Business statistics for Data Science

population vs sample

population size	N	$\frac{\sum_{i=1}^{i=N} x_i}{N}$
population mean	μ	
population variance	σ^2	$\frac{\sum_{i=1}^{i=N} (x_i - \mu)^2}{N}$
sample size	N	
sample mean	$\bar{X}$	$\frac{\sum_{i=1}^{i=N} X_i}{N}$
sample variance	s^2	$\frac{\sum_{i=1}^{i=N} (x_i - \bar{x})^2}{n-1}$

Why $n-1$ for sample variance and not n, because of Bessel's correction

Bessel's correction: you know the value of the n^{th} book or item (use a group of books thickness)

for measurement of central tendency, we uses median, mode mean,
we use median when mean is skewed by outliers.

When handling missing dataset, if normal distribution we use mean
if skewed we used median
if missing categorical data mode.
- " " " " is to large
find the ratio of categories and
apply the ratio to missing data

Probability

coin → Head

(H) (T)

→ $\frac{1}{2}$ (nu:) = 0.5

probability = $\frac{\text{no. of favourable outcome}}{\text{Total no. of possible outcome}}$

Sample space

Event

Mutually exclusive events — Pr of 2 or 5 on a die

$P(A \cup B) = P(A) + P(B)$ — or

$P(2)$ $P(5)$

not mutually exclusive event

$$P(A \cap B) = P(A) + P(B) - P(A \cap B)$$

- $P(A)$: probability of being a boy
- $P(B)$: Pr of being a A
- $P(A \cap B)$: Pr of being a boy and A

Independent events

$$P(A \& B) = P(A) * P(B)$$

$$P\left(\frac{A}{B}\right) = P(A)$$

$$P\left(\frac{B}{A}\right) = P(B)$$

rolling a dice and getting heads on a coin

$SS = \{1, 2, 3, 4, 5, 6\}$

$A \rightarrow \{2, 4, 6\}$

$P(A)$

Cumulative probability: likelihood a value of a random is within a given range.

flip to coins, probability of one or fewer heads

(x ≤ 1)

Sample space = { HH, HT, TH, TT }

(x = 0)

$P(X \leq 1) \rightarrow$

$P(X \leq 1) = P(X=0) + P(X=1)$

$= \frac{1}{4} + \frac{2}{4} = \frac{3}{4}$

Conditional probability

return allowed

6 Red marbles → R/B 6/10

4 black marbles ← 4/10

R/B 6/10 × R/B 4/10 = 16/100

return not allowed

6 Red marbles → 6/10

4 black marbles → 4/9

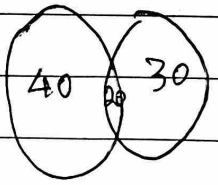
6/10 × 3/9

$P(B|A) = P(A \cap B) / P(A)$

$\frac{4}{10} \times \frac{3}{9} = \frac{12}{90}$

100 people 40 bought sports drinks 30 bought snacks

20 bought sport drinks and snacks



① $P(A) = \frac{40}{100} = 0.4$

Bayes Theorem
Extension of Conditional Probability

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

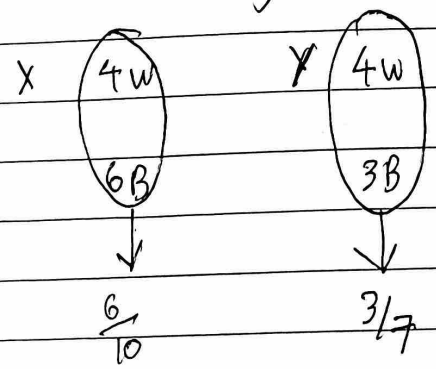
$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

$$P(A \cap B) = P(B|A) \times P(A)$$

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

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

A Bag X contains 4 white & 6 black balls while another Bag Y contains 4 white and 3 black balls. one ball is drawn at random from one of the bags, and it is found to be black find the probability that it was drawn from Bag X.



$E_1 \rightarrow$ Event of choosing Bag X

$E_2 \rightarrow$ " " " " " Y

$$P(E_1) = P(E_2) = \frac{1}{2}$$

$P(A|E_1) = P(\text{Black from Bag X}) = 6/10$

$P(A|E_2) = P(\text{Black from Bag Y}) = 3/7$

$$P(E_1|A) = \frac{P(E_1) \times P(A|E_1)}{P(E_1)P(A|E_1) + P(E_2)P(A|E_2)}$$

probability of getting one head

$$P(x=1) = \frac{5}{32} = \frac{{}^5C_1}{32} = \frac{5!}{1!\,(5-1)!} \times \frac{1}{32}$$

$$= \frac{5!}{4!} \times \frac{1}{32}$$

$$= \frac{5}{32}$$

$$P(x=2) = \frac{{}^5C_2}{32} = \frac{5!}{2!\,(5-2)!} \times \frac{1}{32}$$

$$= \frac{120}{2 \times 6} \times \frac{1}{32} = \frac{10}{32} = \frac{5}{16}$$

$$P(x=3) = \frac{{}^5C_3}{32} = \frac{5!}{2!\,3!} \times \frac{1}{32} = \frac{10}{32}$$

$$P(x=5) = \frac{5}{32}$$

$$P(x=5) = \frac{1}{3}$$

10/32

5/32

1/32

$$\frac{\frac{1}{2}\times\frac{6}{10}}{\frac{1}{2}\times\frac{6}{10}+\frac{1}{2}+\frac{3}{7}} = \frac{6}{20} \div \left(\frac{6}{20}+\frac{3}{14}\right)$$

$$= \frac{6}{20} \div \left(\frac{42+30}{140}\right)$$

$$\frac{6}{20} \div \left(\frac{72}{140}\right)$$

$$\frac{6}{20} \times \frac{7}{72} = \frac{84}{900}$$

$$= \frac{42}{450} = \frac{14}{150}$$

$$\frac{7}{12} = \frac{7}{75}$$

ner Bag
awn at
be blac

X

Y

probability distribution

uniform Distribution

- Discrete uniform : discrete integers
- continuous uniform : continuous float

Binomial distribution

$$P(x) = {}_nC_x p^x (1-p)^{n-x}$$

$2\times2\times2\times2\times2 = 32$

Flip 5 coins probability

Heads = 0 possible outcome $= 2^5 = 32$

$P(x=0) = TTTT = \frac{1}{32} = \frac{5C_0}{32} = \frac{5!}{0!5!} \times \frac{1}{32} = \frac{1}{32}$

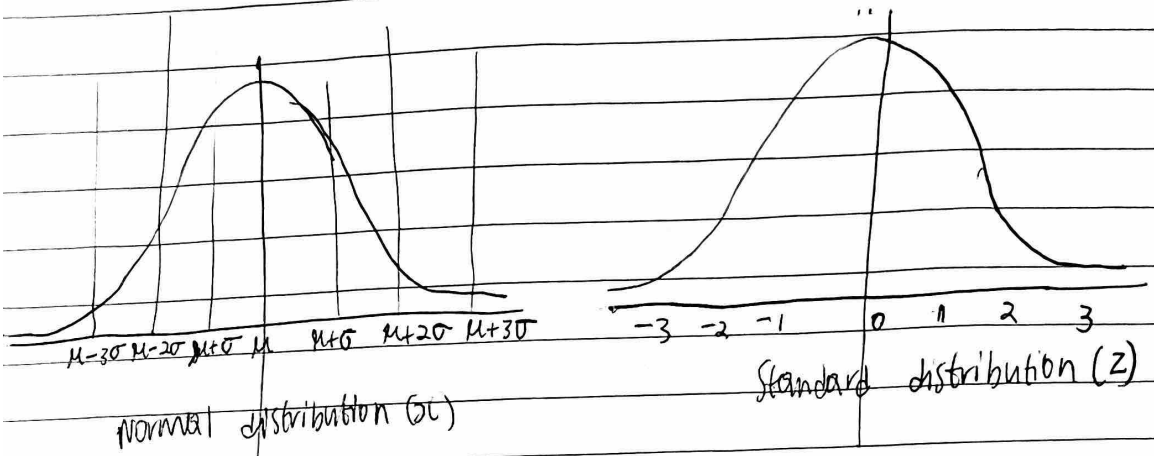
Calculating probability with Z-score for Normal Distribution

- Find out how far the value of X is from μ (in terms of "number of σ")
- The number is called a "Z-score" or "standard score" or "Z value"

$$Z = \frac{X - \mu}{\sigma}$$

$X = \mu$

$$Z = \frac{\mu - \mu}{\sigma} = 0$$



Using the Z table to make predictions

200 students

John scored 700/1000 in a test

How many students did John do better than

Average score is 600

S.D is 150

$$\frac{700 - 600}{150} = 0.67$$

Check Z table

$0.7486 \times 100 = 74.86\%$ He did better than

Kurtosis

osis measure the thick
degree of tailed

Types of Kurtosis

platy kurtic
meso kurtic

Poisson distribution

This is a discrete distribution, occuring over period like an hour

$$P(X=x) = \frac{\lambda^x e^{-\lambda}}{x!} \qquad e = 2.71828$$

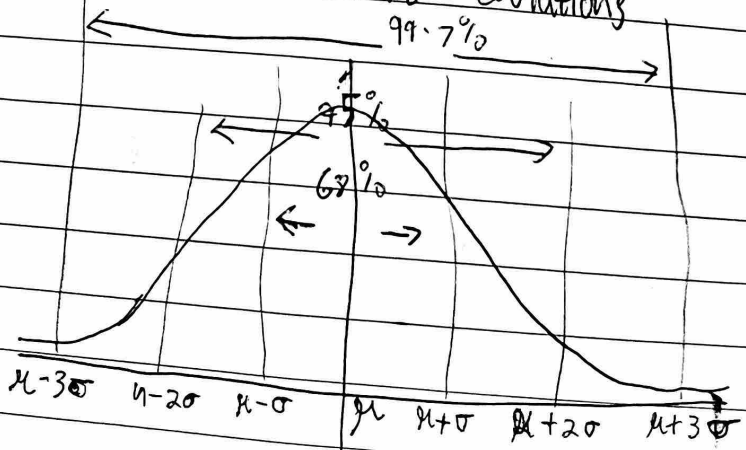
- Average number of calls for a call center is per hour $\lambda = 10$
- probability of having different number of calls in an hour could be calculated as:

$$P(x = 0 \text{ calls}) = \frac{10^0 e^{-10}}{0!} = \frac{1 \cdot 0.000045}{1} = 0.0045\%$$

$$P(x = 3 \text{ calls}) = \frac{\lambda^x e^{-\lambda}}{x!} = \frac{10^3 e^{-10}}{3!} = 0.75\%$$

Normal distribution (Gaussian distribution)

- Graphically, normal distribution appears as a bell curve
- for all normal distribution, 68.2% of the observations will appear within plus or minus one standard deviation of the mean, 95.4% of observations will fall within +/- two standard deviations; and 99.7% within +/- three standard deviations



- Null Hypothesis: In Inferential Statistics, null hypothesis is a general statement that there is no relationship between two measured phenomena (An assumption)

a company production is = 50 unit/ per day

- Alternative hypothesis: The alternative hypothesis is the hypothesis used in hypothesis testing that is contrary to the null hypothesis.

a company production is != 50 unit/day etc

Null hypothesis: =, greater than or equal, less than or equal

Alternative hypothesis: !=, <, >

Tailed test

One tailed test: A test of statistical hypothesis, where the region of rejection is on only one side of the sampling distribution

>= 800

Two tailed test: two sided critical areas

!= 800

p-Value: p value or calculated probability, is the probability of finding the observed, or more extreme, results when the null hypothesis of study question is true

Types of test

1) t-test
2) z-test
3) ANOVA
4) Chi Square Test
5) correlation test

...ver Treatment

Correlation

Correlation is similar to covariance, also for measuring how two variable are related or moving together

+1 to −1 Ranges

Direction and magnitude are useful

$$\rho_{xy} = \text{correlation}(x,y) = \frac{\text{Cov}(x,y)}{\sqrt{\text{var}(x)}\ \sqrt{\text{var}(y)}}$$

$$\frac{\text{Cov}(x,y)}{\sigma_x\ \sigma_y}$$

$$\text{Cov}(x,y) = \frac{1}{n-1} \sum \frac{(x-\bar{x})(y-\bar{y})}{\frac{29.8}{4} = 7.45}$$

Standard deviation

$$\sigma = \sqrt{\sum \frac{(x_i - \mu)^2}{n-1}}$$ → Sample mean

0 − 0.3	weak linear relationship
0.3 − 0.7	moderate linear relationship
0.7 − 1	strong linear relationship

Hypothesis testing?

statistical method that is used in making statistical decisions using experimental data. used for two mutually exclusive events

- Null Hypothesis statement to phenomena
- a comp
- Alternative used in he
- a comp
- null hypo
- Alternative
- one tailed of regres
- Two tailed
- p-value finding of study

Types

1. t − t
2. z −
3. Anov
4. Chi
5. Corre

$\frac{75}{100} \times 200 = 150$ students John did better than

) % of students shorter than 1.5m in a normal distribution
$\sigma = 0.06\,m$ mean $= 1.6\,m$

$$z = \frac{X-\mu}{\sigma} = \frac{1.5-1.6}{0.06} = \frac{-0.1}{0.06} = \frac{100}{6} = \frac{5}{3} = -1.67$$

0.0475

4.75% shorter than 1.5m

4.5 Covariance

Covariance is a statistical term that refers to a relationship between two random variables on how one variable changes would impact the other one

Covariance can either be positive or negative
$-\infty$ to $+\infty$

Magnitude does not really matter with covariance

number of data

$$\text{Covariance}(x,y) = \left(\frac{1}{n}\right)\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$$

mean of x — mean of y

1st value — 2nd value

n = number of data

$\left(\frac{1}{n-1}\right)\Sigma$

within sample size (nu:)

ANOVA (Analysis of Variance)

It's a statistical technique that is used to check if means of two or more group are significantly different from each other. It's parametric test for hypothesis testing

H_0 : mean1 = mean2 = mean3. There is no significant difference between mean values of each group

H_1 : means aren't equal

Sum of Squares (SS): Quantifies the variability within the group

Degree of freedom (df):

df (between groups): no of groups − 1

df (within groups): no of observations − no of groups

mean squared (MS): It is the average variation either between groups or within groups. MS = SS/df

F_ statistics: It is the test statistics in one way ANOVA

F = MS (between groups) / MS (within groups)

if p-value > significance value → we accept the null hypothesis else not

Significance level : 0.05

T-test

makes use of hypothesis test and alternative test to make decision, if $p > 0.05$, then we follow the null hypothesis

H_0 – Null Hypothesis

H_1 – Alternative Hypothesis

Z-test

A Z-test is used to compare the mean of two given samples and infer whether they are from the same distribution or not. we do not implement Z-test when the sample size is less than 30.

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Chi-Square

observed

$$x^2 = \sum \frac{(O - E)^2}{E}$$

Expected

Expected

H_0 : There is no association between gender & location

H_1 : There is association between gender & location

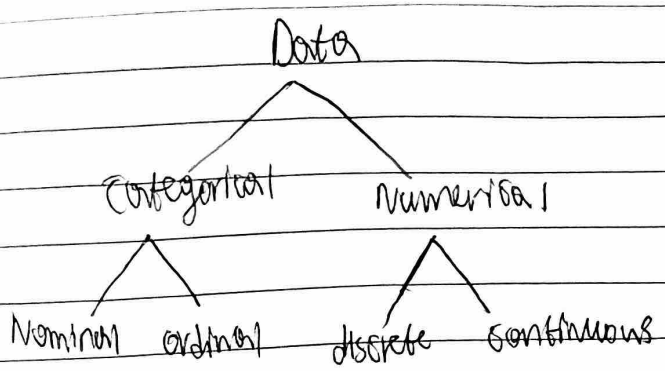
p-value > 0.05 --> Reject H_0

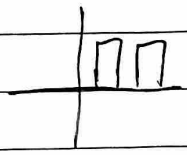
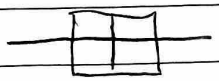
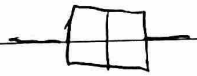
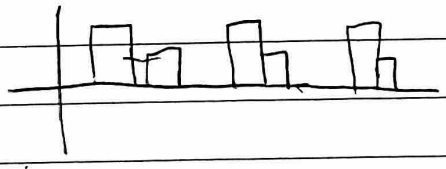
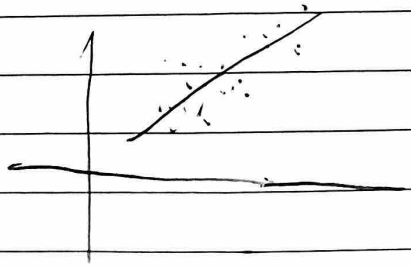
p-value < 0.05 --> Accept H_0

Expected = (Row total * column Total) / Grand Total

df = no row - 1 * no column - 1

other Treatment



Type of Data	what we see	Type of test
1 Categorical		1 sample test
1 Numerical		t-test
1 Numerical and 1 categorical		t-test or Annova
2 categorical		Chi square test
2 Numerical		Corelation

Practical Cleaning of Churning model

Row number	int
Customer ID	object *
Surname	object
Credit score	int
Geography	object
Gender	object
Age	int
Tenure	int
Balance	float
Num of Products	int
Hascrcard	object (binary) *
IsActivemember	object (binary) *
Estimated salary	float
Exit	object (binary) *

feature Scaling

making large numbers more digestable for our model using scalers

Standard scaler (standardization) when you have a normal distribution ensures that for each feature, the mean is zero and standard deviation is 1, bringing all feature to the same magnitude. Standardization helps scale down feature based on the standard normal distribution

Min-max scaler (normalization) when your distribution is not normal

Normalization helps you to scale down your features between a range 0 to 1

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Handling missing values.

(1) Delete Rows / columns
delete column if it has 75% missing values and no dependence

2% - 3% of data missing you can input or delete

(2) Replace with (mean) / (median) / (mode)
- mean: when there are outliers
- mode: uses for categorical data

(3) Algorithm implementation
KNN, Naive Bayes, Random forest are algorithms that have ability to handle missing value.

(4) predicting the missing values
prediction model is one of the advanced method to handle missing values. dataset with no missing values become training set and data set with missing value become the test set and the missing values is treated as target variable.

Practical handling missing values
- filling with mean, median, mode
- forward fill or back fill
- using Time series algorithm (too techious for this stage)

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{0.2}{0.4} = 0.5$$

$$\frac{0.2}{0.6} = 0.33$$

$$\frac{P(A \cap B)\,P(B)}{P(B)\,P(A \cap B) + P(A)\,P(A \cap B)}$$

$$\frac{0.2\,(0.4)}{0.2(0.4) + 0.6(0.2)}$$

$$\frac{0.08}{0.08 + 0.12}$$

$$\frac{0.08}{0.20}$$

EDA

Data Sourcing gathering data from multiple sources as external or internal data collection.

1. public data
2. private data

Data cleaning

Clean the data to improve the quality of the data for further data analysis

Handling missing values → Standardization of the data → outlier Treatment → Handle invalid values

Types of analysis

1. Univariate analysis
   one variable analysis
   can be used for numerical and categorical data
   mostly categorical data. distribution analysis using box plot and histogram

2. Bivariate analysis
   Analysis of two variables, can be done for both numerical and categorical

   two numerical — Scatterplot

   2 categorical — Bar chart

   Numerical & categorical — box plot, histogram

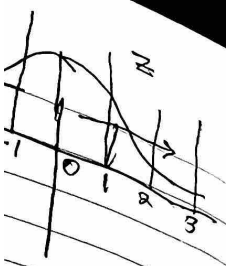
3. multivariate Analysis
   multiple variable analysis

Numerical Analysis

various analysis of numerical data

1 numerical — Box plot
2 numerical — Scatter plot

for multiple numerical functions you can just use corr() in python which gives the correlation in form of heat maps



Outlier Treatment

outliers are the most extreme values in data

Detect outliers using following methods:

1. Boxplot
2. Histogram
3. Scatter plot
4. Z-Score
5. Inter quartile range (values out of 1.5 time of IQR)

outlier treatment is important for time series forecasting

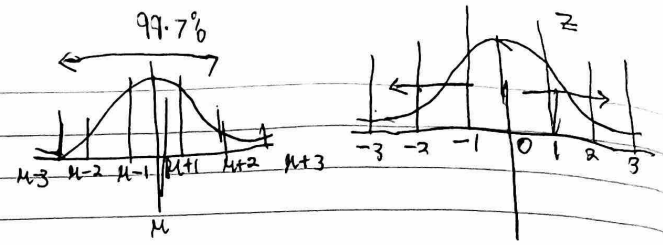
Handle outlier using following methods

1. Remove the outliers
2. replace outlier with suitable values by using the following methods
   - Quantile methods
   - inter quartile range
3. use ~~mode~~ ML model which are not sensitive to outliers. like:- KNN, Decision Tree, SVM, Naive Bayes, Ensemble methods.

Types of Data

- Qualitative
  - Nominal
    - Gender M/F
    - Location
  - ~~ordinal~~
    - Economic status
    - Grade
- Quantitative
  - Discrete
    - Whole number
  - Continuous
    - 37.7
    - 33.2

Example
Standardization
$\mu = 0$
Age



$$z\text{-score} = \frac{X - \mu}{\sigma}$$

$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}}$$

$(N-1)$ — sample

Age	Income (£)	New value
24	15000	$(15000 - 19000)/9643.65 = -0.4147$
30	12000	$(12000 - 19000)/9643.65 = -0.7258$
28	30000	$(30000 - 19000)/9643.65 = 1.1406$

SKlearn = has libraries to accomodate feature Scaling

Example
Normalization

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Age	Income (£)	New value
24	15000	$(15000 - 12000)/18000 = 0.16667$
30	12000	$(12000 - 12000)/18000 = 0$
28	30000	$(30000 - 12000)/18000 = 1$

Restart Windows
Shift + ESC
mount -n -o remount, rw/
passwd your-username
reboot

- model + dimensionality reduction + visualization

~~model + latent representation + dimensionality reduction + visualization with DAG~~

- model + Gradient methods

- model + dimensionality reduction + Lime

- model + latent representation + visualization

One - Hot Encoding Example

Location	loc-m	loc-Indo	loc-IND
malaysia	1	0	0
Indonesia	0	1	0
India	0	0	1
India	0	0	1
mal	1	0	0
mal	1	0	0

Dummy features

- Dummy encoding : an extension of one hot encoding, use dummy features. in here one of the features (columns) can be ignores
  n-1 = dummy features

Select ~~Date~~ id from ~~record Date~~ weather
Where

Select ID from weather
where

1
+1

Derived metrics

create a new variable from the existing variable to get a insightful information from the data

Age	Age months
25	25 x 12
28	
29	

derived metrics

- calculated from Data
- From Domain Knowledge
- Feature Binning
- feature encoding

Feature Binning

Feature binning converts or transforms continuous/numerical variable to categorical variable. It can also identify missing values and outliers.

- Equal width binning (grouping by width, size)
- Equal frequency binning (grouping by count)

0-20, 21-40, 41-60

0-20 → 2K

21-40 → 2K

46-50 → 2K

Feature Encoding

- Label encoding: label encoding is technique to transform categorical variables by assigning a numerical value to each of the categories only used when it the target/Y/dependent variable, bcos large number has priority
- One-Hot encoding: This technique is used when independent variables are nominal. It creates K different columns each for a category and replaces one column with 1 rest of the columns is 0
0, represents the absence and 1 represents the presence of that category

Corr() in heat maps

both numerical

box plot

al data

g

Activation function: Step functions

Threshold based activation function

< 70 Performed well

70 (>70) — poor

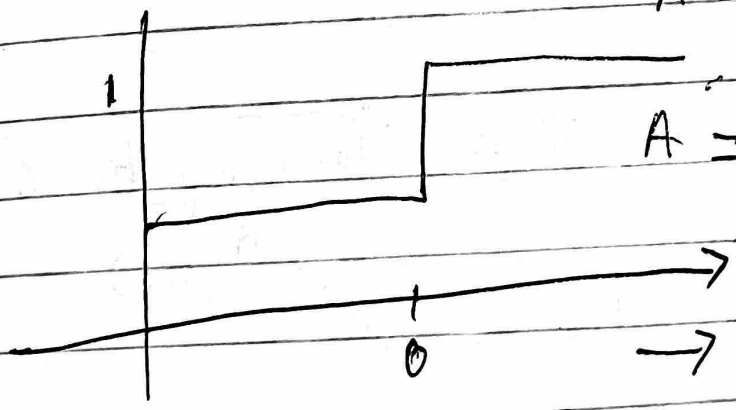
Step function

A = y > threshold

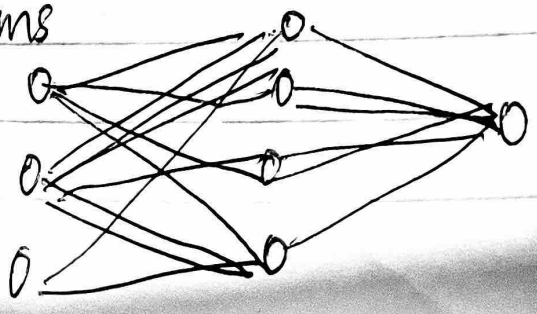
A = 1, y > 0

otherwise

A = 0



Step function use Binary classification problems



1 ↑ activated

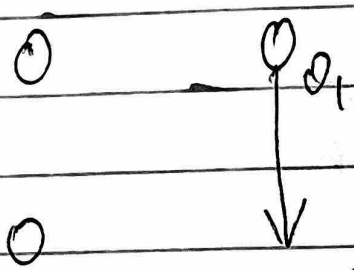
0 ↓ not activated

$\phi = 0$ (node is inactive)

$\phi \neq 0$ (node is active)

Activation functions

part of the hidden layers which decides if neuron should be fired or not



$$o_1 = \phi \sum (w_i x_i) + \text{bias}$$

ϕ ranges from $-\infty$ to ∞

Types of Activation functions

Step functions

Linear functions

sigmoid functions

Tan H functions

ReLu ✓✓

visualization

~~dimensionality~~

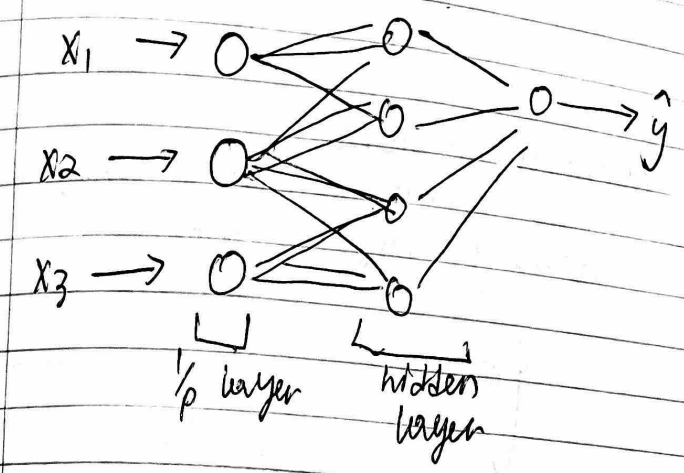
~~PCA~~

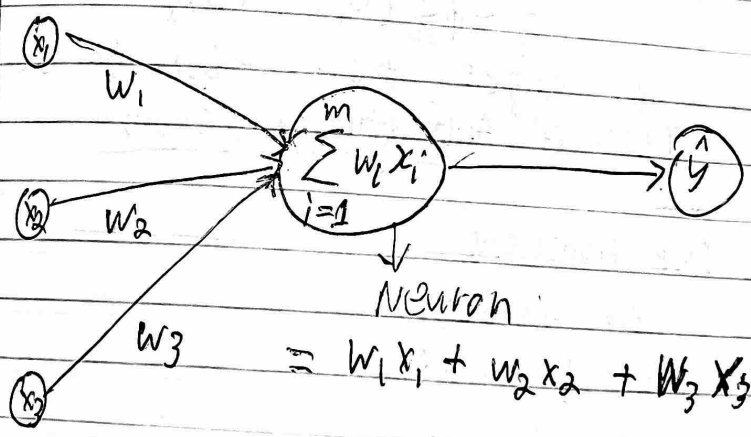
+ Lime

visualization

Deep learning

Neuron





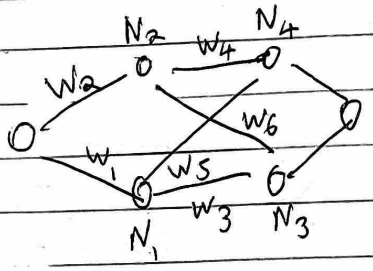
Neuron

$= w_1 x_1 + w_2 x_2 + w_3 x_3$

$w_1, w_2, w_3 \longrightarrow$ weights (connection strength)

$\hat{y} = \phi(\sum w_i \cdot x_i)$

$\phi \rightarrow$ Activation function



$$N_1 = w_1 x_1$$

$$N_2 = w_2 x_2$$

$$N_3 = w_3 N_1 + w_6 N_2$$

$$N_3 = w_3 w_1 x_1 + w_6 w_2 x_2$$

$$= m_1 x_1 + m_2 x_2$$

- Rarely used
- Advised not to be used for multiple hidden layers. (mostly used for first layer)

Activat

~~Sigmoid~~

Non

activation

last la

Δy

to

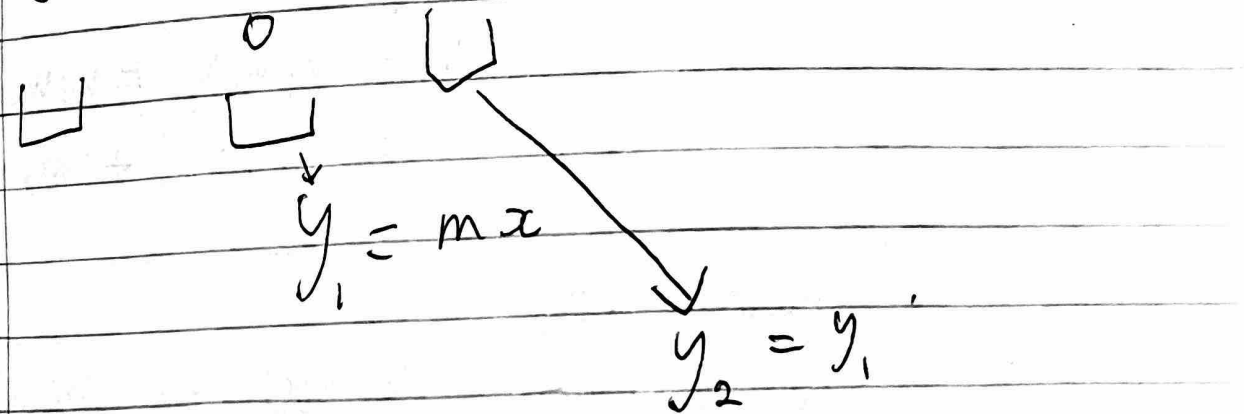
$= \sqrt{36 + 9}$

$= \sqrt{45}$

$= 6.7$

gradient $y = \frac{\Delta y}{\Delta x} = \frac{OB}{OA} = \frac{\sqrt{(8-8)^2 + (6-3)^2}}{(2-8)^2 + (3-3)^2}$

$= \frac{3}{6} = 0.5$



- Step function cannot be used for multi-class problem

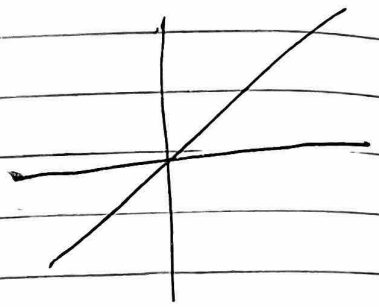
- Activation function : Linear function

$y = mx$

$x=0 \quad y=0$

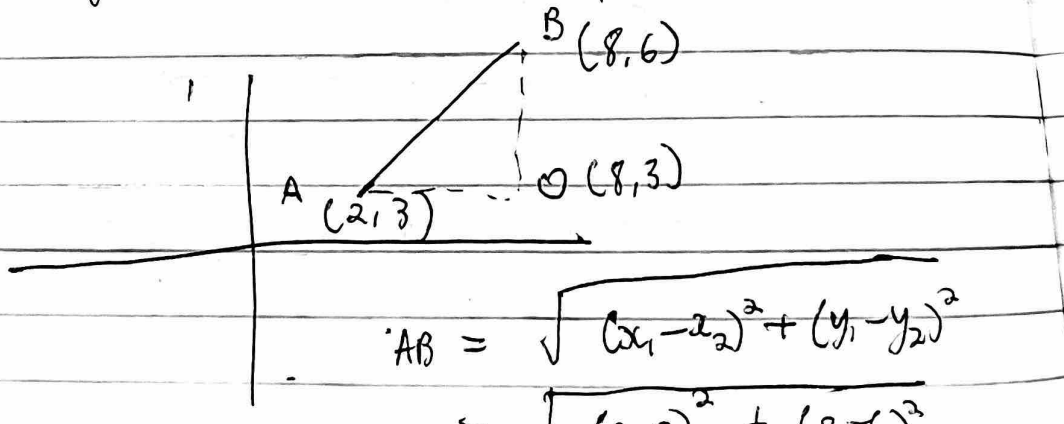
$x=1 \quad y=m$

$x=2 \quad y=2m$



$\frac{dy}{dx} = m$

gradient has no relationship to x



$$AB = \sqrt{(x_1-x_2)^2 + (y_1-y_2)^2}$$

$$= \sqrt{(2-8)^2 + (3-6)^2}$$

$= \sqrt{36 +$

$= \sqrt{$

$= 6.$

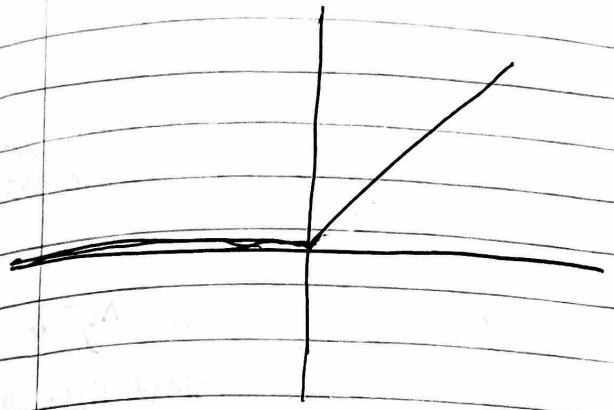
$y =$

gradient

0

0

0



① Non-Linear

$y = \max(0, x)$ range

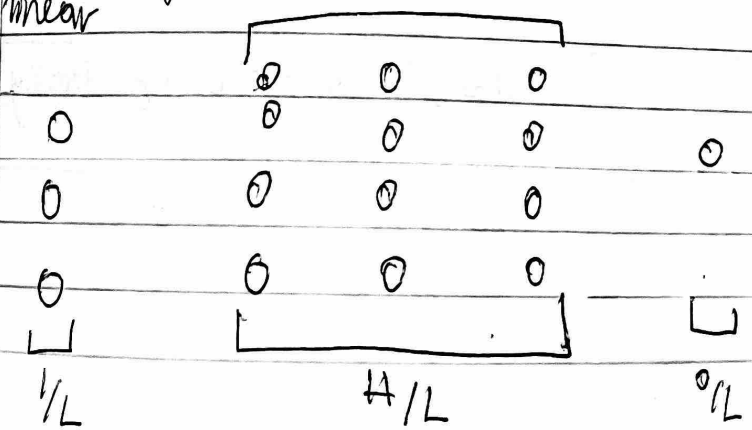
if $x < 0$ $y = 0$

if $x > 0$ $y = x$

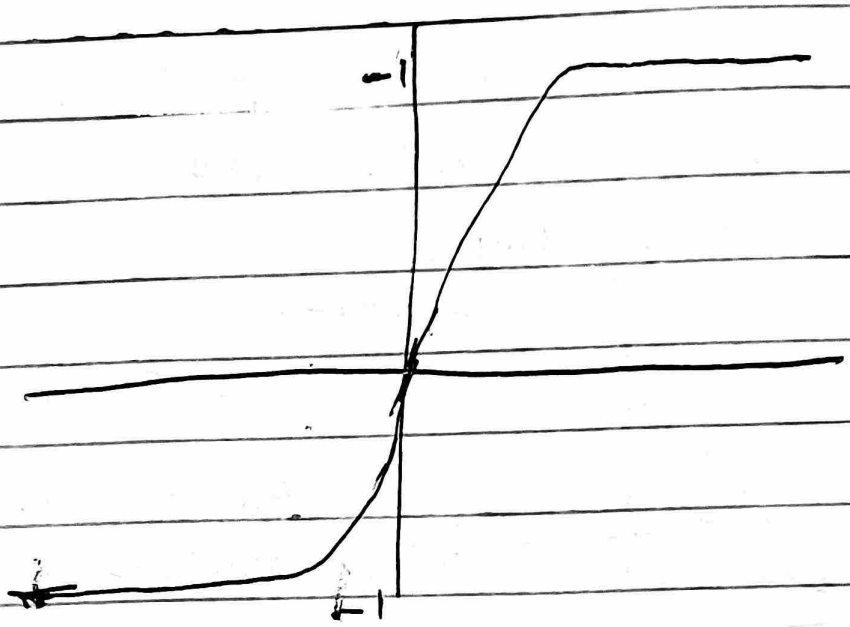
$(0, \infty)$

~~1 Sigmoid~~

Layers can be stacked because it's non linear



Tan H Function



Similar to Sigmoid

non-linear

Range (-1, 1)

~~Relationship btw Tanh and Sigmoid~~

$\tanh(x) = 2\phi(2x) - 1$

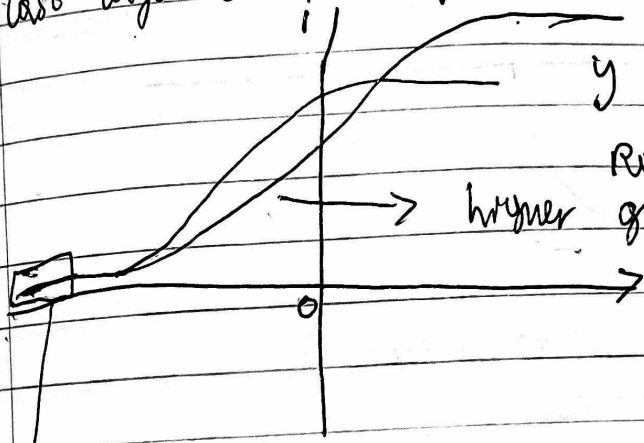
ReLu Function

Rectified linear unit (ReLU)

Activation function sigmoid function (ϕ)

~~Sigmoid~~

Non-linear in nature (one of the best activation function), mostly used in the last layer (output layer)



$$y = \frac{1}{1+e^{-x}}$$

Range always $(0,1]$

higher gradient at steepest part

$\Delta y \downarrow \simeq 0$ gradient $\simeq 0$, here network refuses to learn or drastically becomes low

$$\frac{dy}{dx} = \frac{d}{dx}\left(\frac{1}{1+e^{-x}}\right) = ?$$

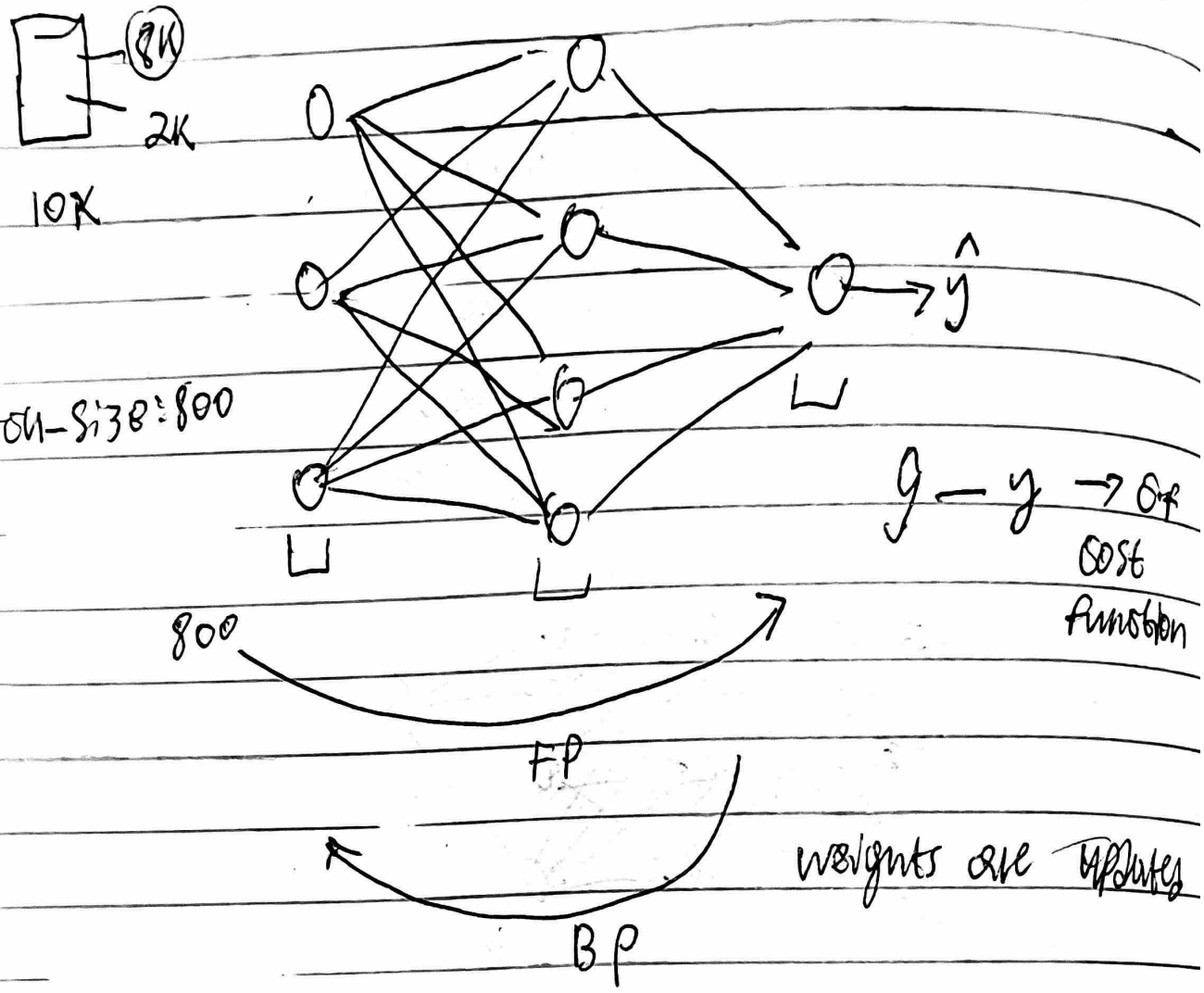
$w_6 w_2$

$w_6 w_2 x_2$

$m_2 x_2$

multiple hidden layer)

w_1 w_2 w_3 w_4



Iteration: one complete FP + BP

800 x 10 = 8000 test

$\frac{8000}{800}$ = 10 cycles

Epoch = the completion of a full training data points

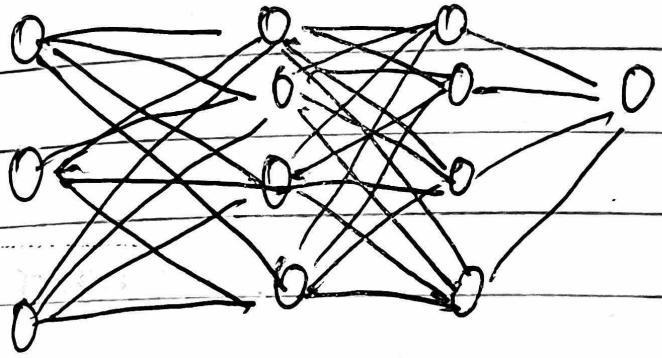
Backpropagation and forward pass

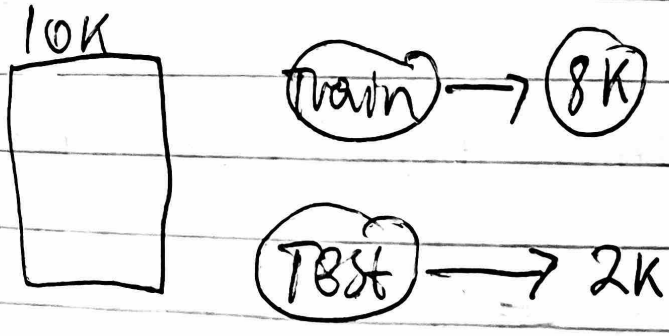
Customer churn prediction

Age

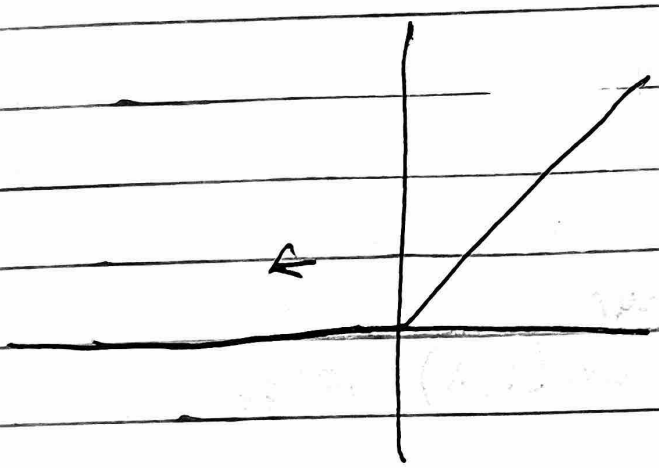
Tenure

Contract type





if sigmoid or tanH is used for large layer activation, it will be very dense and computationally expensive. to ensure sparse activation we use ReLu



neurons in this state are not active

TO solve dying Relu problem ensure neurons are not completely passive to create leaky Relu

$y \neq 0$ $y = 0.001x$ leaky relu

time

Artificial neural networks

features are inputs for input layer

- Traditional machine learning steps
+ Install libraries & importing them
+ importing dataset
+ Data preprocessing
+ feature scaling
+ Feature extraction
+ splitting data into train & Test
+ fitting the classifier to training data predictions
+ Analyzing predictions with actual test data

In ANN, Step # 4 to #6 is replaced by below:

+ initializing ANN
+ Adding input, Hidden & output layers
+ compiling the ANN
+ Fitting the ANN to the training dataset

NN tuning hyperparameters

Epochs is an hyperparameter

Optimizer is an hyperparameter

erative

minimum of

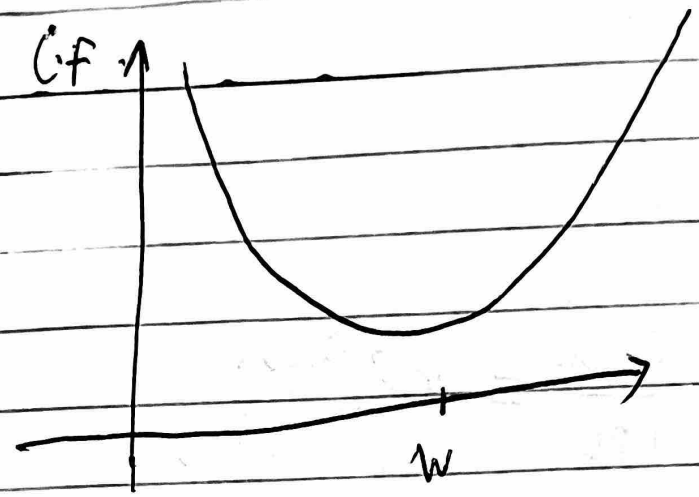
the steepest desce

y finding

using gradient

nimum

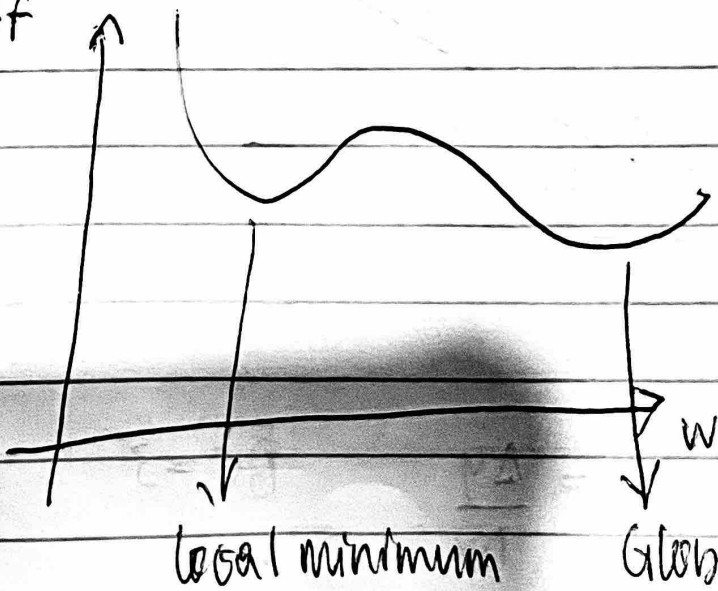
Cost function vs weight over time



Descent

Gradient: cost function decreasing, iterative optimization algorithm used to find the minimum of a function by taking steps in the direction of the steepest descent

Stochastic Gradient descent: continuously finding minimum past the local minimum by using gradient from mini-batch



w_1 w_2 w_3 w_4 ... w_{16}

$\hat{y}$

$y - \hat{y}$ → Cost function

weights are updates

plete FP + BP

st

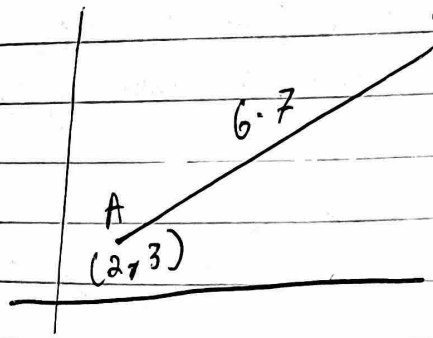
tion of a fun training

Training 10000
batch-size : 100
epochs : 20

$\frac{10000}{100} = 100$ cycles → 1 epoch

2000 cycles is 20 epochs

Gradient Descent



$y = mx + c$

$AB = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$

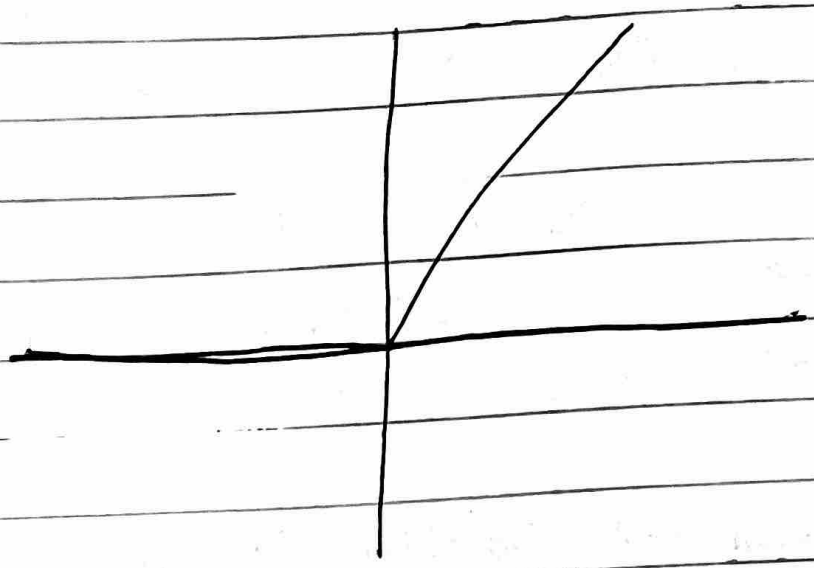
$= \sqrt{6^2 + 3^2} = \sqrt{45} = 6.7$

$g = \frac{\Delta y}{\Delta x}$ $\quad BC = 3$

$AC = 6$

$g = \frac{3}{6} = 0.5$

Relu

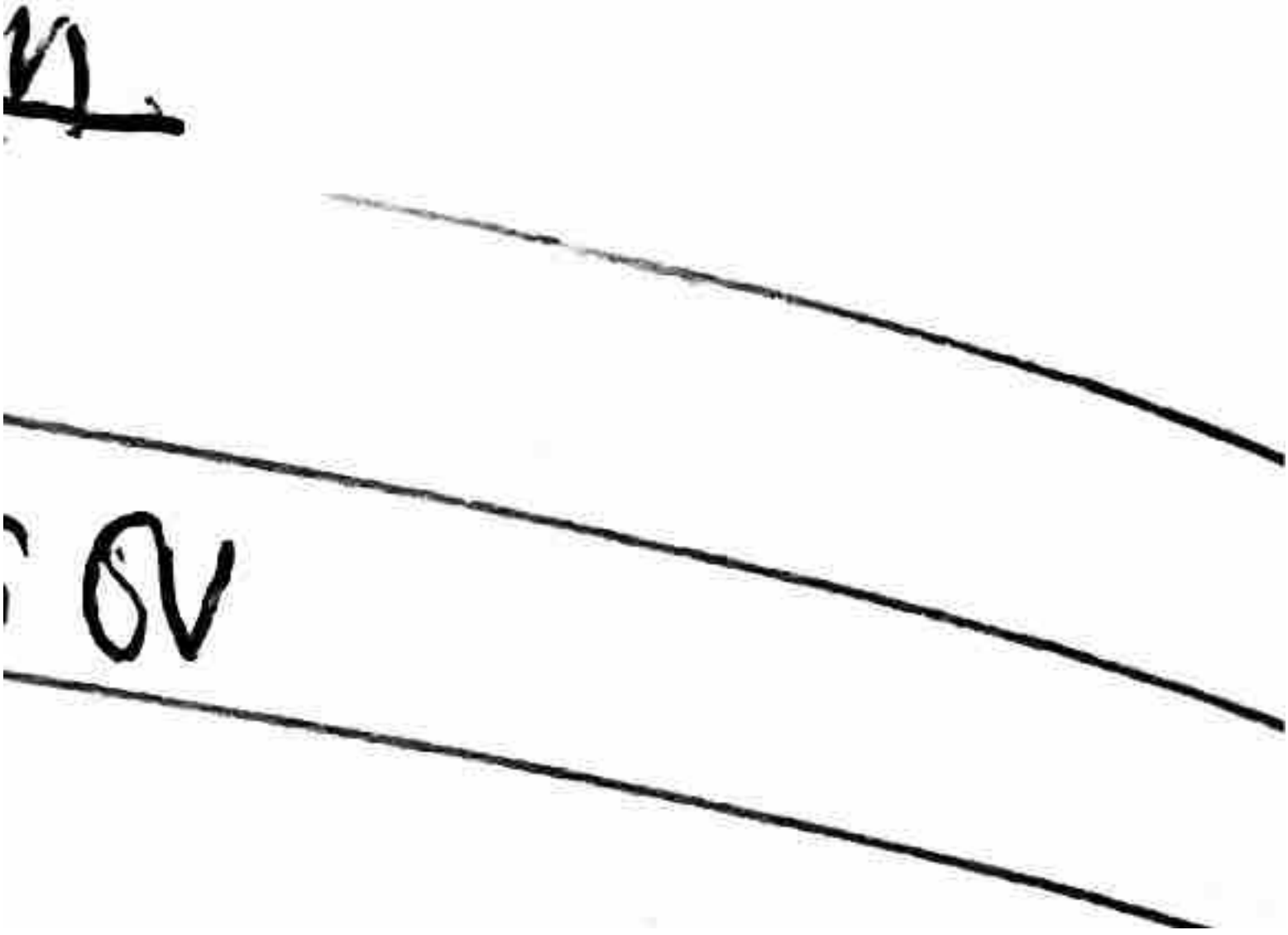


$$y = \max(0, x)$$

$x < 0 \; ; \; y = 0$

$x \geq 0 \; , \; y = x$

- Pooling: will tell where most of the information is. there are 4 types of pooling in Keras
- max
- Average
- Global max
- Global Average



Hyperparameter Optimization

manual search

~~Random search~~ Randomised search CV

Grid search CV

manual

coding your model again and again on Adam optimizer and Rms optimizer

Randomised search CV

defined param grid

```
opt = ["adam", "Rms op"
epochs = [5, 10, 15, 30, 50]
```

→ 10 possible combination

n_iter = iteration

↓

5 number of iteration

Grid search

Running each and every combination

pooling

In pooling we use the pooling methods

0	6	1	0	0
4	1	2	3	1
1	0	2	1	1
0	0	2	3	1
1	0	0	3	1

Max

6	3	1
1	3	1
1	3	1

Average

2.75	1.5	0.25
0.25	2	0.5
0.25	0.75	0.25

Flattening

6	3	1
1	3	1
1	3	1

→

6
3
1
1
3
1
1
3
1

feature detector goes over our image to detect features

we used multiple feature detectors to develop trust 2^n, 3^n

7x7 image with 3x3 feature detector, output will be 5x5

* strides are hyperparameters 1 stride is 5x5, 2 strides is 3x3 matrix

Relu

$x < 0 \quad y = 0$

$x \geq 0 \quad y = x$

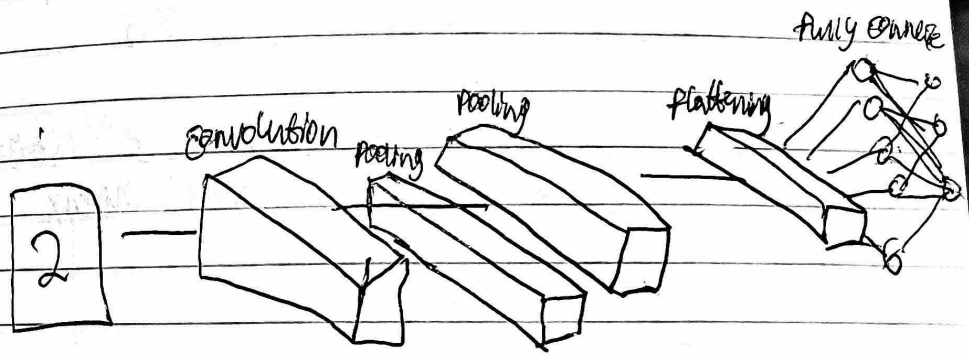
values less than 0 be come 0 with relu

Flattening

1	2	3
5	6	7
3	2	1

→

1
2
3
5
3
2
1



most of the
+ types of pooling

Convolution neural network architecture

0 - 255

0 - black

1 - white

(Kernel, filters)

Feature detector! detects features.

Complexity of data determines the epochs

28x28x1 → Gray scale

28x28x3 → Colored

↓

R, G, B

Convolutional Neural Networks: Batch Size vs Iterations vs Epochs

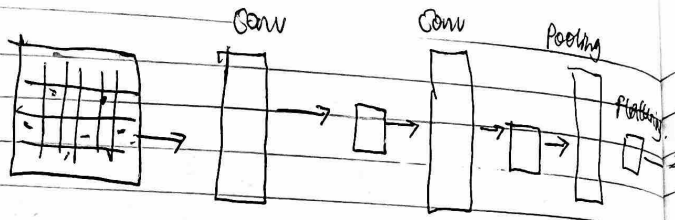
Gradient Descent: optimization algorithm in machine learning used to find the best resu

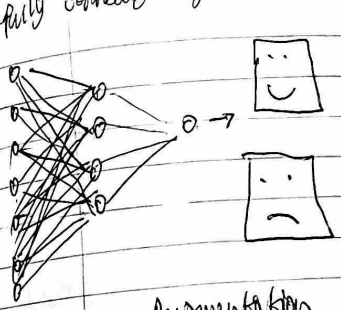
Epochs: when the entire Dataset is passed forward and backward through the neural network ~~only once~~. ~~one epoch is when the entire~~

batch: the first flow of data is a ~~batch~~ single forward and backward propagation

iteration is the number of batches required to complete one epoch

training data = 1000
batch size = 100
iteration = $\frac{\text{training data}}{\text{batch size}} = 10$





Imbalanced Dataset is the major issue while ML / DL and Data Science

Accuracy is not used for imbalances dataset

We use classification report, confusion matrix

1000 Datasets

Class A 100 ↑ 900 Class B 900 → 1800 → M1 → Accuracy

up sampling

Down sampling 100 900 ↓ 100 → M2 →

Augmentation

used to create new data from existing data

Voice data → amplitude vs time Pitch shift
Adding noise
Removing empty space
Adding volume (+db, -db)

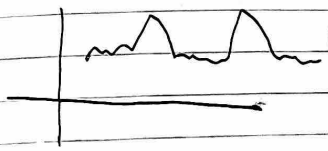


Image → Tilting (vertical, horizontal)
Adding noise
Blurring
Zooming in/out