Introduction to Data Science (Lecture)

02/09 (2)

Def An experiment is any activity that produces distinct outcomes.

* The set of all outcomes from an experiment is called Ω (sample space)
* subsets of Ω are called events. An event occurs if the result of the experiment $\omega \in \Omega$ if $\omega \in A$

ex: Toss a dice $\Omega = \{1,2,3,4,5,6\}$

$A = \{$"an odd number"$\} = \{1,3,5\}$ → Event

$\omega = 1 \Rightarrow$ A occurs / happens $\omega = 2 \Rightarrow$ A does not happen

$\omega = 3 \Rightarrow$

What's probability?

LTRF idea (Long Term Relative Frequency)

Let's say we repeat an experiment n times, we have an event A

$N(A, n) = \frac{\text{"number of times A happens"}}{n}$ → max. = 1, min = 0

ex: $A = \{H\}$ $\Omega = \{H, T\}$

Flip a coin 10 times and 5 heads $N(A, 10) = \frac{5}{10} = \frac{1}{2}$

LTRF $\lim_{n \to \infty} N(A, n) = p$ if you repeat the experiment a multiple of times then it becomes a number (p) → this is valid if you assume that the event are independent from each other

Rules of N:

1) $0 \leq N(A, n) \leq 1$, $N(\Omega, n) = 1$

2) Let A, B be two events $A \cap B = \emptyset$ (disjointed) then

$N(A \cup B, n) = N(A, n) + N(B, n)$ (addition rule)

3) Each experiment is independent of each other

single experiment → Let's assume that all events are "observable" $A \in \mathcal{F}$

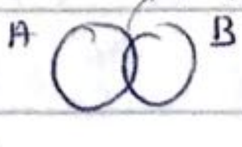
Def: A function $\mathbb{P}: \mathcal{F} \to [0,1]$ is a probability distribution if $\mathbb{P}$ satisfies rules 1,2

→ 2*) Let $A_1, A_2, \ldots, A_i \cap A_j = \emptyset$

$\mathbb{P}(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$

Rule 1,2 ⇒ a bunch of formulas

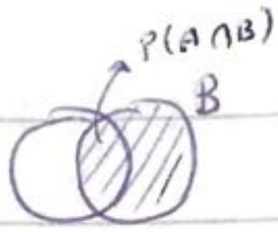
1) $\mathbb{P}(A^c) = 1 - P(A)$ A^c: ~~$\Omega \setminus A$~~ $\Omega \setminus A$

2) $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$  Inclusion / Exclusion Principle

3) $\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$ ~~Bodes~~ Boole's Principle / inequality

Conditional probability: (given)

$$\mathbb{P}(A|B) := \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \quad (1)$$



Independence: $\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B)$

$\Downarrow$ (1)

$\mathbb{P}(A|B) = \mathbb{P}(A)$ if $\mathbb{P}(B) > 0$

04/08 Vector Space:

vector = is an ordered list (measured in dimensions)

{bread, butter(kg), eggs}

$v = (3, 1, 7)$

$\mathbb{V}$ = ordered set = $\{v_1, v_2, v_3, \ldots, v_n\}$

$(x, y, z \to z\text{-axis}, \ldots)$

x-axis, y-axis

Properties:

$V = (v_1, v_2, v_3, \ldots, v_n) \in \mathbb{V}$; $u = (u_1, u_2, u_3, \ldots, u_n) \in \mathbb{V}$ ① $v+u = (v_1+u_1, v_2+u_2, v_3+u_3, \ldots, v_n+u_n) = u+v$

② if v, u, w: vectors $v+u+w = \underbrace{(v+u)}_{z} + w = z + w$

③ 0 ← zero vector $= (0, 0, \ldots, 0_n) \in \mathbb{V}$

④ $u + 0 = u$ identity element in $\mathbb{V}$

⑤ for every $u \in \mathbb{V}$ (there exists) $\exists\ -u \in \mathbb{V}$: $u + (-u) = 0$ → zero vector → the inverse of every vector should be included in the $\mathbb{V}$

R $\mathbb{V}$ $a \in R$ $u \in \mathbb{V}$ $au \in \mathbb{V}$ eg. $u = (1, 5, -2, \ldots, 7)$ eg. $a = 2$ $2 \cdot u = 2(1, 5, -2, \ldots, 7) = (2, 10, -4, \ldots)$

Warning! If $u \in \mathbb{V}$, $v \in \mathbb{V}$ we cannot write $u \cdot v$, when multiply 2 vectors, we don't get a vector but a scalar (dot product)

$a, b \in \mathbb{R}$ $u, v \in \mathbb{V}$ $a(u+v) = (au + av) \in \mathbb{V}$ $(a+b) \cdot u = a \cdot u + bu$

$a(bu) = (ab) \cdot u$

$a \cdot y$

* In R and $\mathbb{V}$ there are only two operations: multiplication and addition → not between two vectors though

~~you cannot add $v+v$~~

$f: \mathbb{V} \times \mathbb{V}$ this means one element of vector 1 and another from vector 2 can be added.

$u, v \xrightarrow{\oplus} u+v \in \mathbb{V}$

$g: R \times \mathbb{V}$

$a, v \xrightarrow{\odot} a \cdot v \in \mathbb{V}$

* If I have coordinates then the vector looks like: $\mathbb{V} = \{(x, 0), (0, y)\}$

vector spaces

$u = (2, 0) \in \mathbb{V}$, $v = (0, 3) \in \mathbb{V}$; $u+v \in \mathbb{V}$? $u+v = (2, 0) + (0, 3) = (2, 3) \notin \mathbb{V}$

if you combine then they are no longer a vector space

Intro to Data Science (Lecture)

c2: Def: A random variable as a mapping "function" $X : \Omega \to \mathbb{R}$ that assigns a real number to each outcome.

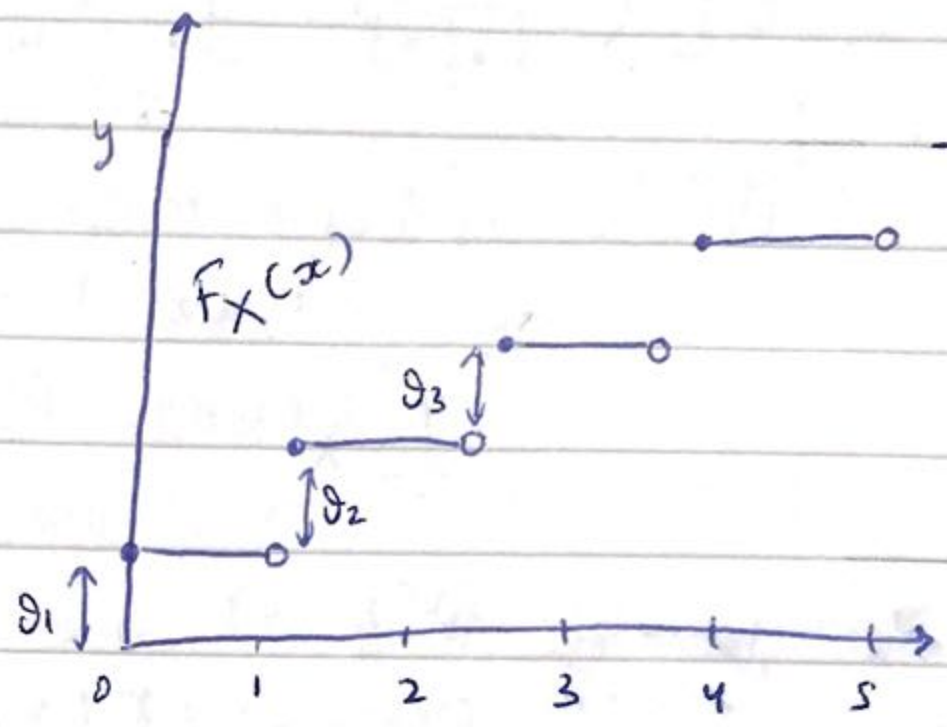
Ex.1 Flip a coin (0) times $\Omega = \underbrace{HHTT \ldots H}_{10 \text{ times}}$

~~$X(\Omega)$~~ $X(\omega)$ = "how many heads"

Ex.2 $\Omega = \{\text{all texts}\}$

$$X(\omega) = \begin{cases} 1, & \text{if text contains "free"} \\ 0, & \text{otherwise} \end{cases}$$

$X(\omega)$ = "count how many times free appears"



Distribution function:

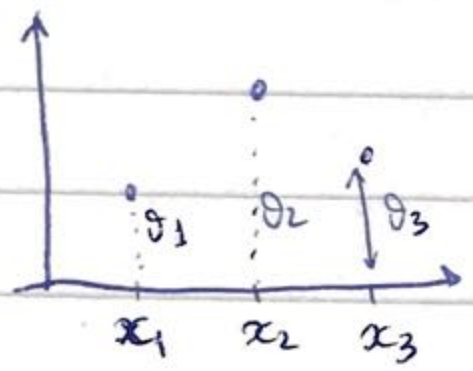
$X : \Omega \to \mathbb{R}$

$$\mathbb{P}(X \le x) = \mathbb{P}(\{\omega : X(\omega) \le x\}) =: F_X(x)$$

Def: A discrete random variable is a random variable that takes discrete ~~variable~~ values, ex. $(0, 1, 2, 3, \ldots)$. The PMF (Probability Mass Function).

$$f_X(x) = P(X = x) = \begin{cases} \vartheta_i, & \text{if } x = x_i \\ 0, & \text{otherwise} \end{cases}$$

X takes values $(x_1, x_2, \ldots)$



Formulas: ① $F_X(x) = P(X \le x) = P(\{X = x_1\} \cup \{X = x_2\} \cup \ldots \cup \{X = x_i\})$

$x_{R+1} > x$

$$= \sum_{x_i \le x} P(X = x_i) = \sum_{x_i \le x} f_X(x_i)$$

② $F_X(b) - F_X(a), \quad b \ge a = \sum_{a < x_i \le b} f_X(x_i)$

③ $\sum_{x_i} f_X(x_i) = 1$

Def: The Expected value of a discrete random variable X is $\mathbb{E}[X] := \sum_{x_i} x_i f_X(x_i)$

based on LTRF idea.

Let $z_1, z_2, \ldots z_{100}$ $\quad \sum_{i=1}^{100} z_i / 100 = \left(\sum_{z_i=1} z_i + \sum_{z_i=2} z_i + \ldots + \sum_{z_i=6} z_i\right) / 100$

$$= 1 \cdot N(z=1, 100) + 2 \cdot N(z=2, 100) + \ldots + 6 \cdot N(z=6, 100)$$

Common expectations: $V[X] := \mathbb{E}[(x - \mathbb{E}[X])^2] = \sum_{x_i} (x_i - m)^2 \cdot f_X(x_i)$

k : th central moments

$\mathbb{E}[(x - \mathbb{E}(X])^k]$

standard deviation $\sqrt{\mathbb{V}[X]} = \sigma(X)$ $\mathbb{E}[X] = \mu$

$$\mathbb{E}\left[\left(\frac{x - \mathbb{E}[X]}{\sqrt{\mathbb{V}[X]}}\right)^k\right] = \mathbb{E}\left[\left(\frac{x-m}{\sigma}\right)^k\right]$$ k : th standardised moment

$k=3$: skewness , $k=4$ = kurtosis

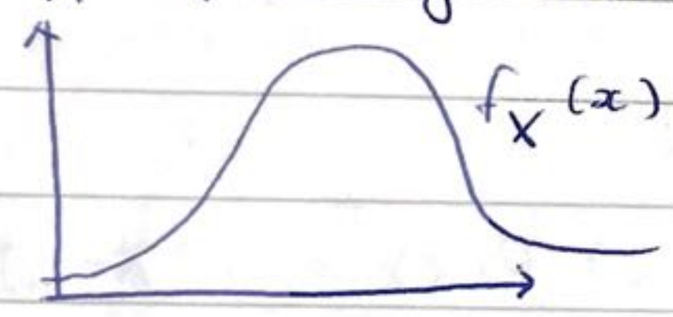
Def: We say that a random variable is continuous if there is a piecewise count function $f_X(x) : \mathbb{R} \to [0, +\infty]$ such that

$$F_X(x) = \int_{-\infty}^{x} f_X(s)\,ds = P(X \le x)$$

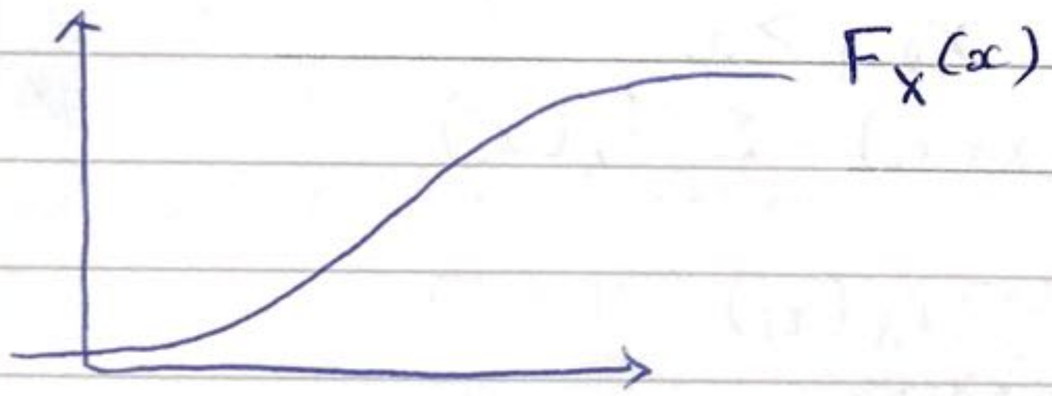
Warning: $\mathbb{P}(X = x) = 0$

$$\mathbb{P}(X = x) = \mathbb{P}(\{X \le x\} \setminus (x < x)\}$$
$$= \mathbb{P}(X \le x) - P(X < x)$$
$$= \int_{-\infty}^{x} - \int_{-\infty}^{x}$$
$$= \int_{x}^{x} f_X(s)\,ds = 0$$

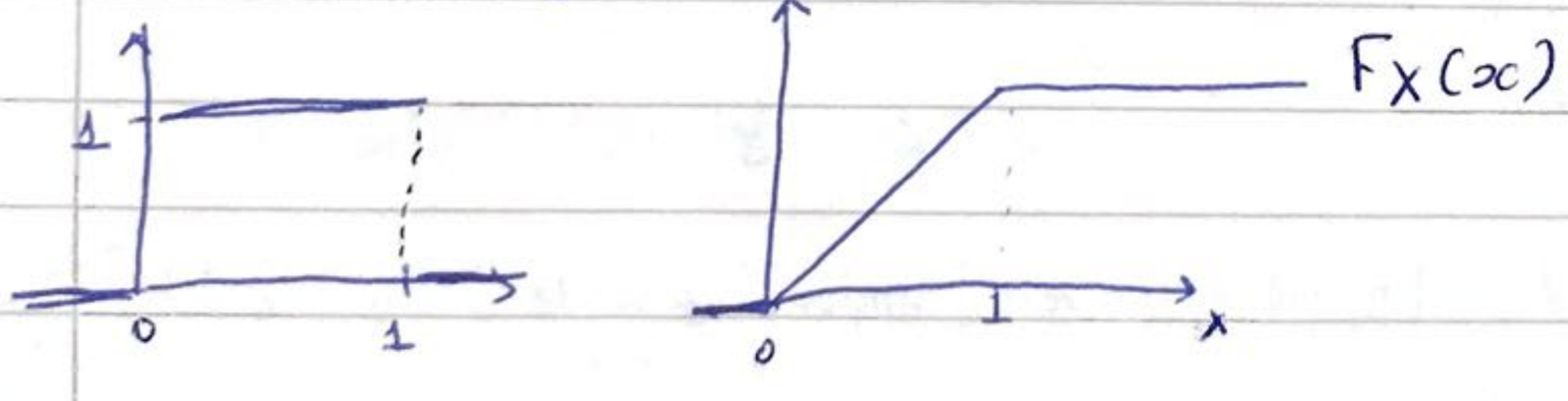
f_X : probability density



Normal / Gaussian



Uniform [0,1]



Transformation : X is discrete Random variable $g : \mathbb{R} \to \mathbb{R}$ $Y = g(x)$

Ex. $g(x) = (x-m)^2$

$F_Y(y) = \mathbb{P}(Y \le y) = \mathbb{P}(g(x) \le y)$ Let's say g is increasing (use the inverse) $= \mathbb{P}(X \le g^{-1}(y)) =$

$P(X \le g^{-1}(y)) = F_X(g^{-1}(y))$

Data Science (Lecture)

Lecture 3

Def: A multivariate Random Variable $\underline{X}$ is a mapping from Ω to $\mathbb{R}^m$ $\underline{X}: \Omega \to \mathbb{R}^m$ That is $\underline{X}$ represents m measurements

$F_{\underline{X}}(x_1, x_2, \ldots, x_m) = \mathbb{P}(\{\underline{X}_1 \leq x_1\} \cap \{\underline{X}_2 \leq x_2\} \cap \ldots \cap \{\underline{X}_m \leq x_m\})$

Joint Distribution Function (JDF)

Def: Let $F_{\underline{X}}(x_1, x_2)$ a JDF for $\underline{X}$ $\underline{X} = (\underline{X}_1, \underline{X}_2)$

$F_{\underline{X}_1}(x_1) := F_{\underline{X}}(x_1, +\infty) = \mathbb{P}(\{\underline{X}_1 \leq x_1\} \cap \underbrace{\{\underline{X}_2 \leq \infty\}}_{\text{all } \Omega}) = \mathbb{P}(\underline{X}_1 \leq x_1)$

↑ Marginal Distribution

↳ it is the 03-Random_Variables.ipynb → def F_X_1? because y takes ∞ which is the largest value (like the ∞ in the JDF)

Def: A 2-dimensions Random Variable $\underline{Z} = (\underline{X}, \underline{Y})$ we can say $\underline{X}$ and $\underline{Y}$ are independent if $F_{\underline{Z}}(x,y) = F_{\underline{X}}(x) \cdot F_{\underline{Y}}(y)$ for all $(x,y) \in \mathbb{R}^2$

Think $\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B)$

Def: Let $\underline{Z} = (\underline{X}, \underline{Y}) \in \mathbb{R}^2$ then the conditional PMF/Density (discrete or continuous)

$$f_{\underline{X}|\underline{Y}}(x|y) = f_{\underline{XY}}(x,y) / f_{\underline{Y}}(y),$$

density (little f) of X conditioned on Y $\quad f_Y(y) > 0$

eg. discrete ~~$f_{XY}(x,y)$~~ $\mathbb{P}(\underbrace{\{\underline{X} \leq x\}}_{A} | \underbrace{\{\underline{Y} = y\}}_{B}) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\sum_{x_i \leq x} \mathbb{P}(\underline{X} = x_i, Y = y)}{\mathbb{P}(\underline{Y} = y)}$

$= \sum_{x_i \leq x} \left(\frac{f_{XY}(x,y)}{f_Y(y)} \right) \rightsquigarrow$ PMF

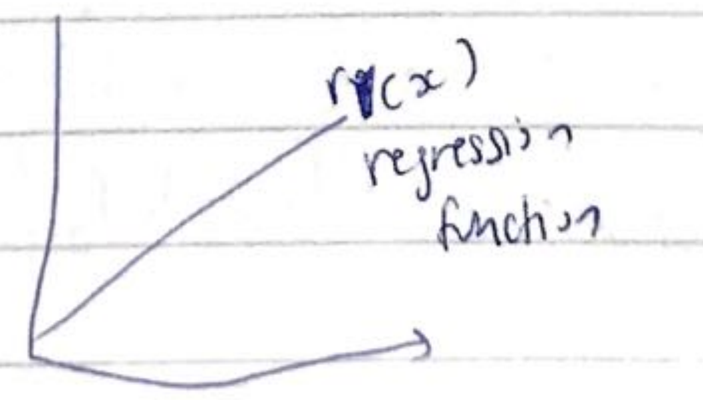
Independence $\Leftrightarrow f_{X|Y} = f_{\underline{X}}(x) \quad f_Y(y) \neq 0$

! head -n 5 data/CORIS.csv

Sequences: We say that we have a sequence of Random Variable $X_1, X_2 \ldots X_n$ if for a fixed m $\underline{X} = (\underline{X}_1, \ldots \underline{X}_m)$ is a RV

* We say sequence is independent if $F_{\underline{X}_1, X_2, x \ldots x_m}(x_1, x_2 \ldots x_m) = F_{x_1}(x) \cdot F_{x_2}(x) \ldots F_{x_m}(x)$ for any m
* We say a sequence is identically distributed if $F_{\underline{X}_i} = F_{\underline{X}_j}$ for all i, j
* We say a sequence is iid if both independent and identically distributed

In regression we want $\mathbb{E}[\underline{Y} | \underline{X} = x] = r(x)$



17/09/24 Intro to Data Science : Lecture 4

Learning from data? What is the avg weight of the population in Sweden

1) Experiment is to select a random person $\Omega = \{w_1, w_2 \dots w_N\}$

weighting person $w \to x(w)$, $X(w) : \Omega \to \mathbb{R}^+$

what do we know about X? X: continuous , $X \geq 0$ X: bounded $\leq M$

Assumptions about reality

2) Choose to select n people $X_1 \dots X_n$ (if there is replacement then i.i.d (identically & independently distributed))

$X_1 \dots X_n$ IID

knowing the weight of 1 person doesn't give you an indication of other person

$\frac{1}{n} \sum_{i=1}^{n} X_i$ LTRF idea

$\mathbb{E}[X_1]$

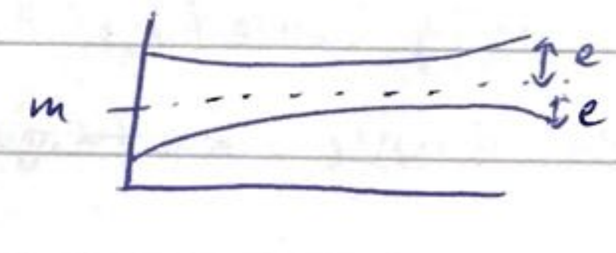
phenomena of concentration (when you do an experiment a lot) ex: coin toss

We started with LTRF we defined probability now we are back at LTRF. We gained that we can calculate things.

Fix ε (error) > 0 and define $m = \mathbb{E}[X_i]$, $\bar{X}_n := \frac{1}{n} \sum_{i=1}^{n} X_i$

$\mathbb{P}(|\bar{X}_n - m| > \varepsilon) \to 0$

as $n \to \infty$



"convergence in probability"

* Remark: Sometimes we can calculate exactly $\mathbb{P}(|\bar{X}_n - m| > \varepsilon)$ $X_i \overset{iid}{\sim} Bernoulli(p)$ then $n \cdot \bar{X}_n \sim Binomial(n, p)$

Probability inequalities: Chebychev: Let $X \in \mathbb{R}$ be a RV and let $\varepsilon > 0$ then $\mathbb{P}(|X - \mathbb{E}[X]| > \varepsilon) \leq \frac{\mathbb{V}[X]}{\varepsilon^2}$ assuming $\mathbb{E}[X]$ exists and $\mathbb{V}[X]$ exists

Small variance $\Rightarrow$ small probability

Large error $\Rightarrow$ —— " ——

$\mathbb{E}\left[\frac{1}{n} \sum_{i=1}^{n} X_i\right] = \frac{1}{n} \sum_{i=1}^{n} \underbrace{\mathbb{E}[X_i]}_{m} = m$

take any one bc they're identically distributed

$\mathbb{P}(|\bar{X}_n - m| > \varepsilon) \leq \frac{\mathbb{V}[\bar{X}_n]}{\varepsilon^2}$ Independence gives $\mathbb{V}[\bar{X}_n] = \frac{\mathbb{V}[X_i]}{n}$

$\mathbb{V}[\bar{X}_n] = \mathbb{E}[(\bar{X}_n - \underbrace{\mathbb{E}[\bar{X}_n]}_{m})^2] = \mathbb{E}\left[\left(\frac{1}{n} \cdot \sum_{i=1}^{n} \underbrace{(X_i - m)}_{Y_i}\right)^2\right] = \frac{1}{n^2} \mathbb{E}\left[\left(\sum_{i=1}^{n} Y_i\right)^2\right]$

$= \frac{1}{n^2} \mathbb{E}\left[\sum_{i,j} Y_i Y_j\right] \overset{\text{indep.}}{=} \frac{1}{n^2} \left(\sum_{i \neq j} \underbrace{\mathbb{E}[Y_i]}_{0} \cdot \underbrace{\mathbb{E}[Y_j]}_{0} + \sum_{i=1}^{n} \mathbb{E}[Y_i^2]\right) = \frac{\mathbb{V}[X_1]}{n}$

indep.

$\mathbb{E}[Y_i^2]$ → variance

$\mathbb{V}[X_1]$

Variance of a sum $\overset{\text{independence}}{=}$ sum of the variances

$$\mathbb{P}(|\bar{X}_n - m| > \varepsilon) \le \frac{\mathbb{V}[\bar{X}_n]}{\varepsilon^2} = \frac{\mathbb{V}[X_1]}{n\cdot\varepsilon^2} \quad \text{as } n\to\infty \quad n\cdot\varepsilon^2 \overset{\text{goes to}}{\rightsquigarrow} \infty$$

Fix a number $a \in (0,1)$ choose ε such that $\frac{\mathbb{V}[X]}{n\cdot\varepsilon^2} = a$

$$\sqrt{\frac{\mathbb{V}[X]}{a}} = \varepsilon$$

$$\mathbb{P}\left(|\bar{X}_n - m| > \overbrace{\frac{\sqrt{\mathbb{V}[X]}}{\sqrt{n}\cdot\sqrt{a}}}^{\text{s.d.}}\right) \le a$$

(look at the proof ← Hoeffdings inequality (will be used for the entire course - we need to learn it).

Let $X_1 \dots X_n$ be IID RVs such that $a \le x_i \le b$ and $\varepsilon > 0$ then

$$\mathbb{P}(|\bar{X}_n - m| > \varepsilon) \le 2e^{\frac{-2n\cdot\varepsilon^2}{(b-a)^2}}$$

the bigger a then all the ε becomes bigger

First thing: Fix a solve $2e^{-\frac{2n\varepsilon}{(b-a)^2}} = a$

$$\varepsilon = \sqrt{\frac{(b-a)^2}{2n}\ln\left(\frac{2}{a}\right)}$$ then form an interval $I = (\bar{X}_n - \varepsilon, \bar{X}_n + \varepsilon)$ ← empirical mean

$\mathbb{P}(m \text{ is in } I) \ge 1-a$ Let us make a decision based on data $(X_1 \dots X_n)$

Decision: I decided that the true mean is inside my interval I. (you are gonna be correct $1-a$ % of times) ← probability

I: is for before you've seen the data then if you see the data is no longer a prob.

↳ It is the fastest you learn as ~~data~~ ~~esp~~ the ~~data~~ errors get smaller when sample size increases eg. weight of a person (there is a max. and we know it is non-negative., coin toss)

Binomial is exact but it is difficult to calculate ~~of~~ for $n \ge 20$. but you can use Hoeffding to approximate.

If you know the variance you can do better: Bennett

$2\exp(\quad)$ ... goes to zero when sample $\to \infty$

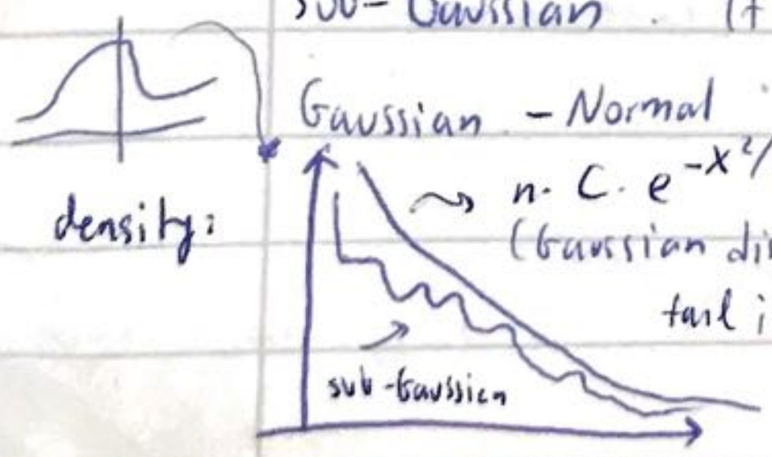
δ

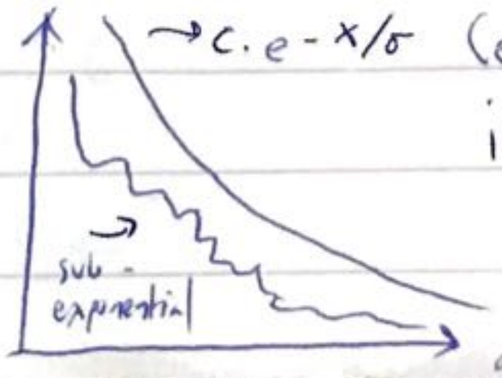
25/09 Lecture 5: Hoeffding's Inequality : true if $a \le x \le b$

$\mathbb{E}[e^{-s(x-\mathbb{E}(x))}] \le ? \; e^{s^2/\sigma^2}$ (assume this is true).

if we assume that is a valid assumption and as soon is true it is called sub-Gaussian. (It's called sub-Gaussian because the tails are smaller than)

density:





If you satisfy Gaussian condition do you automatically satisfy exponential? yes because you are taking the sqr → smaller always

Ex. 1: Normal is sub-Gaussian

Ex. 2: $a \le x \le b$ also sub-Gaussian

Ex. 3: $X \sim N(0,1) \to X^2$ is sub-exponential (χ^2 ~ Chi-squared ↵ the tail gets an exponential distribution

if you want to estimate the variance not only the mean, then you are using squared values

$\hat{\mathbb{V}}_n := \sum_{i=1}^{n} (X_i - \mathbb{E}[X_i])^2$ the rate with respect to epsilon for sub exponential is lower.

for sub-Exponential: $\mathbb{P}(|\bar{X_n} - \mathbb{E}[\bar{X_n}]| > \varepsilon) \le \max\left(2e^{-\frac{\varepsilon^2 n}{c_1}}, 2e^{-\frac{\varepsilon n}{c_2}}\right)$

as the error increases (for large ε) you get fatter tail ↗ that is why there isn't ε^2).

Def: A statistical model is a set of distributions (densities/ regression functions)
↳ Wasserman.

$\mathcal{F} := \{F\}$ -set of distributions

These are our assumptions: We are implicitly assuming that our data comes from $F^* \in \mathcal{F}$

Ex. 1: $\mathcal{F} := \{\underbrace{\frac{1}{\sigma\sqrt{2\pi}} e^{-\left(\frac{(x-\mu)}{\sigma}\right)^2}}_{\phi_{\mu,\sigma}}, \mu \in \mathbb{R}, \sigma > 0\}$ it is not a big set as I can describe it using only 2 numbers (mean & variance). Then I will have a good approximation. This is called a parametric model

Non-parametric ⇔ not parametric

Ex. $\mathcal{F} := \{F_X : F_X(a) = 0, F_X(b) = 1\}$ all functions from a to b - you need huge amount of parameters to describe it. you take a RV between a and b

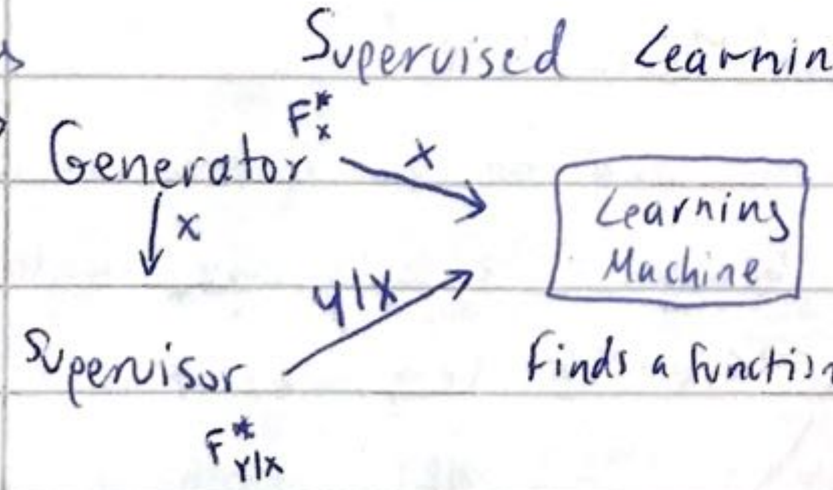
What can we learn?

* $\mathbb{E}[X] = \int x \, dF^*(x)$ - sum or integral based on if it is a discrete/continuous variable.
* learn $(F^*)' = f^*$ density - "Histogram" example of learning density

learn $F^*(x)$: $\hat{F}(x) = \frac{1}{n} \sum_{x_i \in x} 1$ (how many sample is below this)

↳ Empirical Distribution function.

Supervised Learning:



Def: Consider a function $L: \mathbb{R}^2 \to \mathbb{R}$ (loss function) Let $(X,Y) \sim F^*_{XY}$, then Risk of a

function $g: \underline{X} \to \underline{Y}$ is $R(g) := \mathbb{E}[L(g(X), Y)]$ ~ it measures how close they are is ; it is a good guess (L) and g: expected error

Ex. 1: $L(a,b) = (a-b)^2$ quadratic loss. (care about large errors ↘)

output is a number

$R(g) = \mathbb{E}[(g(x) - Y)^2]$ least square problem if $g(x) =$ line (error very big when diff is big because ^2)

Ex. 2: $L(a,b) = |a-b|$ absolut loss

$R(g) = \mathbb{E}[|g(x) - Y|]$ → measures ↓ everything the same (proportional)

EX. 3: if the output $\underline{Y} = \{0,1\}$:

$$L(a,b) = \begin{cases} 1, & \text{if } a \neq b \\ 0, & \text{otherwise} \end{cases}$$

$R(g) = \mathbb{E}[L(g(x), Y)] = \mathbb{P}(g(x) \neq Y)$ ← prob of making a mistake

↳ Pattern recognition

$1 - R(g) =$ "accuracy" It is called 0-1 loss

Objective: The learning machine tries to minise risk by searching a $g^* \in \mathcal{M}$ [~ model space] that solves $R(g) = \inf_{g \in \mathcal{M}} R(a)$ → optimisation problem

Ex. 1: $\mathcal{M} = \{kx + m,\ k \in \mathbb{R},\ m \in \mathbb{R}\}$ → linear function } Least squares problem

$L(a,b) = (a-b)^2$ ↳ Ordinary Linear Regression

The LM doesn't have access to $R(g)$ because it doesn't know F^*. To solve this we use data (examples) which is a bunch of pairs $((x_1,Y_1), (x_2,Y_2) \ldots (x_n, Y_n))$.

Instead solves $\hat{R}(\hat{g}^*) = \inf_{g \in \mathcal{M}} \hat{R}(g)$

where $\hat{R}_n(g) = \frac{1}{n} \cdot \sum_{i=1}^{n} L(g(x_i), Y_i)$

Once you have found $\hat{g}^*$ (if you do the same thing again with new data then $\hat{g}^*$ changes)

↳ $\hat{g}^*$ is random (depends on size of $\mathcal{M}$ and # of data points)

How do we know its any good? Take new data (testing data)

$((x_{n+1}, Y_{n+1}), \ldots, (x_{n+m}, Y_{n+m}))$ Then you calculate:

$\hat{R}_m(\hat{g}^*) = \frac{1}{m} \sum_{j=1}^{m} L(\hat{g}(x_{n+j}), Y_{n+j})$ ← testing error

30/09 $\mathcal{F} = \{$ all admisable distributions $\mathcal{F}$: Statistical Model

$\mathcal{M} = \{$ all guessing functions g for the LM$\}$

Loss function $L(a,b)$

Data: $((x_1,Y_1), \ldots (x_n,Y_n)) =: D_{Tr}$ "Training Data"

↳ IID $- F_{XY}^*$ (unknown to us) (if there is sorting in the data there is longer IID → need scheduling)

$((x_{n+1}, Y_{n+1}), \ldots, (x_{n+m}, Y_{n+m}) =: D_{TE}$ "Testing Data"

The LM minimises the "empirical risk" (because they are actual data) $\hat{g} = a + \min_{g \in \mathcal{M}} \overbrace{\frac{1}{n} \sum_{i=1}^{n} L(g(x_i), Y_i)}^{\text{empirical risk}}$

We are thinking that the learning Machine is approximately minimising the true risk

$R(g) = \mathbb{E}[L(g(X), Y)]$

↑ Risk

what is the approximation of this? (of the true risk)

we test $\hat{g}$ on D_{Te} by calculating $\frac{1}{m}\sum_{i=1}^{m} L(\hat{g}(x_{n+i}, y_{n+i}) \approx \mathbb{E}[R(\hat{g})|$ ↩

$\hat{g}$: is not random anymore → you guessed it $= \mathbb{E}[R(\hat{g})|D_{Tr}]$

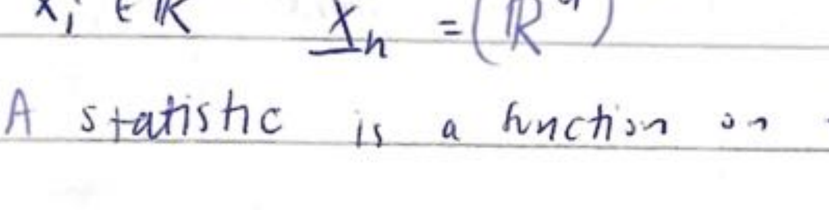
actual function — based on the training data and you want to test it on the testing data

trained model (Posterior error)

This is an approximation of the true risk of $\hat{g}$ given $(\hat{g}|D_{Tr})$

$\hat{g}$: is fixed since you know the training data.

Ex.



given data

your prediction based on the given data

$L(a,b) = (a-b)^2$ training error is always very small ~ usually 0

but when you applied it in testing data increases.

Def: Given a statistical model S the data space $\underline{\mathbb{X}}_n$ is where $((x_1,y_1)\ldots(x_n,y_n)) \overset{IID}{\sim} F_X^*$ like $F_X^* \in S$ → true distribution but we don't know it.

* x_i = includes both x and

Ex. $x_i \in \mathbb{R}$ $\underline{\mathbb{X}}_n = \mathbb{R}^n$

$x_i \in \mathbb{R}^d$ $\underline{\mathbb{X}}_n = (\mathbb{R}^d)^n$

Def: A statistic is a function on the data space $\underline{\mathbb{X}}_n$: $T: \underline{\mathbb{X}}_n \to \mathbb{T}$ ↵ the possible values of T.

Ex: Supervised Learning $\hat{g} = \underset{g\in\mathcal{M}}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n} L(g(x_i), y_i))$ $\hat{g}$: is statistic can be calculated based on data.

data space $\underline{\mathbb{X}}_n = (\mathbb{R}^2)^n$ $\Rightarrow \mathbb{T}$

↳ all straight lines that $\hat{g} \in \mathcal{M}$ are in the model space.

Ex: Testing error $\frac{1}{m}\sum_{i=1}^{m} L(\hat{g}(x_{n+i}), y_{n+i})$. this is statistic as we calculated it using data

$=: T(D_{Te})$ a statistic $\underline{\mathbb{X}}_m$ $\mathbb{T}$: target space $= \mathbb{R}_+$

Properties: "An estimator" of ϑ is a statistic T that approximates ϑ"

bias $(T) = \mathbb{E}[T] - \vartheta$ if >0 positively-biased, <0 negatively-biased,

↑ "T estimates ϑ" if $=0$ ⇒ unbiased

Ex: $T(x_1 \ldots x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i$ $\vartheta = \mathbb{E}[x_i]$ bias $(T) = 0$

Ex 2: $T(x_1 \ldots x_n) = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x}_n)^2$ $\vartheta = \mathrm{Var}[x_1]$

bias $(T) = -\frac{1}{n}$ if you have a lot of data, the bias is very small (bias goes to zero if you collect more data). If bias $(T) \to 0$ as $n \to +\infty$ we say T is asymptotically unbiased.

Ex.3 $T(x_1, \ldots x_n) = x_1$ $\vartheta = \mathbb{E}[x_1]$ what is the bias? Bias $(T) = 0$

Def: standard error $se(T) = \sqrt{\mathbb{V}[T]}$

Ex.1 $se(T) = \frac{\sqrt{\mathbb{V}[x_1]}}{\sqrt{n}}$ ↳ depends on n.

Ex.3 $se(T) = \sqrt{\mathbb{V}[x_1]}$ ← it doesn't matter how much data you get ⇒ will be the same

$\mathbb{E}[R(\hat{g})]$

Mean squared error (MSE(T)) $= \mathbb{E}[(T-\theta)^2] = \mathbb{E}[(T-\mathbb{E}[T])^2 + (\mathbb{E}[T]-\theta)^2$

(T: calculated, θ: true, how far they are)

$= (se(T))^2 + (bias(T))^2$

Bias - Variance decomposition

you can split the error into uncertainty (how far you are from the true value - if you push one to zero the other goes higher, the smaller the variance the higher the variance will be

Bias can go to zero as you can find the true but $\Rightarrow$ variance will go up

4/10 Lab Estimation : $X_1, \dots, X_n$

- estimate parameters

for example : $X_1, \dots X_n \sim N(m, \sigma^2)$ → use parametric model

Point estimator : $\vartheta = g(X_1, \dots X_n)$

$T(x)$: estimator

Properties : 1) bias $E_\vartheta(T(x)) - \vartheta$ if $E(T(x)) = \vartheta \Rightarrow$ unbiased

2) $se = \sqrt{var(T(x))}$

MSE : Mean Squared Error :

$Bias^2(\hat{\vartheta}) + Var(\hat{\vartheta})$

Example : ① $x \sim Po(\lambda)$ $\vartheta = [P(x=0)]^2 = e^{-2\lambda}$

Choose unbiased estimator $(T(x))$

$E(T(x)) = \vartheta = e^{-2\lambda}$

$\sum_{x=0}^{\infty} \frac{e^{-\lambda}\lambda^x}{x!} T(x) = e^{-2\lambda}$

$T(x) = -1^x$

② AOS : Chapter 6. 1) $X_1 \dots X_n \sim Poi(\lambda)$

$\hat{\lambda} = \frac{\sum x_i}{n}$ find bias, SE, MSE 1) Expectation : $E\hat{\lambda} = E\left[\frac{\sum X_i}{n}\right] = \sum \frac{EX_i}{n} = \frac{n \cdot EX_i}{n} = \lambda$

sample mean is always unbiased bias = 0.

$MSE = Var(\hat{\vartheta}) = Var\left(\frac{\sum x_i}{n}\right) = \frac{1}{n^2} \sum (Var X_i) = \frac{n\lambda}{n^2} = \frac{\lambda}{n}$

$\sum \frac{EX_i}{n} = \frac{n EX_i}{n} = \lambda \iff SE = \frac{\sqrt{\lambda}}{n}$

③ $X_1 \dots X_n \sim Uni(0, \vartheta)$ $\hat{\vartheta} = 2\bar{X}_n$ compute SE, bias, MSE

$\bar{X}_n = \frac{\sum x_i}{n}$ $E\hat{\theta} = \frac{2\sum X_i}{n} = \frac{2\sum(EX)}{n} = \frac{2\frac{\vartheta}{2}}{n} n = \vartheta \rightsquigarrow$ unbiased estimator bias = 0

$MSE = Var(\hat{\theta}) = Var\left(\frac{2\sum x_n}{n}\right) = \frac{4}{n^2} \sum Var(x) = \frac{4}{n^2} \cdot \frac{\vartheta^2}{2} \frac{\vartheta^2}{3n^2} \Rightarrow SE = \frac{\sqrt{\vartheta^2}}{3n^2} = \frac{\vartheta}{n\sqrt{3}}$

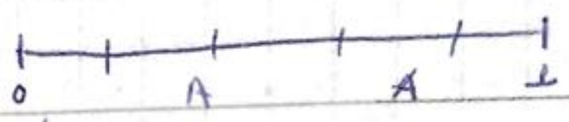
L7: Estimation Risk : std error if ↓ ⇒ better

L8: Generating RVs Markov Chains: Pseudo Random Number Generation (PRNG) → behaves as it was random

Minimum requirements (Naive): to ~~Naive~~ Uniform ([0,1])

Create $V_0, V_1, \dots, U_n$ for any set $A \subset [0,1]$ $\quad \frac{N_n(A)}{n} \to \int_A dx$

$N_n(A)$ is the number of $U_i \in A$ $\quad 0 \le i \le n-1$



$X \sim \text{unif}([0,1])$ $\quad \mathbb{P}(X \in A) = \int_A dx.$

To hard to do unif([0,1]) so let's do uniform $\mathcal{M} = \{0, 1, \dots M-1\}$

Def. A sequence $V_0, V_1 \dots$ is pseudorandom if for any number $a \in \mathcal{M}$: $\frac{N_n(a)}{n} \to \frac{1}{M}$ (frequency)

(every number should appear an equal amount of times).

eg. 0,1,0,1... is pseudo random on {0,1}

1,1,0,0,1,1,0,0 : — // —

Def: A congrevential generator with parameters (a,b,M) : $D(x) = (ax+b) \cdot \text{mod } M$

start at some u_0 - random seed $U_i = D(U_{i-1})$: $V_1 = D(U_0)$, $V_2 = D(U_1)$

Ex.1: $a=2$, $b=1$, $M=3$ $\quad U_0 = 1$, $V_1 = (2 \cdot 1 + 1) \text{ mod } 3 = 0$, $V_2 = (2 \cdot 0 + 1) \text{ mod } 3 = 1$

$V_0 = 2$: $V_1 = (2 \cdot 2 + 1) \text{ mod } 3 = 2$

Ex. 2: $a=3$ $\quad b=1$ $\quad M=7$ $\quad V_0 = 1$: $V_1 = (3 \cdot 1 + 1) \text{ mod } 7 = 4$

$V_1 = 4$: $V_2 = (3 \cdot 4 + 1) \text{ mod } 7 = 6$

$V_2 = 6$: $V_3 = (3 \cdot 6 + 1) \text{ mod } 7 = 5$

$V_3 = 5$: $V_4 = (3 \cdot 5 + 1) \text{ mod } 7 = 2$

$V_5 = 0$: $V_6 = (3 \cdot 0 + 1) \text{ mod } 7 = 1$.

$V_6 = 1$

$V_7 = 3$: $V_1 = (3 \cdot 3 + 1) \text{ mod } 7 = 3$

Def: The period of D starting at V_0 is the smallest number T_0 such that $U_{i+T_0} = U_i$, for all i

Ex.1: Period of D starting in 1 is 2 and if you start from ~~2~~ $V_0 = 2$ is 1

EX.2: $V_0 = 1$, period 6, $V_0 = 3$ period 1.

Full period on $\{0, 1, \dots M-1\}$ is M

Start at $V_0 = 0$ then after M steps we get 0 $\quad \frac{N_n(0)}{n}$ or $\frac{N_{k \cdot M}(0)}{k \cdot M} = \frac{k}{kM} = \frac{1}{M}$ $\quad \frac{N_n(0)}{n} \to \frac{1}{M}$

is pseudo-random sequence

Choose a very large M and assume you have a full period $V_0, V_1 \dots V_n$. Fix K such that K divides M and construct: $V_i = [V_i \cdot \frac{K}{M}]$ → floor you take just the integer and leave behind the decimals (same as using int in Python).

V_i is pseudorandom on $\{0, 1, \dots, K-1\}$

The period of V_i is M (good candidate for randomness).

To create uniform [0,1] we take $V_i = \frac{U_i}{M}$

Sampling. say we want A_0 sample from F, calculate F^{-1}

Generate Uniform([0,1]) and consider $X = F^{-1}(V)$ then $X \sim F$

Lecture: Introduction to Data Science

Regression: Consider a pair $(X,Y) \sim F^{*}_{XY}$ — $R(f)$ → risk = expected loss

$L(a,b) = (a-b)^2$

Let's say we have a guess f: our guess

$R(f) = \mathbb{E}[L(\bar{Y}, f(\bar{X}))] = \mathbb{E}[(\bar{Y} - f(\bar{X}))^2]$

Define $r(x) = \mathbb{E}[\bar{Y} \mid \bar{X} = x]$

$R(f) = \mathbb{E}[((Y - r(\bar{X})) + (r(\bar{X}) - f(\bar{X})))^2]$

$\mathbb{E}[(Y - r(X))^2] + \mathbb{E}[(r(\bar{X}) - f(\bar{X}))^2] + 2\mathbb{E}[(\bar{Y} - r(\bar{X}))(r(\bar{X}) - f(\bar{X}))]$ — $f(\bar{X})$

↳ Tower Rule

$= \mathbb{E}[\mathbb{E}(Y - r(\bar{X}))(r(\bar{X}) - f(\bar{X}) \mid X]]$

$= \mathbb{E}[\mathbb{E}[Y \mid X] - r(\bar{X}))(r(\bar{X}) - f(\bar{X})] = 0$

$\mathbb{E}[Y \mid X] = r(\bar{X})$

Conclusion: $R(f) = \mathbb{E}[(Y - r(X))^2]$ ① $+ E[(r(\bar{X}) - f(\bar{X}))^2]$ ②

If you change f: nothing happens to the first term. If you choose $r = f$: the second term will be zero ⇒ will be the best that you can do.

$R(r)$ = is the smallest Risk.

you cannot do anything about it

① : calculates the measurement noise (how fat it is) → "noise variance"

② : calculates the bias^2 (you can do something about it)

r -> is called a regression function $\bar{Y}$

second problem we need to solve

Find "f" Now we have $E[(\bar{Y} - r(\bar{X}))^2] = 0$ this can happen when you know the real label.

good for image classification (eg. NN) and text generation _sequence of words

(usually much easier)

Pattern Recognition. $Y \in \{0,1\}$ $L(a,b) = \begin{cases} 1 & \text{if you make a mistake } a \neq b \\ 0 & \text{otherwise (no mistake)} \end{cases}$

$r(x) = \mathbb{E}[\bar{Y} \mid X = x] = \mathbb{P}(Y = 1 \mid X = x)$

↳ class ↳ images → true probability that class is in the image based on the image.

$h^*(x) = \begin{cases} 1 & \text{if } r(x) > 1/2 \\ 0 & \text{if } r(x) \leq 1/2 \end{cases}$ Bayes Classifying Rule

Property: $R(h^*) \leq R(h)$ for any guessing function h.

Maximum Likelihood Method:

Let $S = \{P_a(x) : a \in \mathbb{R}^d\}$ densities; it can be a Gaussian or any other distribution

a: parametric model

Loss: $L(x,a) = \ln P_a(x)$ log-loss

$X \sim P_{a^*} \in S$ $R(a) = \mathbb{E}[L(X,a)] = \mathbb{E}[-\ln P_a(x)] = \int \ln P_a(x) \cdot P_{a^*}(x)\,dx$

we cannot calculate the risk but we can calculate the empirical risk:

Empirical Risk : $X_1, \dots, X_n \overset{IID}{\sim} P_{a^*}$

$$\hat{R}(a) = \frac{1}{n} \sum_{i=1}^{n} \underbrace{-\ln P_a(\bar{x}_i)}_{\text{neg-log-likelihood}}$$

This is also called entropy - we are measuring $R(a) - R(a^*)$

Regression: If we assume $f_{XY} = f_{Y|X} \cdot f_X \Leftrightarrow f_{XY} = \ln f_{Y|X} + \ln f_X$

If we think that $f_{Y|X} = P_{a^*}(y|x)$

Statistical Model is assumed to be $\mathcal{J} := \{ P_a(x,y) : P_a(x) = P_0(x), P_a(y|x) = f_a(x,y) \}$

usually we do an assumption that this is poicson distribution and plug it as below.

Ex. Linear regression

$$P_a(y|x) = \frac{1}{\sqrt{2\pi \cdot \sigma^2}} \cdot e^{-\frac{(y-(ax+b))^2}{\sigma^2}}$$

$a = (a, b, \sigma)$

$$-\ln P_a(y|x) = \ln \sqrt{2\pi} + \ln|\sigma| + \frac{(y-(ax+b))^2}{\sigma^2}$$

→ just a number; doesn't depend on data (a); ~ only these part depends on the data.

$$\hat{R}(a) = \ln|\sigma| + \frac{1}{n} \sum_{i=1}^{n} \frac{(\bar{y}_i - (a \cdot \bar{x}_i + b))^2}{\sigma^2} + \ln \sqrt{2\pi}$$

always same result when using Gaussian

Logistic Regression : $P_a(y|x) = \text{Bernoulli}(p)$

$$p = \frac{1}{1+e^{-(ax+b)}} = G(ax+b)$$

$$P_a(y|x) = p^y (1-p)^{1-y}$$

$$G(x) = \frac{1}{1+e^{-x}} \rightsquigarrow \text{logistic function}$$

$$\ln P_a(y|x) = y \ln p + (1-y) \cdot \ln(1-p)$$

when you guess the p then you plug it to the pattern recognition part and you get 1 or 0 based on the result.

→ prob. that $X \le x = P(\bar{X} \le x) = P(F^{-1}(u) \le x) = P(U \le F(x)) = F(x)$

because $F_u(u) = u$

$x) = 1 - e^{-\lambda x} \Rightarrow y = 1 - e^{-\lambda x} \Rightarrow e^{-\lambda x} = 1 - y \Leftrightarrow x = \frac{-1}{\lambda} \ln(1-y) = F^{-1}(y)$

Inversion sampling (works for discrete too - but use a different def for discr

Accept - Reject Method : Input target density f

sampling density g

$f(x) \le Ng(x)$ for all x.

Repeat: Draw $\bar{X} \sim g$ and calculate $r(\bar{X}) = \frac{f(\bar{X})}{Ng(\bar{X})}$, Draw a Uniform ([0, 1]) and accept X if $U \le r(\bar{X})$ → accept or reject.

$X \sim G$ $U \sim unif\ [0,1]$

$I = \begin{cases} 1, & \text{if } U \le r(x) \\ 0, & \text{otherwise} \end{cases}$ if we accept then $I = 1$ distribution that we are looking for : $P(X \le x \mid I = 1)$

Bayes then $f_{\bar{X}/I}(x/i) = \dfrac{f_{I/\bar{X}}(i/x) \cdot f_{\bar{X}}(x)}{f_I(i)}$

$f_{I/\bar{X}}(i/x) = P(U \le r(\bar{X}) / X = x) = r(x)$

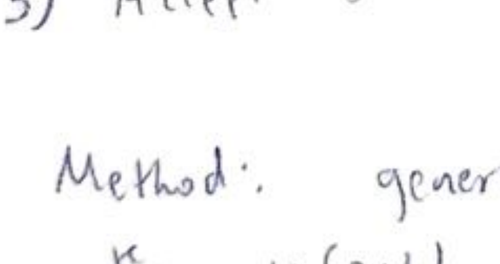
$f_{\bar{X}}(x) = g(x)$

$f_I(i) =$ using the Law of total probability $= \int f_{I/\bar{X}}(i/x) \cdot f_{\bar{X}}(x)\, dx = \int \underbrace{r(x)}_{\frac{f(x)}{N \cdot g(x)}} g(x)\, dx$

$= \frac{1}{N}$

$f_{\bar{X}/I}(x/i) = \dfrac{r(x) \cdot g(x)}{1/N} = f(x)$ That is $P(\bar{X} \le x / I = i) = F(x)$

if density is very thin and sampling distribution is uniform:



Lab session: Random number generation:

1) inverse transform method: $u \sim u(0,1)$, $F^{-1}(u) = F \sim$ target distribution

2) Box - Muller algorithm : $u_1, u_2 \sim u(0,1)$ - iid $z_0 = \sqrt{-2\ln(u_1)} \cdot \cos(2\pi \cdot u_2)$

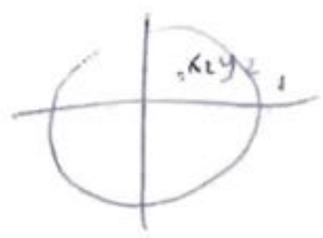
$z_1 = \sqrt{-2\ln(u_1)} \cdot \sin(2\pi \cdot u_2)$

$z_0, z_1 \sim N(0,1)$

$z = z_i \cdot \delta^2 + m$ $z \sim N(m, \delta^2)$

3) Accept - reject : target distribution $f(x)$ & sampling distribution $g(x)$

$f(x) \le c\,g(x)$



$x_0^2 + y^2 \le 1$ (x_1, y_1)

Method: generate X from $g(x)$ $r(x) = \dfrac{f(x)}{c\,g(x)}$

generate $u \sim u(0,1)$ if $u \le r(x)$ $x^n = x^a$

else:

reject generation until $u \le r(x)$

Markov's Chains:

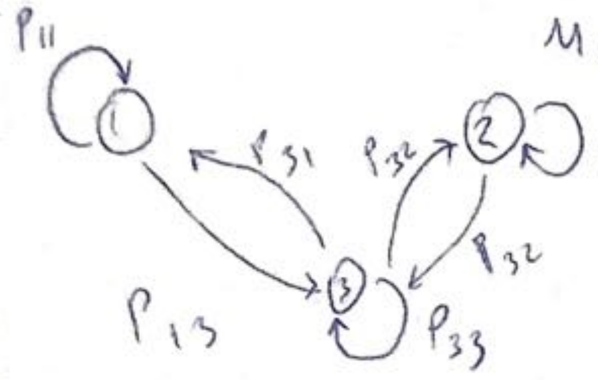
→ specialise of stochastic process.

Finite MC . $X_n \sim$ stochastic probability

X_n is a MC if $P(X_{t+1} \leq x \mid f_t) = P(X_{t+1} \mid X_t)$

MC is defined with a transition matrix.

$$M_n = \{\hat{M}_{ij}\}$$



Key Properties: $\sum_j P_{ij} = 1$ for any i

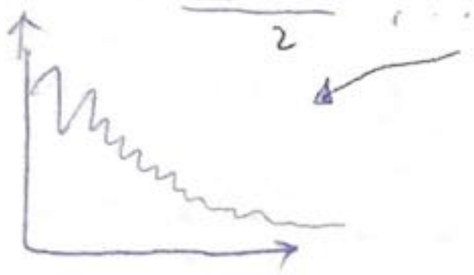
6/11/24 · Lecture 9: Markov's Chain

Def: A stohastic process is indexed set of random variables $\{X_i\}_{i \in \mathbb{I}}$ → index can be a time or any number

If $\mathbb{I} = \{1, 2, 3 \dots\}$ we say its discrete

Ex: $\overline{X_1}, \overline{X_2} \dots \overline{X_n} \overset{IID}{\sim} F$ (having independent samples) → you cannot shuffle it ⇝ you will destroy it.

Ex. $\overline{X_1}$, $\frac{\overline{X_1} + \overline{X_2}}{2}$, ... , $\frac{1}{n} \cdot \sum_{i=1} \overline{X_i}$ all connected as the next value depends on the previous.

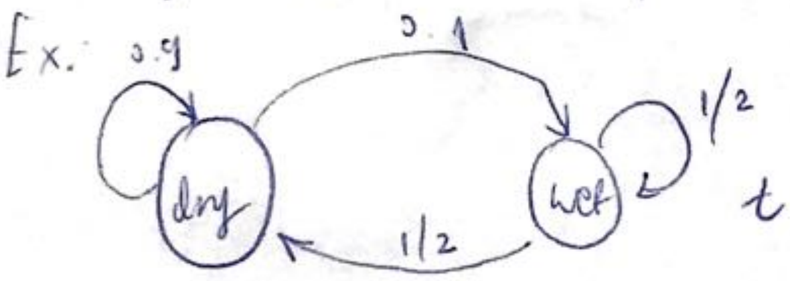


Def: Markov process: $\{X_i\}_{i \in \mathbb{N}}$, $\overline{X_i} \in \overline{\mathbb{X}}$ → fix this to be the same in a Markov process

$\overline{\mathbb{X}} = \{S_1, \dots, S_n\}$

$\mathbb{P}(X_t = y \mid X_1, X_2, \dots, X_{t-1}) = \mathbb{P}(X_t = y \mid X_{t-1})$ ⇝ "Markov Property" (holds a memory of one)

"Memory only of one step"

Ex. 

Transition diagram:

$$P^{t-1} = \begin{bmatrix} \mathbb{P}(X_t = dry \mid X_{t-1} = dry) & \mathbb{P}(X_t = wet \mid X_{t-1} = dry) \\ \mathbb{P}(X_t = dry \mid X_{t-1} = wet) & \mathbb{P}(X_t = wet \mid X_{t-1} = wet) \end{bmatrix} = \begin{bmatrix} 0.9 & 0.5 \\ 0.1 & 0.5 \end{bmatrix}$$

Say that we know:

$P_{t-1} = [\mathbb{P}(X_{t-1} = dry), \mathbb{P}(X_{t-1} = wet)]$

$\mathbb{P}(X_t = dry)$ = if you on the dry, you could have been previously to wet or dry and using the law of total probability $= \mathbb{P}(X_t = dry \mid X_{t-1} = dry) \cdot P(X_t = dry) + P(X_t = dry \mid X_{t-1} = wet)\, \mathbb{P}(X_{t-1} = wet)$

$\mathbb{P}(X_t = wet) = P(X_t = wet \mid X_{t-1} = dry) \cdot P(X_{t-1} = dry) + P(X_t = wet \mid X_{t-1} = wet)\, \mathbb{P}(X_{t-1} = wet)$

$P_{t-1} = (1, 2)$ vector

$\overline{P} = (2, 2)$ $\quad P_{t-1} \cdot \overline{P} = (1, 2)$

$P_t = [\mathbb{P}(X_t = dry), \mathbb{P}(X_t = wet)] = P_t = P_{t-1} \cdot \underline{P}$

Recursive formula: $P_t = P_{t-1} \cdot \underline{P} = (P_{t-2} \cdot P) \cdot P = P_0 \cdot \underline{P}^t$ → matrix (multiplication) power

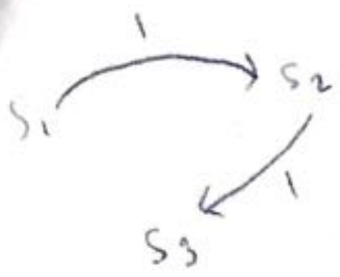
$P_0 = [0, 1]$ → we start on the wet state

$P_{10} = [0, 1] \cdot \underline{P}^{10}$

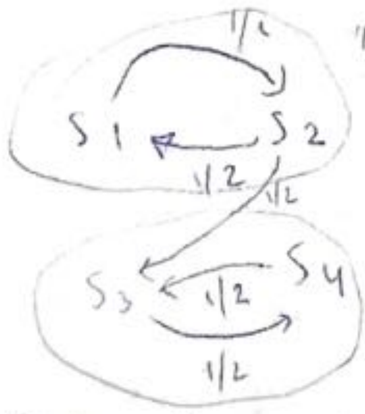
Homogeneous Markov Chains we can do something as the transition matrix is the same over time.

$\{S_1, \dots, S_n\}$ we say $S_i \to S_j$ "S_i communicates with S_j" if $\mathbb{P}(X_t = S_j \mid X_1 = S_i)$ for some t.

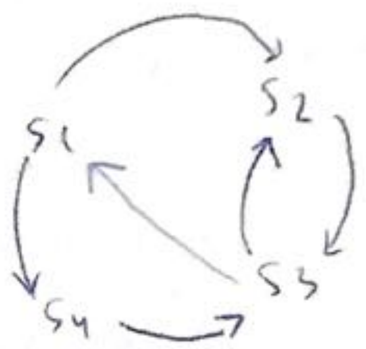
$S_i \to S_j$ and $S_j \to S_i$ intercommunicate



$S_1 \to S_2$
$S_2 \to S_3$
$S_1 \to S_3$

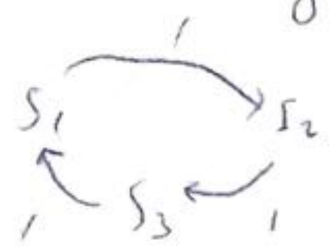


"two classes" check if there is a place you can get stuck



when everything communicates we say chain is irreducible.

you can stuck in a loop:

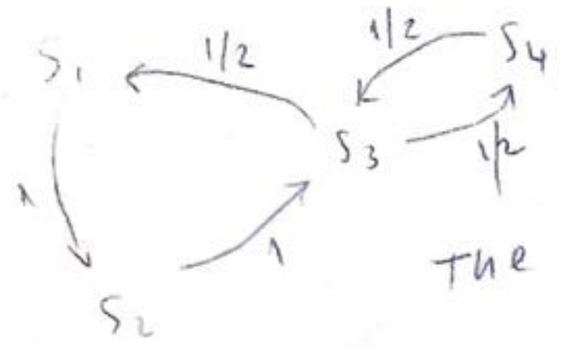


you will just going around. → avoid having

so mixing → problematic

Returning times:

$T(x) = \{t \in \mathbb{N} : (P^t)_{xx} > 0\}$



$T(S_1) = \{3, 5, 6, 7, \dots\}$

$T(S_3) = \{2, 3, 4, 5 \dots\}$.

The period of a state is $\gcd(T(x))$

we say that a chain is aperiodic if $\gcd(T(S_i)) = 1$ for all states

↳ a lot of mixing going on

Theorem: If irreducible and aperiodic then $P_0 P^t \xrightarrow[t\to\infty]{} \pi$ "mixing" ← unique

also $\pi = \pi \cdot P$ ~ stationary distribution if this holds

$\pi^T = (\pi \cdot P)^T = P^T \cdot \pi^T$ ~ eigenvector π^T of P^T

* remember to take P^T in the computer with an eigenvalue of 1.)

Peron-Frobenions theorem $\pi_i \geq 0$ for all i

because the ~~rows~~ columns sum to 1. (Columns if Transpose)

how fast it goes to the stationary distribution?

use eigenvectors as a basis

$$(P_0 - \pi) \cdot P = \left(\sum (p_0 - v_i) v_i - \pi\right) \cdot P = \sum_{i=1}^{N} (P_0 \cdot v_i) \lambda_i - \pi \quad \text{(since } \pi \cdot P = \pi\text{)} = \sum_{i=2}^{N} \lambda_i \cdot (P_0 \cdot v_i)$$

λ_2 is the largest one

$\|(P_0 - \pi) \cdot P\| \leq |\lambda_2| \|P_0 - \pi\|$ ~ converges exponentially fast

< 1.

"Exponential convergence to equilibrium"

Lab 8/11 Markov Chains:

1) Irreducibility: $MC = X = \{S_0, \dots, S_n\}$ is irreducible if:

$P(X_t = S_j \mid X_0 = S_i) > 0$

$P(X_t = S_i \mid X_0 = S_j) > 0$ $\quad S_i, S_j \in \mathbb{R}^2$

2) Aperiodicity: A period of the state - greatest common division of $T(x)$

$T(x) = \{t \in N, p^t(x, x) > 0\}$ State is aperiodic if a period is one.

Stationary Distribution: $MC: X = \{s_0, \dots, s_n\}$, $P = (P(x, y))$ has a stationary distribution $\pi = \{\pi(s_0), \pi \dots, \pi(s_n)\}$

if: 1) $\pi(x) > 0$ for any x, $\sum \pi(x) = 1$

2) $\pi \cdot P = \pi$ (fixed point equation)

Homogeneity: $P_{ij}^n = P_{ij}^t$, any $n, t \in \mathbb{N}$

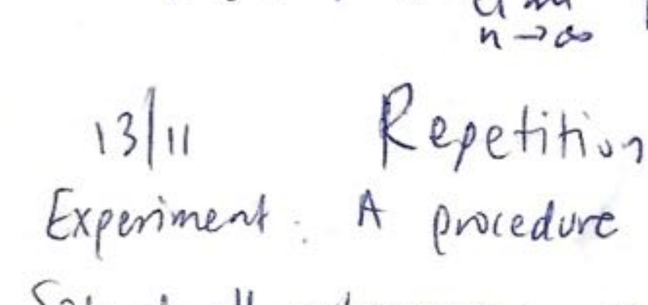
exercises from Lecture notes: 7.35 : some bacteria - splits in two
- stays the same
- dies
} states (0, 1, 2)

$X \sim$ offspring from each bacteria

$Z_{n+1} = \sum_{i=1}^{Z_n} X_{n,i}$ 1) is Z_{n+1} a MC? 2) simulate the process 3) Estimate the expectancy

it is a MC as it depends only on the previous state

Galton-Watson process



$X \sim$ # ofspring each node generates

$p = \{1/3, 1/3, 1/3\}$

Extinction prob $(\leq X) = 1$

Ex. $< 1 \Rightarrow \lim_{n\to\infty} P(Z_n = 0) = 1$

Ex. > 1 : $\lim_{n\to\infty} P(Z_n = 0) < 1$

13/11 Repetition Chapter 1-3

Experiment: A procedure that produces outcomes distinct.

Set of all outcomes: Ω "sample space" $\Omega = \{\omega\}$ → outcomes from the experiment

Trial: One round of an experiment: → n-fold trial (number of times you do the trial)

Events: subset of Ω , $A \subset \Omega$

A happens when $\omega \in A$

Ex. Coin toss $\Omega = \{0, 1\}$ $A = \{0\}$ = "we get a 0 in the trial"

Events A, B: $A \cup B$ = A or B , $A \cap B$ = "A and B"

$\mathbb{P}(\max\{X_1, X_2\} \leq x) = P(\max\{X_1 \leq x\} \cap \{X_2 \leq x\})$ if min then it is OR. and if $\geq <$

then is again OR

$\mathbb{P}\left(\left|\frac{1}{n}\sum X_i - \mathbb{E}[X_0]\right| > \varepsilon\right) = \mathbb{P}\left(\{\frac{1}{n}\sum X_j - E[X] > \varepsilon\} \cup \{E[X] - \frac{1}{n}\sum X_i > \varepsilon\}\right)$ → continue this & find the Hoeffding's inequality

Def: Let Ω be a set then we call $\mathcal{F}$ (sigma (σ)-algebra, observables). $\mathcal{F}$: set of events/subsets that are observable. To write it down you need 3 rules: 1) $\Omega \in \mathcal{F}$ (check if the trial happened),

2) If $A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$ → $\Omega \setminus A$ 3) if $A_1, A_2 \dots A_n \in \mathcal{F} \Rightarrow \bigcup_i A_i \in \mathcal{F}$ (at least one happens)

x. Markov's Chain: $X_1, X_2 \dots X_n \dots$

$\mathcal{F}_i = \{\omega : \underline{X_1} = x_1, \underline{X_2} = x_2 \dots, \underline{X_i} = x_i \mid$ for all $X_1 \in X, X_2 \in X \dots X_i \in X\}$

time t, that is $\mathcal{F}_t$

} observable
$\in F_t$

$\{X_{t+1} \le x\} \notin F_t$ not observable

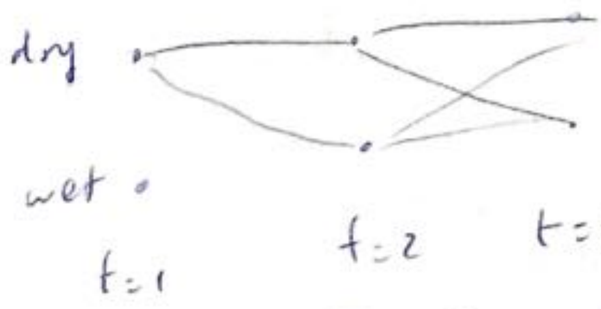
$F_1 \subset F_2 \subset F_3 \subset \dots$ (It is growing, more time the more information)

↳ Filtration

$F_1 = \{\{X_1 = 1\}, \{X_1 = 0\}, \{X_1 = 1, X_1 = 0\}\}$ ← union

$\{X_1 = 1\} = \{w_j : X_i = 1\} = \bigcup_{i=0} \{w_j : X_1 = 1, X_2 = 1\} \in F_1$

$\{X_1 = 1, X_2 = -1\} \in F_2$



Def: Let Ω be a set and given an F (observables) then: $\mathbb{P}: F \to [0,1]$.

1) $\mathbb{P}(\Omega) = 1$ 2) $A, B \in F$ $A \cap B = \emptyset$ then $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$

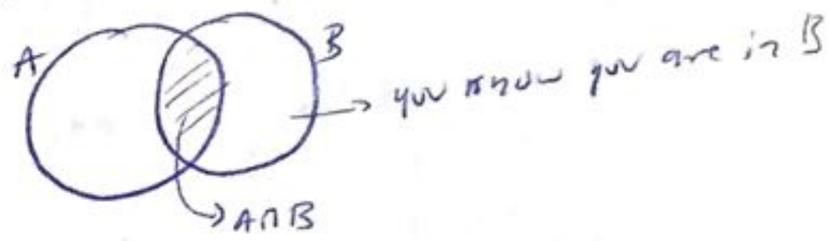
3*) Rule 2 holds for sequences: $A_1, A_2 \dots \in F$ $A_i \cap A_j = \emptyset$ $i \neq j$

$\mathbb{P}(\bigcup A_i) = \sum_i \mathbb{P}(A_i)$

⊛ Once you have Ω you can construct F and then construct $\mathbb{P}$

$(\Omega, F, \mathbb{P})$: probability triple

Def: Conditional probability: $\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$, given that A, B: observables $\in F$



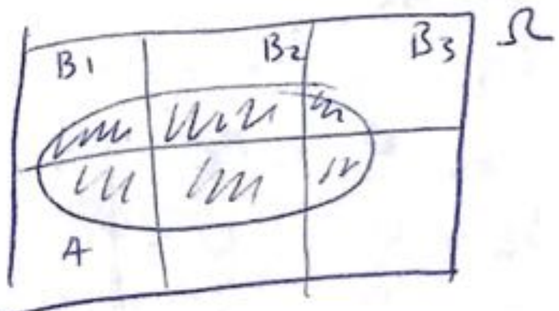
Law of Total Probability: $A, B_1, \dots B_n \in F$ $\bigcup B_i = \Omega$ $B_i \cap B_j = \emptyset$ $i \neq j$

$$\mathbb{P}(A) = \sum_{i=1}^{\infty} \mathbb{P}(A|B_i)\cdot\mathbb{P}(B_i)$$

$$= \sum_{i=1}^{n} \frac{\mathbb{P}(A \cap B_i)}{\mathbb{P}(B_i)} \mathbb{P}(B_i)$$

$$= \sum_{i=1}^{n} \mathbb{P}(A \cap B_i) \qquad (A \cap B_i) \cap (A \cap B_j) = \emptyset$$

$$= \mathbb{P}(A)$$



Def: A RV is a function from Ω to $\mathbb{R}$ 1) $\underline{X}: \Omega \to \mathbb{R}$

2) $\{w : \underline{X} \le x\} \in F$ (the little X I can observe), for all $x \in \mathbb{R}$

$(\Omega, F, \mathbb{P})$ $F_{\underline{X}}(x) = \mathbb{P}(\underline{X} \le x)$ CDF

Let $X_1 \dots X_n$ of RV's joint CDF $F_{\underline{X}_1 \dots \underline{X}_n}(x_1, x_2, \dots x_n) := \mathbb{P}(\underline{X} \le x_1, \dots \underline{X}_n \le x_n)$ ← defined AND, AND

$\underline{X}$ ∩ ... ∩ ...

we say X and Y are independent if $F_{XY}(x,y) = F_X(x) \cdot F_Y(y)$

we can now write the conditional density (continuous RVs): $F_{Y|X}(y|x) = \frac{f_{\underline{X}\underline{Y}}(x,y)}{f_{\underline{X}}(x)}$

$f_{\underline{X}}(x) > 0$

$r(X) = \mathbb{E}[Y|X=x] = \int_{-\infty}^{\infty} y \cdot f_{Y|X}(y|x)\,dy$ ⇝ Regression function

prediction is always conditional

$r(X) =$ "$\mathbb{E}[Y|X]$" you calculate this for every X and then plug it in $r(X)$ (you will more)

$\mathbb{E}[(Y - r(\bar{X}))^2]$

14/11/2024

Markov Chains:

- We have some M: (with states $\{1,2,3,4\}$)

$$P = \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{3} & 0 & 0 & \frac{2}{3} \\ \frac{1}{3} & \frac{1}{3} & 0 & \frac{1}{3} \\ \frac{1}{5} & \frac{2}{5} & \frac{2}{5} & 0 \end{bmatrix}$$

(initial state $X_0 = 4$)

- check if the rows sum to one

a) Prob to return to X_0 in 2 steps?

$$P(X_2 = 4 \mid X_0 = 4) = \sum_{1}^{4} P(X_2 = 4 \mid X_0 = 4,\ X_1 = 1) \cdot P(X_1 = 1 \mid X_0 = 4)$$

$$= \sum_{1}^{4} P(X_2 = 4 \mid X_1 = 1) \cdot P(X_1 = 1 \mid X_0 = 4) = 0 + \frac{2}{3} \cdot \frac{2}{5} + \frac{2}{5} \cdot \frac{1}{3} + 0 = \frac{2}{5}$$

- Find stationary distribution:

$$\pi = P\pi$$

$$\pi_1 = \frac{1}{2}\pi_2 + \frac{1}{2}\pi_3$$
$$\pi_2 = \frac{1}{3}\pi_1 + \frac{2}{3}\pi_4$$
$$\pi_3 = \frac{1}{3}\pi_1 + \frac{1}{3}\pi_2 + \frac{1}{3}\pi_4$$
$$\pi_4 = \frac{1}{5}\pi_1 + \frac{2}{5}\pi_2 + \frac{2}{5}\pi_3$$

$$\pi_1 = \frac{46}{206},\ \pi_2 = \frac{60}{206},\ \pi_3 = \frac{45}{206},\ \pi_4 = \frac{55}{206}$$

(2) Coin Flip: You flip a coin until you get 3 tails in the row

Define a MC with the states:

S - init (before flipping)

T = one tail

H = a head (which means your stunk of T is over)

TT = 2 tails in a row

TTT = 3 tails in a row

Compute transition matrix

$$P = \begin{array}{c} \\ S \\ H \\ T \\ TT \\ TTT \end{array} \begin{array}{c} \begin{array}{ccccc} S & H & T & TT & TTT \end{array} \\ \begin{bmatrix} 0 & 1/2 & 1/2 & 0 & 0 \\ 0 & 1/2 & 1/2 & 0 & 0 \\ 0 & 1/2 & 0 & 1/2 & 0 \\ 0 & 1/2 & 0 & 0 & 1/2 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \end{array}$$

$E[S]$ = expected number of steps to get TTT from state S. Compute $E(H), E(S), E(T), E(TT), E(TTT)$

$E(TTT) = 0$: you are already in this state - you don't need to do anything else

$E(S) = E(H)$: if you get head is at you start again.

Supervised Learning

(X, y) $\quad y \in \{0, 1, \dots 1_{C-1}\}$: classes

X → comes from the generator

y → comes from the supervisor

$\Gamma_{Y|X}(y|x)$ conditional distribution given x

Model space $M = \{g_\lambda(x) : g_\lambda(x) \in \{0, 1, \dots, 1_{C-1}\}\}$?

0-1 loss:

$$L(z, u) = \begin{cases} 1 & \text{if } u \neq z \\ 0 & \text{if } u = z \end{cases}$$

$R(\lambda) = \mathbb{E}[L(g_\lambda(X), Y)]$ ⇝ Risk

* Goal is always minimise Risk $\quad \lambda^* = \arg\min_\lambda R(\lambda)$

Linear Classifier

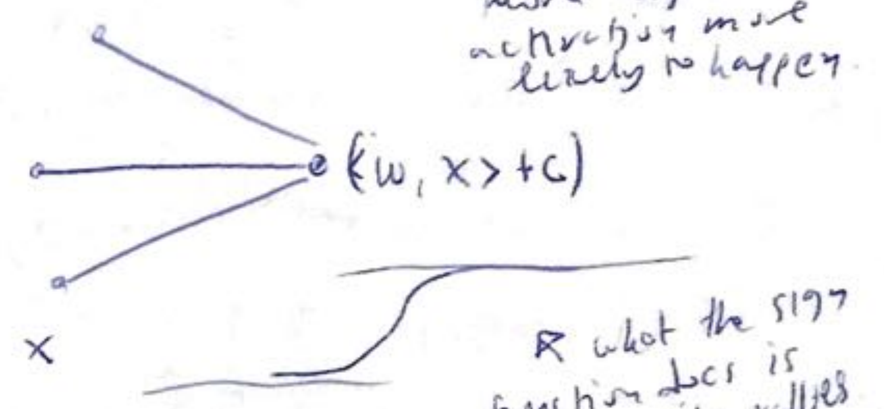
consider $x \in \mathbb{R}^d$, $y \in \{-1, 1\}$ and let $w \in \mathbb{R}^d$ and $c \in \mathbb{R}$

$g_w(x) = \text{sign}(\langle w, x \rangle + c)$ ← (McCullah Pitts 43)

$$L(y, g_w(x)) = \frac{-\text{sign}(\langle w, x \rangle + c)\, y) + 1}{2}$$

$g \neq g_w(x)$ $\quad L(y, g_w(x)) = \frac{-(-1)+1}{2} = 1$

$y = g_w(x)$ $\quad L(y, g_w(x)) = \frac{-(-1)+1}{2} = 0$



Dimension Trick:

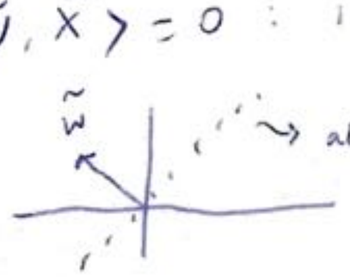
Redefine $\tilde{w} = (w_1, \dots w_d, c)$ $\quad \tilde{x} = (x_1, \dots, x_d, 1)$ to remove the c (simplify the problem)

$\langle x, w \rangle + c = \langle \tilde{x}, \tilde{w} \rangle$

The problem to solve:

(there is a distance between them) well-seperated

$\langle \tilde{w}, \tilde{x} \rangle = 0$: is a plane passing through the origin



→ all vectors in this line are perpendicular so we get the zero vector

Perception Algorithm: (Rosenblatt 60)

① set $w = (0, 0, \dots 0)$ $\quad d+1$

② we have points (x_i, y_i) $i = 1, 2 \dots n$ if there is a point (x_i, y_i) such that $\langle w, x_i \rangle \cdot y_i < 0$

(classification was wrong) : $w \leftarrow w + x_i y_i$ $\quad \in \mathbb{R}^{d+1}$, $\in \{-1, 1\}$

Repeat until all satisfy $\langle w, x_i \rangle y_i > 0$

Why this update rule?

we have $\langle w, x_i \rangle y_i < 0$, we want $\langle w, x_i \rangle y_i > 0$

This is a function of w, find which direction we need to change w to create the correct sign

$\nabla_w (\langle w, x_i \rangle y_i) =$

$\nabla_w (\underbrace{f(w)}_{\mathbb{R}^{d+1}}) = \left(\frac{\partial f}{\partial w_1}, \frac{\partial f}{\partial w_2} \cdots \frac{\partial f}{\partial w_{d+1}} \right)$ ⇝ direction of steepest (best) increase

$=$ we have $\nabla_w (\langle w, x_i \rangle y_i) = (w_1 (x_1)_1 + w_2 (x_2)_2 + \cdots w_{d+1} \cdot (x_i)_{d+1}$

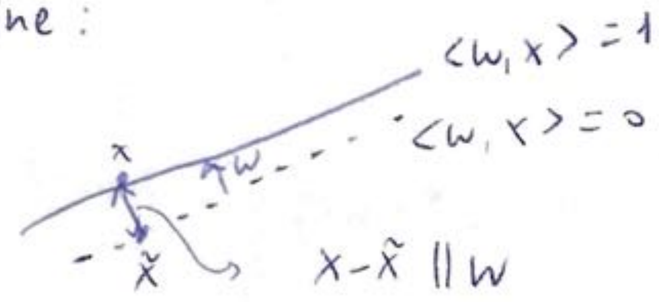
$((x_i)_1 \cdot y_{i_1}) \cdots (x_i)_{d+1} \cdot y_{i_{d+1}} = x_i y_i$

If there are many solutions to a problem this algorithm will stop once it found one solution. (Hard Margin).

Support Vector Machine:

Take $\|w\| = 1$

$\langle w, x \rangle = 1$



$|(x - \tilde{x}) w| = \|x - \tilde{x}\| \underbrace{\|w\|}_{1} = \|x - \tilde{x}\|$

if $\|w\| \neq 1$ then $\|x - \tilde{x}\| = \frac{1}{\|w\|}$

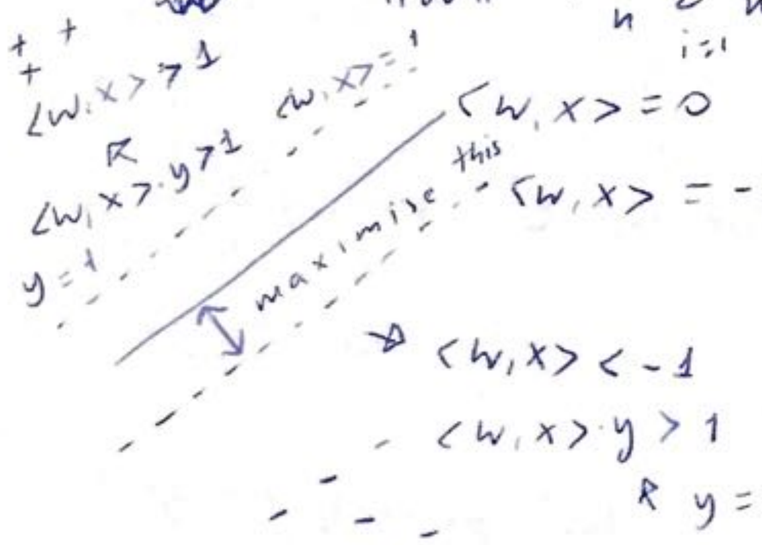
Consider the function : $\max\{0, 1 - \langle w, x \rangle y\}$ "Hinge loss"

minimise $\|w\|$ (you maximise the distance in other words)

under the constraint $\langle w, x_i \rangle y_i \geq 1$ for all $i = 1, \ldots, n$

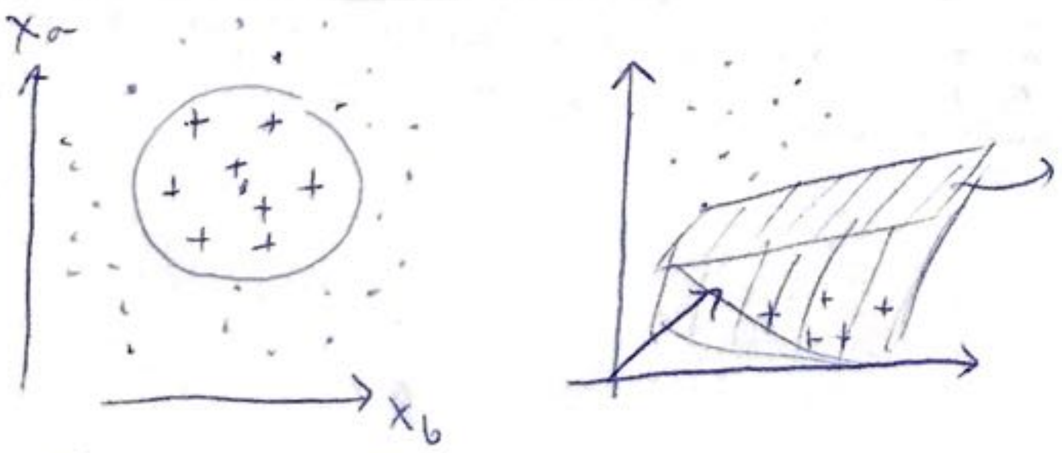
Soft margin SVM ⇝ constant

minimise : $\|w\|^2 + \frac{c}{n} \sum_{i=1}^{n} \max\{0, 1 - \langle w, x_i \rangle y_i\}$



When data do not have a linear seperator

$\Phi(x_a, x_b) = (x_a, x_b, x_a^2 + x_b^2)$



The Kernel Trick : Perception update rule : intial $w = 0$ $\quad y_i = +1$ or -1

$w \leftarrow w + x_i y_i$ we will find w of the form $w = \sum_{i=1}^{n} c_i x_i$

$\langle w, x_k \rangle = \sum_{i=1}^{n} c_i \underbrace{\langle x_i, x_k \rangle}_{K_{ij}}$ ← Gram matrix

we have a ϕ

$\phi(x_i)$

$\langle z_i, z_k \rangle = \sum_{i=1}^{n} \tilde{c}_i \underbrace{\langle \phi(x_i), \phi(x_k) \rangle}_{K_{ik}}$

if we have function $k(a,b)$

such that $\tilde{K}_{ik} = k(x_i, x_k)$ ~> we just need K and we don't need the ϕ anymore

Def: A kernel is a function $k(a,b)$ symmetric, if there is a function ϕ such that

$\phi: \mathbb{R}^k \to \mathbb{R}^d$: $K(a,b) = \langle \phi(x), \phi(y) \rangle$, $x,y \in \mathbb{R}^k$

Ex. 1) $K(a,b) = \langle a,b \rangle$ linear

2) $K(a,b) = (\gamma \langle a,b \rangle + m)^r$ polynomial

3) $K(a,b) = e^{-\|a-b\|^2}$ Radial Basis function (RBF)

4) $K(a,b) = \tanh(\gamma \langle a,b \rangle + m)$: sigmoidal → infinite dimension => very complicated functions

RBF:

$\langle w, x \rangle = 1$

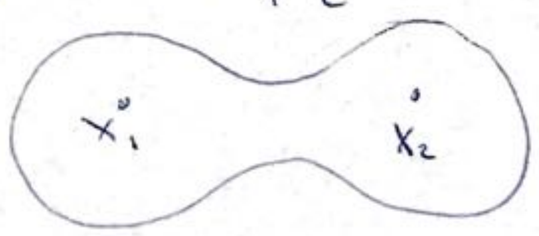
$1 = \sum_{i=1}^{n} c_i K(x_i, x)$ ~> find K such that this is true

1-point:

$\langle w, \hat{x}_i \rangle = 1$

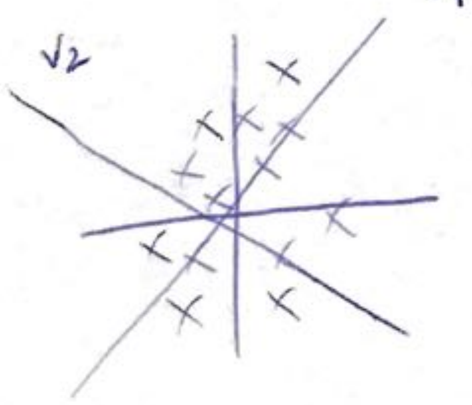
2-points:

$e^{-\|x_1 - x\|^2} + e^{-\|x_2 - x\|^2} = \frac{1}{2}$

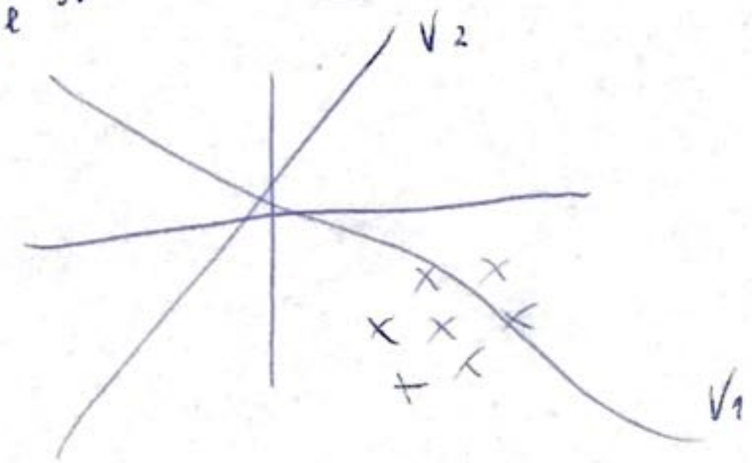


9/12

SVD
PCA = change of basis



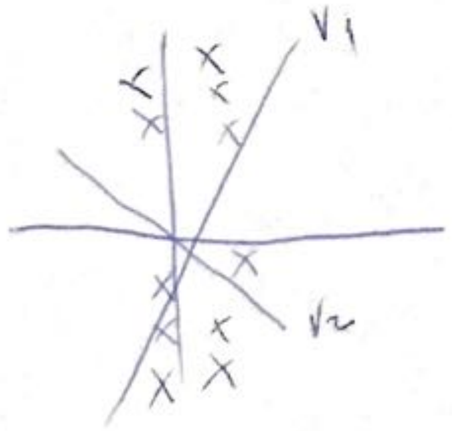
whiten
in sklearn

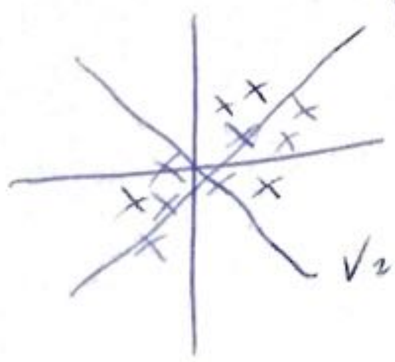


Centre your data!

standardised

If you standardise your data ⇒



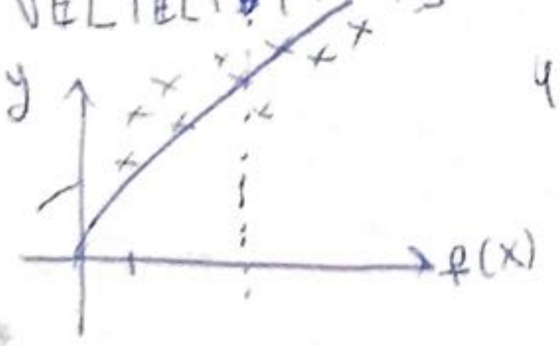


* does the data exploration in the train set only

Calibration: error:

$\sqrt{\mathbb{E}[|\mathbb{E}[Y|f(x)] - f(x)|^2]} =: c(f)$

you want it to be 45° $y=x$ otherwise you are overpredicting or underpredicting



$y \in \{0, 1\}$

$\mathbb{E}[|\mathbb{P}(Y=1 \mid f(x)) - f(x)|^2]$ → usually on test set → calibration set

$f(x)$: predicted probability

Feature engineering

* Construct features from data such that a "model" can understand it.
* Construct more features to make a model more powerful.

Ex: $f_{a_0, a_1}(x) = a_0 + a_1 \cdot x$

$z = x^2$

$f_{a_0, a_1, a_2}(x, z) = a_0 + a_1 x + a_2 x^2$

* Transformations of features

Ex: replace x with $\log(x)$

* Transformation of the target:

Ex: Y is price

$y' = \log(y)$

log scale $Y_1' - Y_2' = \log\left(\frac{Y_1}{Y_2}\right)$: proportional error

Test data:

Dictionary: an enumeration of all possible words in your data

Bag of words model: ignore order / content ⇒ how many times does word x appears?

Term frequency - Inverse Document Frequency (Tfid Vectorizer in python)

↳ is better than count: $f_{t,d}$ ← document
↳ term

Count of how many times t appears in d.

$TF(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t,d}}$ ← number of words in d.

$IDF(t) = \log\left(\frac{N}{|\{d : t \in d\}|}\right)$ N: number of docs

↳ in how many document does this word appear?

$TFIDF(t,d) = TF(t,d) \cdot IDF(t)$