\hat{\mu}_{ML}) = \frac{\sigma^2}{n}$

and thus $MSE = \sigma^2/n$

CENTRAL LIMIT THEOREM

Central Limit Theorem

$$\sum_{i=1}^{n} Y_i \xrightarrow{d} N(n\mu, n\sigma^2)$$

and $\bar{Y} \xrightarrow{d} N(\mu, \sigma^2/n)$ as $n \to \infty$

a lot of distributions converge into normal

$bin(\theta, n) \approx N(n\theta, n\theta(1-\theta))$
$Pois(\lambda) \approx N(\lambda, \lambda)$
any $Poi(\lambda)$ RV is sum of λ $Poi(1)$ RV

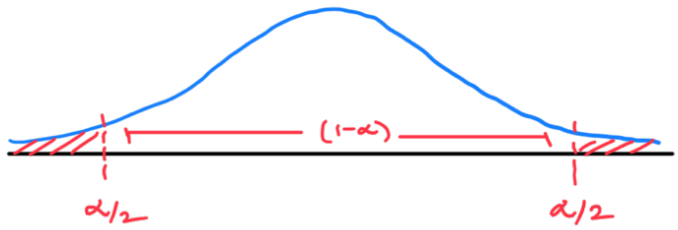
Confidence Interval

as point estimation is not highly accurate due to randomness we use interval est^n.

$$\hat{\theta}_m(y) \longrightarrow (\hat{\theta}^-(y), \hat{\theta}^+(y))$$
point → interval

i.e. 2 sided CI: $[\theta(\alpha/2), \theta(1-\alpha/2)]$

we are $(1-\alpha)$ confident $Z \sim (0,1)$ falls inside $(-z_{1-\alpha/2}, z_{1-\alpha/2})$



$$\Rightarrow [-z_{1-\alpha/2}, z_{1-\alpha/2}]$$

CI for Normal Means

① unknown μ, known σ^2

—

$$\hat{\mu}_{ML} = \bar{Y} = \frac{1}{n}\sum_{i=1} y_i \quad \text{(Sample mean)}$$

$$\left(\hat{\mu}_{ML} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \hat{\mu}_{ML} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

(2) unknown μ, unkown σ^2

we can not assume sample mean to be equivalent to population mean

Sample variance.

$$\begin{cases} \hat{\sigma}_\mu^2 = s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{Y})^2 \\ \text{but is no longer Normally dist.} \end{cases}$$

So we use student-t dist^n

$$\left(\hat{\mu}_{ML} - t_{\alpha/2,n-1}\frac{s}{\sqrt{n}},\ \hat{\mu}_{ML} + t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}\right)$$

HYPOTHESIS TESTING

- propose a null hypothesis (H_0) and an alternate one (H_A)
- try to reject H_0 using p-value
- low p-value : evidence against Null

Goal — Reject Null.

$$p = \begin{cases} 2P(-|x|) & H_0 \text{ is } = \\ 1 - P(x) & \leq \\ P(x) & \geq \end{cases}$$

just like in C1, σ is either

(1) known

$$z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

calculate pr. under the curve :
$p < 0.01$ STRONG
$0.01 < p < 0.05$ MODERATE
$p > 0.05$ WEAK evidence against NULL

(2) unknown

$$t_{n-1} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

unlike above, we don't care about p-value here

if $|t_{\hat{\mu}}|$ > required t-score then Reject H_0

else we can't reject H_0

REGRESSION & CORRELATION

Linear Regression

$$y = \beta_0 + \beta_1 x$$

$$SS_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = n(\overline{x^2} - \bar{x}^2)$$

$$SS_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = n(\overline{xy} - \bar{x}\bar{y})$$

$$SS_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = n(\overline{y^2} - \bar{y}^2)$$

$$\beta_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2}$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x} = \frac{\bar{y}\,\overline{x^2} - \overline{xy}\,\bar{x}}{\overline{x^2} - \bar{x}^2}$$

$$RSS(\hat{\beta}_0, \hat{\beta}_1) = \frac{SS_{yy} SS_{xx} - SS_{xy}^2}{SS_{xx}}$$

$$= SS_{yy} - SS \hat{\beta}^2$$

$$SS_{yy} \quad SS_{xx} \quad p_1$$

$$R^2 = 1 - \frac{RSS}{SS_{yy}} = \frac{SS_{xy}^2}{SS_{xx} SS_{yy}} = r_{xy}^2$$

$$s.e.(\beta_0) = \sqrt{\frac{RSS}{n(n-2)} \frac{\overline{x^2}}{\overline{x^2} - \bar{x}^2}}$$

$$s.e.(\beta_1) = \sqrt{\frac{RSS}{n(n-2)} \frac{1}{\overline{x^2} - \bar{x}^2}}$$

Multi Linear Reg.

$$E[y_i | x_{i,1} \cdots x_{i,p}] = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}$$

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & & & & \\ \vdots & x_{21} & x_{22} & \cdots & x_{2p} \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

CLASSIFICATION

Naive Bayes.

$$P(y | x_1 \cdots x_p) = \frac{P(y) \prod_{j=1}^{p} P(x_j | y)}{P(x_1 \cdots x_p)}$$

Logistic Regression

$$\alpha_i = w_0 + w_1 x_{i1} + \cdots w_p x_{ip}$$

$$\alpha_i = w_0 + \sum_{j=1}^{p} w_j x_{ij}$$

$$\hat{y}_i = \frac{1}{1 + e^{-\alpha_i}} \quad \text{bound to } [0, 1]$$

$$\text{cost}(w) = \frac{-1}{n} \sum_{i=1}^{n} y_i \log \hat{y}_i + (1 - y_i) \log (1 - \hat{y}_i)$$

-ve log likelihood / log-loss

$$dw_j = \sum_{k=1}^{p} (\hat{y}_i - y_i)\, x_{ik}$$

$$w_j = w_j - dw_j \times \eta$$

η — learning rate

w_0 — bias parameter

if $w_0 < 0$: class 0
$w_0 > 0$: class 1

$$\Rightarrow p(Y_i = 1 \mid x_{i1}, \ldots, x_{ip}) = \frac{1}{1 + e^{\eta_i}}$$

$$\eta_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}$$

Decision Trees

$$H(x) = E\left[\log_2 \frac{1}{p(x)}\right]$$

$$H(x|y) = \sum_{y} p(y = y)\, H(x \mid y = y)$$

Information Gain = $H(x)$ − $H(x|y)$
initial — after split

maximise IG by minimising $H(x|y)$

if x, y are independent :

$$H(x, y) = H(x) + H(y)$$

MACHINE LEARNING

Linear Regression

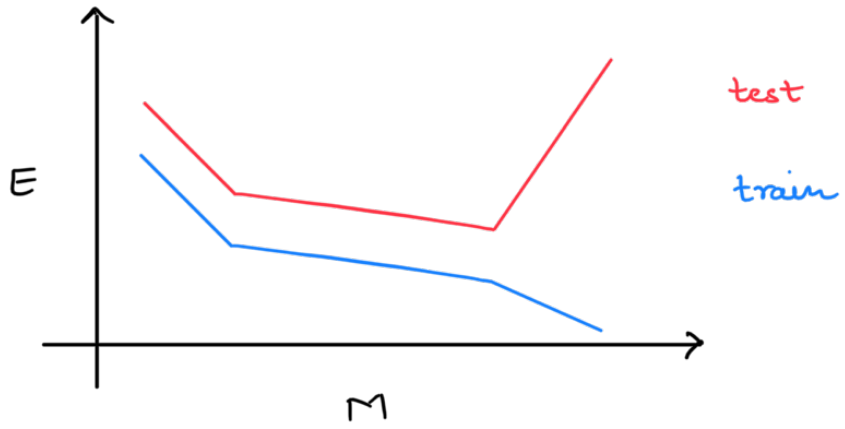
$$y(x, w) = \sum_{j=0}^{M} w_j x^j$$

i.e. $w_0 x^0 + w_1 x^1 + w_2 x^2 - - - w_M x^M$

- Linear w.r.t weights (w_j) not parameters (x)
- Parametric Model : no. of param.s are fixed

  (ex. of non param — KNN where K may change)
- M refers to model complexity

  i.e. as order increases, complexity increases.



underfitting overfitting

- To overcome overfitting :

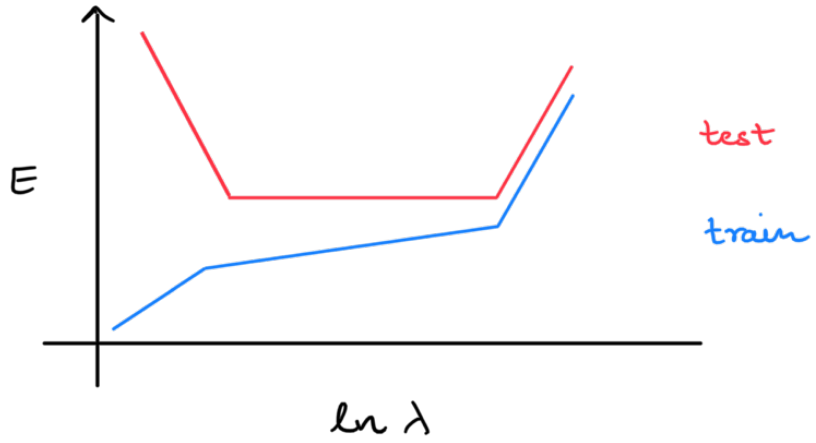
  1. Increase Training data

2. Regularisation

$$E(w) := \frac{1}{2}\sum_{n=1}^{N}[y(x_n, w) - t_n]^2 + \frac{\lambda}{2}\|w\|^2$$

i.e. adding a penalty

∴ trade off between error and complexity



- Optimisation:

$$w^* := \arg\min_w E(w)$$

- Model Selection:

1. Hold-out Validation

→ wasteful of training set

2. K-Fold Cross Val.

→ complicated but good

uses avg. of val. errors for est.

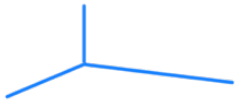
3. Leave-One-Out

→ Special case of K-fold

K is set to no. of training points

- Uncertainty due to : 1. Noise
2. finite dataset

can be measured using probability



Bootstrap Bayesian

Bootstrapping –

We generate w for given D but will this w work for D' ?

i.e. generalisation problem due to finite dataset

Sol^n : generate alternate datasets (D') from original (D)

→ random sel^n of N datapoints one by one with replacement to create bootstrap sample

i.e. simulate infinity

Bayesian Approach

$$p(M|D) = \frac{p(D|M)\,p(M)}{p(D)}$$

i.e. $$p(w|D) = \frac{p(D|w)\,p(w)}{p(D)}$$

posterior $\propto$ likelihood $\times$ prior

$p(D|w) = w^{|H|}(1-w)^{|T|}$ MLE

$$p(w) \propto \underbrace{w^{a-1}(1-w)^{b-1}}_{\text{beta dist}^n} \qquad \text{prior . belief}$$

$$\Rightarrow p(w|D) \propto w^{|H|+a-1}(1-w)^{|T|+b-1}$$

$$\propto \text{Beta}(w|\ |H|+a, |T|+b)$$

$$E[w|D] = \frac{|H| + a}{|H| + |T| + a + b}$$

and $p(D) = \int p(D, w)\, p(w)$

$$= \int p(D|w) \cdot p(w)\, dw$$

$a = b =$ ↑ strong belief
fair coin

$a = b =$ ↓ weak belief

$a >> b$ biased coin

- Basis Function

as stated earlier $y(x, w)$ has to be linear funct^n of w not x

thereby $y(x, w) = \sum_{j=0}^{M-1} w_j \phi_j(x)$

where $\phi_j \equiv$ basis funct^n

and can be non linear as well

popular basis functions :

1. Gaussian

$$\phi_j(x) = e^{\left(-(x-\mu_j)^2/2s^2\right)}$$

2. Sigmoidal

$$\phi_j(x) = \sigma\left(\frac{x - \mu_j}{s}\right)$$

3. Hyperbolic tangent

$$\tanh(a) = \frac{1 - e^{-2a}}{1 + e^{2a}}$$

- Loss function (error or misfit)

Data $D = \{(x_n, t_n)\}_{n=1}^{N}$

Function : $y(x, w)$

Noise : $\varepsilon = \mathcal{N}(0, \sigma^2)$

t_n : $y(x, w) + \varepsilon$

$$p(t \mid x, w, \sigma^2) = \mathcal{N}(t \mid y(x, w), \sigma^2)$$

- Likelihood Function

$$p(t_1 \dots t_N \mid x_1 \dots x_N, w, \sigma) = \prod_{n=1}^{N} \mathcal{N}(t_n \mid y(x_n, w), \sigma^2)$$

- Log - likelihood function

$$\mathcal{L}(w) := \log p(t_1 \dots t_N \mid x_1 \dots x_N, w, \sigma)$$

$$= \log \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} e^{\left(-\frac{1}{2\sigma^2}[t_n - w \cdot \phi(x_n)]\right.}$$

$$= N \log \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{2\sigma^2} \sum_{n=1}^{N} [t_n - w \cdot \phi(x_n)$$

Const. wrt w | SSE

MLE under Gaussian assumptn
→ Rigorous choice

$$\nabla L(w) = \begin{bmatrix} \frac{\partial L(w)}{\partial w_0} \\ \vdots \\ \frac{\partial L(w)}{\partial w_{M-1}} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

∴ optimal solutions requires solving multiple equations

→ can be solved using matrix eqn but expensive.

soln : Iterative optimizatn algo (s)

GD SGD

- Gradient Descent Algorithm

* initialise $w^{(0)}$ and $t = 1$
* while (stopping condn not met) do:
  * $\eta' = \eta$
  * while $\eta' > \varepsilon$ do
    * $w := w^{(t-1)} - \eta' \nabla E(w^{(t-1)})$
    * if $E(w) < E(w^{t-1})$ then break
    * $\eta' = \eta'/2$
  * $w^t := w$
  * $t = t + 1$

processing entire data set in each step is very expensive.

∴ when large dataset or entire data not accessible, we use SGD

- Stochastic Gradient Descent

* initialise $w^{(0)}$ and $t=1$
* while (stopping cond^n not met) do:
  * for each training point:
    * $\eta' = \eta$
    * $w^{\tau} := w^{\tau-1} - \eta' \nabla E(w^{\tau-1})$
    * if $E(w) < E(w^{t-1})$ then break
    * $\eta' = \eta'/2$
  * $w^t := w^{\tau}$
  * $t = t+1$

- Regularisation

1. Ridge (L_2- norm)

$$\frac{1}{2}\sum_{n=1}^{N}\left(t_n - w\cdot\phi(x_n)\right)^2 + \frac{\lambda}{2}\sum_{j=1}^{M-1}(w_j)^2$$

2. Lasso (L_1- norm)

$$\frac{1}{2}\sum_{n=1}^{N}\left(t_n - w\cdot\phi(x_n)\right)^2 + \frac{\lambda}{2}\sum_{j=1}^{M-1}(w_j)$$

Both shrink the w but Lasso also allows sel^n of input var(s)

∴ Lasso gives sparse sol^n (s) but

useful when 'large no. of input var(s)'

→ ignore those with 0 params.

for optimal soln:

Ridge: Matrix opn

LASSO: NON Differentiable

∴ sub-gradient.

- Bias Variance Analysis

Bias refers to accuracy

i.e. comparison of predicted values to true value/label

Variance refers to consistency

i.e. how the predictions for each data sample compare with each other

predicted value refers to avg or expected value.

∴ low bias, low variance

- Generalisation Error

on one data-set:

$$[y(x, D) - h(x)]^2$$

∴ over all datasets:

$$E_D [y(x, D) - h(x)]^2$$

$$= E_D\left[y(x,D) - E_D[y(x,D)] + E_D[y(x,D)] - h(x)\right]^2$$

$$= E_D\left[\left(y(x,D) - E_D[y(x,D)]\right)^2 + \left(E_D[y(x,D)] - h(x)\right)^2 + 2\left(y(x,D) - E_D[y(x,D)]\right)\left(E_D[y(x,D)] - h(x)\right)\right]$$

$$= E_D\left[\left(y(x,D) - E_D[y(x,D)]\right)^2\right] + E_D\left[\left(E_D[y(x,D) - h(x)\right)^2\right]$$

$$= E_D\left[\left(y(x,D) - E_D[y(x,D)]\right)^2\right] + \underbrace{\left[E_D[y(x,D) - h(x)\right]^2}_{\text{bias}^2}$$

$E_D\left[\left(y(x,D) - E_D[y(x,D)\right]^2\right)$ ↓ variance

$\therefore$ general^ error = bias^2 + variance

- Model flexibility

bias-variance trade off

rigid model : low var. high bias

flexible model : high var low bias

managing bias and variance :

if high test error ⇒ overfitting

⇒ M↑ ⇒ too flexible model

⇒ high variance, low bias

in such a case, increase training data or reduce input features

else if high bias, add new features

Classification

- terminology : input space

decision boundary (k-1) dim

decision regions (labels)

linear model

linear separability

binary/multiclass

decision boundary contains all points on $f(x, w) = 0$

$\Rightarrow w_0 + w_1 x_1 + w_2 x_2 - - - w_d x_d = 0$

if d dim input space, d-1 for boundary

→ line for 2D
hyperplane for 3D

- Type of Models

Discriminative	Generative
* assign x to C_k	$P_w(C_k \mid x)$ for each C_k
* direct modelling	indirect, gen. x for C

$p(C_k|x) \sim p(x|C_k) \cdot p(C_k)$

likelihood prior

- Discriminative Models

1. 2 class discriminant $f(x)$

C_1 if $y(x) \geq 0$ } sign based

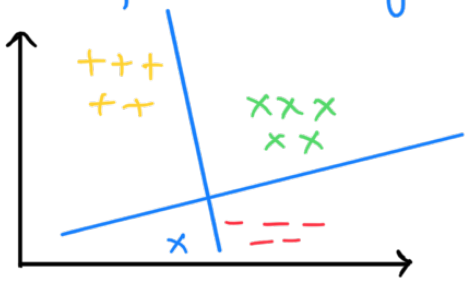
C_2 otherwise

$y(x, w) = 0$ is $\perp$ to w

2. Multiclass prob

A) One vs Rest

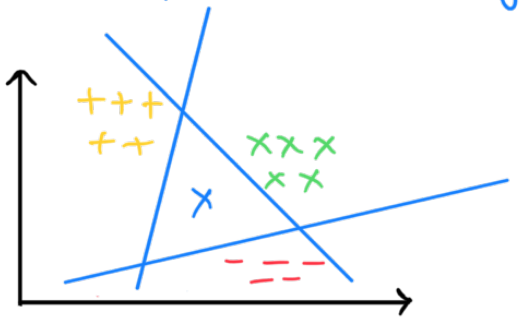
$(K-1)$ disc. classifier
each sep. one C_K from rest



ambigous region X

B) One vs One

$(k(k-1))/2$ disc. classifiers



C) K Discriminant $f(x)$ — $y_K(x)$

use magnitude instead of just sign of $y_k(x)$

for each k

$$y_k(x) = \tilde{w}_k^T x + w_{k0}$$

i.e. put x to c_k if

$$y_k(x) > y_j(x)$$

$$\Rightarrow \quad y = \underset{k}{\text{argmax}}\ y_k(x)$$

- Generalised Linear Model

applying non-linear funct^n on lin regression

$$y(x) = f(w_0 + w^T x)$$

$f \equiv$ activation funct^n, examples :

Step (perceptron)

$$f(a) = \begin{cases} 1 & \text{if } a \geq 0 \\ -1 & \text{else} \end{cases}$$

logistic

$$f(a) = \frac{1}{1 + e^{-(a)}}$$

Perceptron : gen. lin model

simplest neural n/w

single neuron

→ weighted sum of inputs

→ threshold the sum using step $f(x)$

input : x, param : w

decision boundary : $w^T x = 0$

or simply $\phi(x) = x$

model : $y = \begin{cases} +1 & \text{if } \vec{w}\vec{x} \geq 0 \\ -1 & \text{else} \end{cases}$

objective : decision surface (wt·s) to minimize misclassified data points (error fn)

$E(w) = -\sum_{n \in M} w \cdot x_n t_n$

$\nabla E(w) = -x_n t_n$

algo : (i) randomly initialise w

(ii) repeat until all data points are correctly classified or error $< \varepsilon$:

* for each data pt. :

* compute y

* if (y=t), ok

* else update wt.s :

$w \leftarrow w + \eta t_n x_n$

perceptron only works for linearly separable data

to find perfect weight vec(s) it may take a lot of iterations

it may NOT converge if data is not linearly separable i.e. may just cycle through wt.(s) w/o stopping

sensitive to initialisation i.e. diff. wt. vec(s) if data points visited in diff order.

- Probabilistic Generative Models

model $p(x|C_K)$ to generate input data

Bayes classifier

$$p(C_K|x) = \frac{p(x|C_K)\cdot p(C_K)}{p(x)}$$

$$\text{argmax}_{C_K} p(C_K|x) = \text{argmax}_{C_K} p(x|C_K)\cdot p(C_K)$$

class conditional PDF prior

- $p(C_K) = \phi^{C_K}(1-\phi)^{1-C_K}$ Bernoulli

- $p(x|C_K) = \frac{1}{\sqrt{2\pi}\sigma_{C_K}}\left(-(x-\mu_{C_K})^2/2\sigma_{C_K}^2\right)$

considering x is continuous, we assume it is normally dist.

"Gaussian Bayes Classifier"

$\therefore$ model params $w = \phi, \mu_{C_K}, \sigma_{C_K}$

$$\prod_{n=1}^{N} p(x_n, C_K : \mu_{C_K}, \sigma_{C_K}, \phi)$$

$$= \prod_{n=1}^{N} p(x_n|C_K : \mu_{C_K}, \sigma_{C_K})\cdot p(C_K : \phi)$$

apply log & partial der. w.r.t to params.

do this sep. for MLE & prior as they have diff. params

$$\Rightarrow \quad \mu_{C_K} = \frac{1}{N}\sum_{n=1}^{N} x_n$$

$$\sigma^2 = \frac{1}{N}\sum_{n=1}^{N} (\quad)^2$$

$$\cdots \frac{1}{N} \sum_{n} (x_n - \mu_{c_K})$$

as $p(x|c_K) \approx p(x|\mu_{c_K}, \sigma^2_{c_K})$

Now for multivariate inputs, we have x as a vector

$x = (x_1 \cdots x_3)$

$\therefore p(x|c_K) \approx p(x|\mu_{c_K}, \Sigma)$

where Σ is the covariance mat.

and μ_{c_K} is diff for each c_K but Σ is same

Prediction rule:

1. $p(c_1|x) > p(c_2|x)$

$\Rightarrow p(x|c_1)\, p(c_1) > p(x|c_2)\, p(c_2)$

2. $a = \ln p(x|c_1)\, p(c_1)$

$\ln p(x|c_2)\, p(c_2)$

$\Rightarrow c_1$ if $a > 0$, c_2 otherwise

solving for a, we get linear form

$a = \frac{1}{2}(x-\mu_2)^T \Sigma^{-1}(x-\mu_2) - \frac{1}{2}(x-\mu_1)^T \Sigma^{-1} + \frac{\ln p(c$
$\ln p(c$

$\therefore$ Linear decision boundary

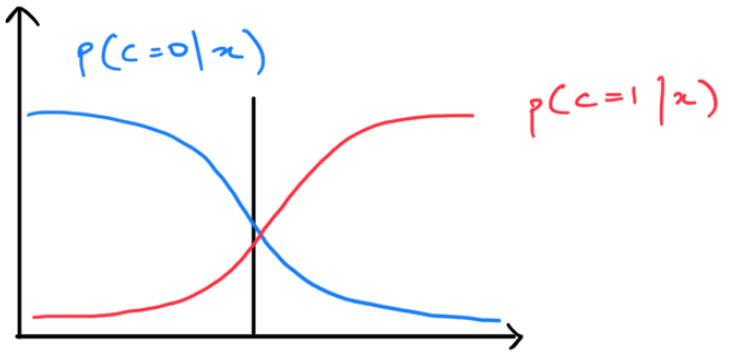
- Logistic Regression

applying sigmoid to linear $f(x)$ as:

$$y(x) = f(w^T x), \sigma(z) = \frac{1}{1+e^{-z}}$$

$$p(c=1|x) = \frac{1}{1+e^{(-w^T x)}} = \frac{e^{(w^T x)}}{1+e^{(w^T x)}}$$

$$p(c=0|x) = 1 - \frac{1}{1+e^{(-w^T x)}} = \frac{e^{(-w^T x)}}{1+e^{(-w^T x)}}$$



⇒ Linear decision boundary.

- Clustering

  hard : a data pt can only belong to 1 cl.

  Soft : a data pt may belong to more than 1 cluster

  ex- KMeans, DBSCAN, Hierarchial, Graph

- KMeans

  non-probabilistic model — only hard cl
  iterative algo. — simplest
  group similar data pt's into K clusters
  ↳ distance based
  1. random initialisⁿ of K cluster centres

  $\mu_1^0, \mu_2^0, - - - \mu_K^0$

2. repeat until stopping criteria:

(i) update data pt. assignment

- cal. distance from cl. centres
- assign to cluster with min di

(ii) update cluster centres

$$\mu_K^{(\tau+1)} = \frac{\sum_n r_{nk} x_n}{\sum_n r_{nk}}$$

$$r_{nK} = \begin{cases} 1 & \text{if } x_n \text{ assigned to } \\ 0 & \text{otherwise} \end{cases}$$

KMeans sensitive to initial cluster cent

→ high inter cluster similarity and low intra cluster.

- Gaussian Mixture Models.

if we are given labels before hand (complete data):

$$p(k, x_n) = p(x_n | k)\, p(k)$$

$$= \mathcal{N}(\mu_K, \Sigma_k) \cdot \varphi_K$$

but in reality, we do not have the labels i.e. incomplete data

$$\therefore p(k, x_n) = p(x_n | k) \sum_{k=1}^{K} p(z_n = k)$$

$$= \sum_{k=1}^{K} \varphi_K \mathcal{N}(x_n | \mu_K, \Sigma_K)$$

this is called GMM

$$\theta = (\phi, \mu_1, \Sigma_1 \dots \mu_K \Sigma_K)$$

for complete data (gaussian clf)

$$p(x,z) = \prod_{n=1}^{N} \prod_{k=1}^{K} p(x_n, k)$$

for incomplete data

$$p(x) = \prod_{n=1}^{N} p(x_n)$$

$$= \prod_{n=1}^{N} \sum_{k=1}^{K} p(x_n, z_n)$$

hard to get global solutions
∵ sum inside log

∴ EM (an iterative algo)

$$L(\theta) = \sum_{n=1}^{N} \ln \sum_{k=1}^{K} p(x_n, z)$$

$$= \sum_{n=1}^{N} \ln \sum_{k=1}^{K} p(z_k) p(x_n | z_k)$$

$$= \sum_{n=1}^{N} \ln \sum_{k}^{K} \varphi_k \mathcal{N}(x_n | \mu_k \Sigma_k)$$

$$\gamma(z_{nk}) := p(z_n = k | x_n)$$

$$= \frac{\varphi_k \mathcal{N}(x_n | \mu_k \Sigma_k)}{\sum_j \varphi_k \mathcal{N}(x_n | \mu_j \Sigma_j)}$$

when derivating w.r.t μ_k, Σ_k & φ_k
→ no sol^n

as all depend on posterior (γ_{nk})

∴ EM for $\ln p(x|\theta) = \ln \sum_z p(x, z|\theta)$

- Expectation Maximization

algo :

i) ... old

(i) choose initial θ

(ii) while convergence not met

E step : Evaluate $p(z|x,\theta^{old})$

M step : Evaluate θ^{new} by

$$\theta^{new} \leftarrow \text{argmax}_\theta \underbrace{\sum_z p(z|x,\theta^{old}) \ln p(x,z|\theta}_{Q(\theta,\theta^{old})}$$