NUST AI Course

Libraries

Seaborn, Matplotlib } Plotting

Pandas, Numpy } Data Manipulation

Scikit-learn, Tensorflow, ByTorch, xGBoost, Statsmodel } → Data prediction

To install kernel

pip install ipykernel

To install libraries

pip install numpy, pandas, scipy, seaborn

↓

libraries name.

Extensions to install in VS code

→ Text explorer UI
→ Python
→ Pylance
→ Jupyter
→ Material icon theme
→ Python extension pack

jis pe cursor lagaen ge usi information ae gi

→ Intellicode.

Exploratory Data analysis

1) Identify patterns, trends and relationships.
2) Spot errors or unusual data points.
3) Potential explanations for observations.
4) Data cleaning and transformation.

(1) Numpy

Numerical computing. Support for arrays, matrices etc.

(2) Pandas

It provides series data (one dimensional) or data frame (2-D)

1) Data cleaning
2) Data transformation
3) Handling Time series data

3) Matplotlib

Line plots, bar charts, histograms, scatter plots etc.

Graphs

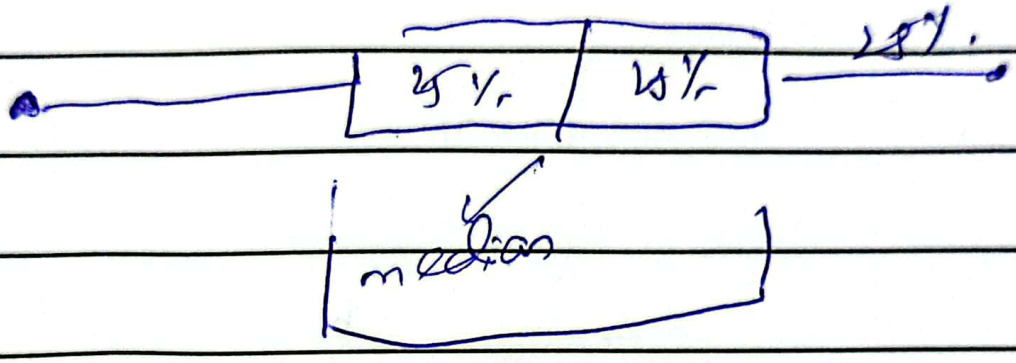
1) Histograms

→ use when data is numerical.
→ use to check normal distribution of data.

2) Box plots

Used for numerical data

2) we have a clear understanding about outliers, data range about lowest and highest and we have an idea about skewness.



3) Scatter plot

To observe and show relationship between two numeric variables. Use to find co-relational relationship.

4) Heatmap

Dark colors indicate stronger co-relation, while light colors indicate weaker co-relations.

1) Positive co-relations are usually represented by warm colors such as red or orange.

2) Negative co-relations are usually represented by cool colors, such as blue or green.

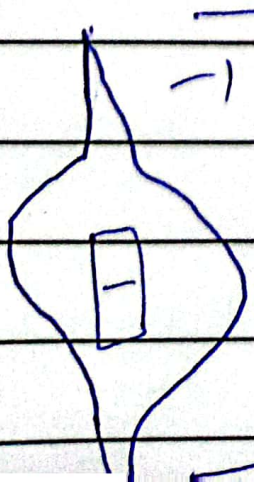
5 Pair plots

It combines histogram/KDE and scatter plots

used because others are uni-variate i.e if specific are focus, use pair plots.

6) Violin plots

used to observe distribution of numeric data.



scipy-stats

Outliers removal

1) If data is normally distributed then Z-score.
2) If data is not normal then IQR.

One hot encoding is not

1) suitable when data is high dimensional it further increases dimensions.

2) Label encoding.

3) Frequency → if many categories then we use this.
   perhaps unique data k then usi frequency.

4) Ordinal encoding

jo order hai, usi order pe encoding

Feature Engineering

→ 1) Creating new feature by combining existing feature.

2) Encoding categorical variable

3) Feature selection.

i) Filter method → filter unnecessary data.

ii) wrapper method → Forward selection or backward elimination one by one variable add kr k prediction model acha aye ge

when target is numeric, then regression problem
when target is categorial, then classification

iii) Dimensionality reduction
   → PCA, T-SNE

iv) Regularization method
   → Ridge and LASSO regression etc

v) Deep learning Based method
   / Random Forest, Decision trees

(4) Feature scaling

(5) Handling missing data

Scale sb column ko kren ge
but transformation sirf un main
jin main issue ho.

Feature selection vs feature extraction

Domain knowledge hai to
feature selection.

Filter method — (WYPI)

1) Pearson / spearman co-relation.

Pearson → numerical

spearman → ordinal, categorical

2) Between features co-relation
ni hna chahiye agar hoga
to multi-colinearity.

if more than 2 variable then 2 or t test more than 2 variables, it features are numerical

② ANOVA → analysis of variance
↳ mean testing

③ chi-square test → it features are categorical.

co-variance tells direction in which variables vary. but co-relation tells the amount between 0 and 1

=) For supervised learning we prefer that variables are independent.

=(4) variance main histogram bna k jisti variability kam hai usko consider krlein and jis main zada usko drop.

cal (R^2 = 0.34 means 34% due to x_1)

nce

wrapper method

(one by one sary include us k dekhen ge Rsq)

i) Forward selection ii) Backward elimination

iii) Recursive feature elimination
iv) Exhaustic feature selection.

Features extraction

→ distance measures (co-relation)

jo feature aek taraah ki information dety hain unko combine krdety hain. → Dimension reduction

i) (Principal component analysis PCA) when component are linearly related. means straight line.

Distance measures

i) Euclidian

ii) Manhattan

non-linear relationship

ii) T-SNE

Tree distributed stochastic neighbourhood embedding.

iii) Linear discriminant analysis LDA

linear relationship

→ Time series data

trend, seasonal, residual

irregular

≠ model → additive, multiplicative

→ perplexity means what you think about the number of close neighbors each point has. Typical value 5 to 50.

→ high dimensional ko 2D main data hai

Null- hypothesis and alternate hypothesis

TO check association between two variables.

if not associated then independent.

Scree plot

If plot is straight then importance of all is equal.

→ jahan bend ae ga, top se wahan tk tetain krnay hain.

Data Science

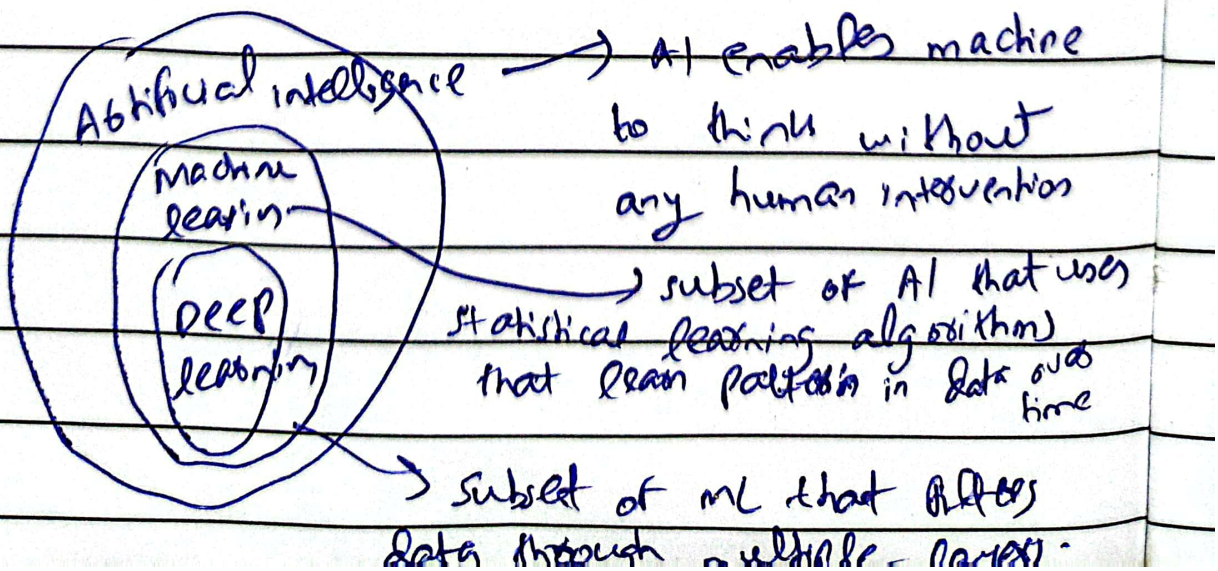
Machine learning

we human learn from experiences and then implement it for the future.

So in machine we train (learn) from data then perform prediction/ classification

→ **Machine learning**

It is a branch of artificial intelligence that focuses on building systems that can learn from and make decisions based on data.



Error → due to some natural factors

Bias → due to human errors

→ whatever kia ja sakta hai through machine

If target variable is numerical then regression and if categorical then classification.

Decision Tree and Random Forest

Random Forest is combination of decision Trees.

For machine learning tasks, we need to split data in two sets.

Training dataset and testing dataset (70,30) or (80,20)

overfitting of model

- Not detectable information because trained on noise (irrelevant data)

i.e if training and testing errors are low then good fitted.

i.e if training error is low and testing error is high then overfit.

Causes

i) complex model.
ii) Too many features
iii) Insufficient training data.

Solution

i) simplify model.
ii) Regularization (L1 shrinkage method and L2 is shrinking as well as

Hyper parameter tuning Ridge and LASSO ↗ features & domain collinearity ma hu.

feature selection method.

iii) Cross validation → 5 to folds

agar hm or test and training
ki aek bracket & thi then second
- phase main usay shuffle kren.
koi oo bracket select kren for training
and one portion for test and
so on.

Underfitting of model

Training and testing both error high.

of

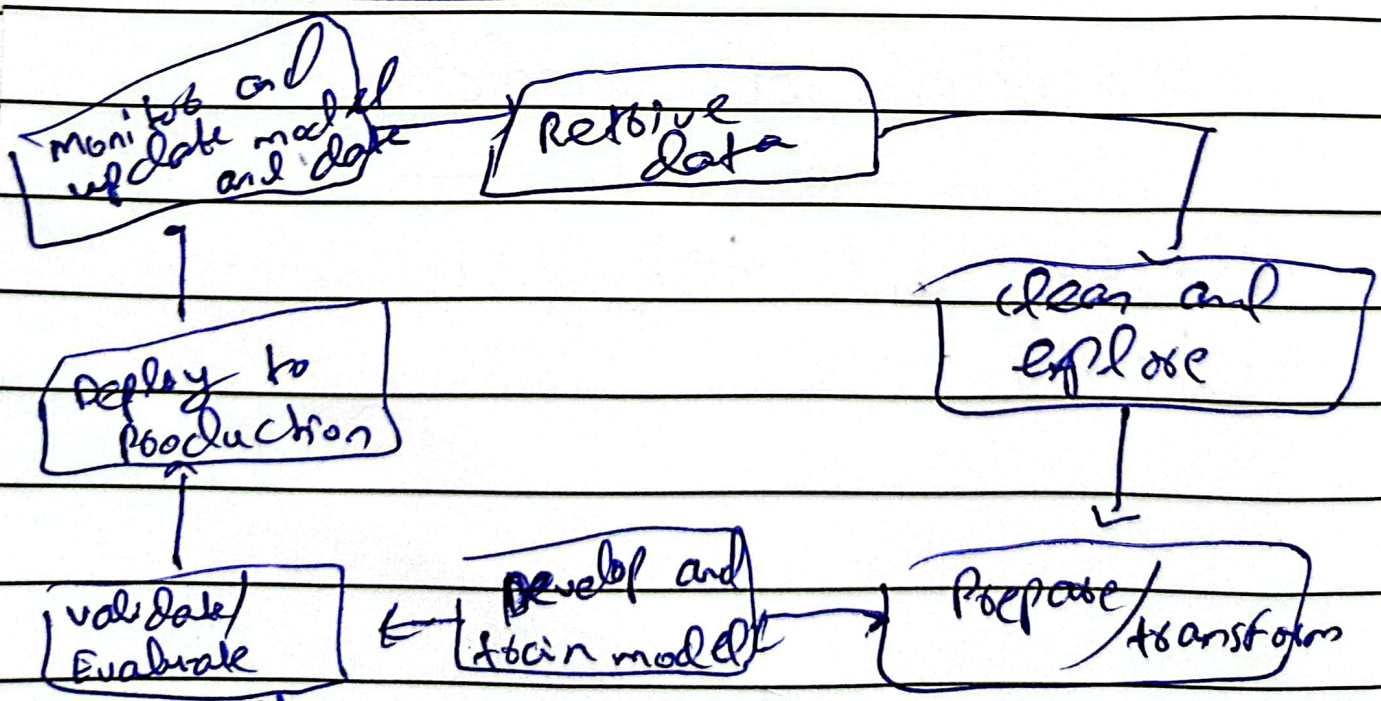
Bias - variance trade off

Bias → underfitting

variance → overfitting
(capture noise and fluctuations)

Balancing bias and variance-

workflow of ML

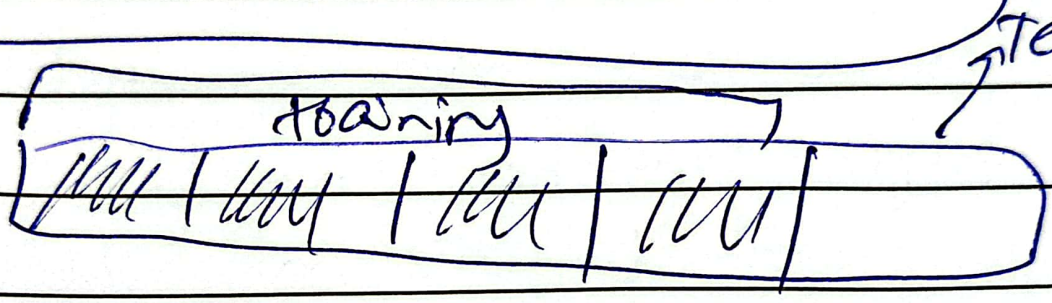


→ Model performance on training data is called calibration

→ Model performance on test data is called validation

Validation Type

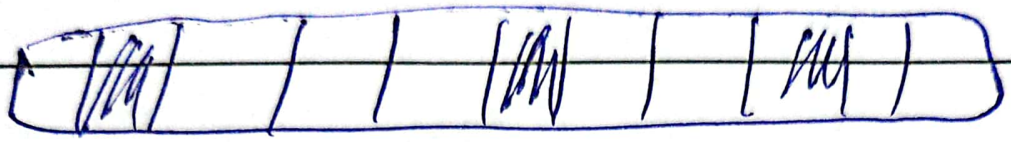
K- fold cross validation



if K is 5 then we divide data into 5 part and changing testing one part in every iteration.

Monte carlo cross validation

K- fold main hm ne fix kiya tha kitna part train and test main but is main randomly



randomly kahi per 4 pahr kahi pe 3 etc.

3) Boot-strapping → data increase
re sampling with replacement

overall data main sa randomly sample pick hotay hain.

jb ~~sample~~ tb use hain jb consistency check use ho ya data kam ho enhance krna chahen.

Evaluation measures (Regression)

(1) R^2 batata hai k model kitna acha fit hoa hai. (0 ~ 1) range.

(2) MSE } Mean square error
RMSE } Root mean square error
MAE } mean absolute error.

↙

hm ne jo prediction ki hain and jo actual data tha ~~uskay~~ jo hm ne test data main predictor ko pesh kia tha uskay sath difference ko kitna aa raha hai.

$$MSE = \frac{y - \hat{y}}{n-1}$$

y = actual value
$\hat{y}$ = predicted value.

Evaluation measures (classification)

Confusion matrix

Actual \ Predicted	0	1
0	TN	FP
1	FN	TP

Sensitivity / Recall

How much were correctly identified as positive to how much were actually active.

Specificity

how much were correctly classified as negative to how much were

actually negative.

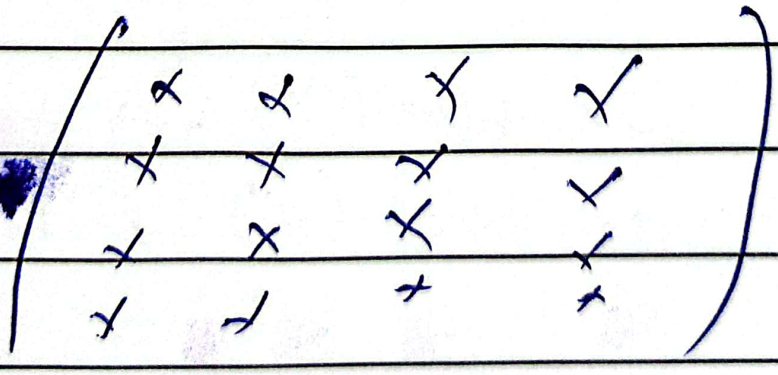
Precision

How much correctly classified as positive out of all positives

F1 score tells overall performance of binary classification model.

PCA

co variance matrix



diagonal = variances

off-diagonal = co-variances

Eigen values and eigen vectors

hum determine kartay hain ki
kitnay components sufficient
hain.
jis ka eigen value greater than
one hai us component ko
retain kartay hain.

Linear Regression

Dependent variable (Y) → target
Independent variables (X) → Features.

Simple linear regression → we have
one dependent and one indepen

Multiple linear regression

when features are more and target value is one.

High dimensional Data →

no of features > no of observation

Multiple linear regression

β_0 = intercept

β_p = slope

ε = Residuals or error.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + \varepsilon$$

intercept → agar effect ni aa raha x pe to y intercept means advertising ni ki lekin phir be kuch na kuch sales ho rae hon.

OLS

Ordinary least squares regression

minimize sum of squares of errors

Ridge regression

$$\beta = \arg\min \left(\sum y_i - \beta_0 - \sum_j \beta_i x_{ij} + \lambda \sum_j \beta_j^2 \right)$$

λ = ridge penalty for multi-collinearity

in ridge, lambda (hyper-parameter) or ridge penalty get higher, co-efficient get closer to zero mean variations lower.

// of λ

we check for optimal value/ jahan pe MSE kam ae.

if lambda = 0, then linear regression

lambda means km bias put krna

test size = 0.3 means 30% for test

random-state = 42 means seed point

OLS Regression Results

Adj R-square → irrelevant info ni leta so ye wala dekhna hai.

F-statistic check significance that model is good fitted or not

→ P(f-stats) < 0.05, these is significant contribution.

AIC, BIC should be low.
Akaike information criterion
Bayesian information criterion.

|p>|t|

if less than 0.05 it mean significant

agar values bracket k bech main aa aae, it means no significant, null hypothesis not rejected

Skewness

if 0 then data is normal

greater than zero positively skewed.

less than zero negatively skewed

Kurtosis

leptokurtic if >3 means heavy tail distribution.

if =3 then mesokurtic means normal

if <3 platykurtic means flat.

Durbin-watson Test (0-4)

Tells about auto co-relation means all observation kisi se pe depend to ni krte hai

if = 2, there is no auto co-relation
if >2 high auto-corelation
if <2 negative co-relation

Jarque-Bara

if P value is less then 0.05 then normally distributed then non-normality

[redacted]

Steps for EDA

1) Investigate each variable (check missing values and impute.)
2) If variables are quantitative then check their distribution. if categorical then label the data like label encoding, one hot encoding etc. and see their pattern by using bar plot, pie charts.
3) Scale and standardize the features for avoiding any influence of high or low magnitude.
4) After scaling and standardizing next step is transformation (if required). when we

plot histograms or pair plots. if quantitative
there is any non-normality then
we need to transform that.

5) Then plot data again to see.
6) if not normalized then apply
other transformation methods.
7) Then remove outliers and feature selection
8) Then modelling.

↗ L2 regression.

Ridge regression → errors shrinkage

Lamda/alpha (ridge penalty) add
us tay hain to re-define
coefficients. taakay predicted
actual line k qareeb aajae.

i) used mostly when large data.
ii) and when multi-collinearity
is a serious issue.

β_1 = slope coefficient

β_0 = intercept

jo unimportant features hotay hain unkay beta co-efficent ko zero k qareeb rakhtay hain.

alpha → determines the weighting to be used.

→ lambda/alpha jb zero hoga to means linear regression because β^2 bhe zero hojae ga.

Ridge main lambda penalty dalnay s bd bhe beta co-efficients kbhi exactly zero ni hotay, bluch zero k qareeb hotay hain.

Ridge cv se best alpha find krtay hain.

L1 regularization

LASSO Regression

Least absolute shrinkage and selection operator.

→ The benefit is it also provides features selection, with the help of ridge, we cannot select features.

→ LASSO shrink unimportant features co-efficients to zero.

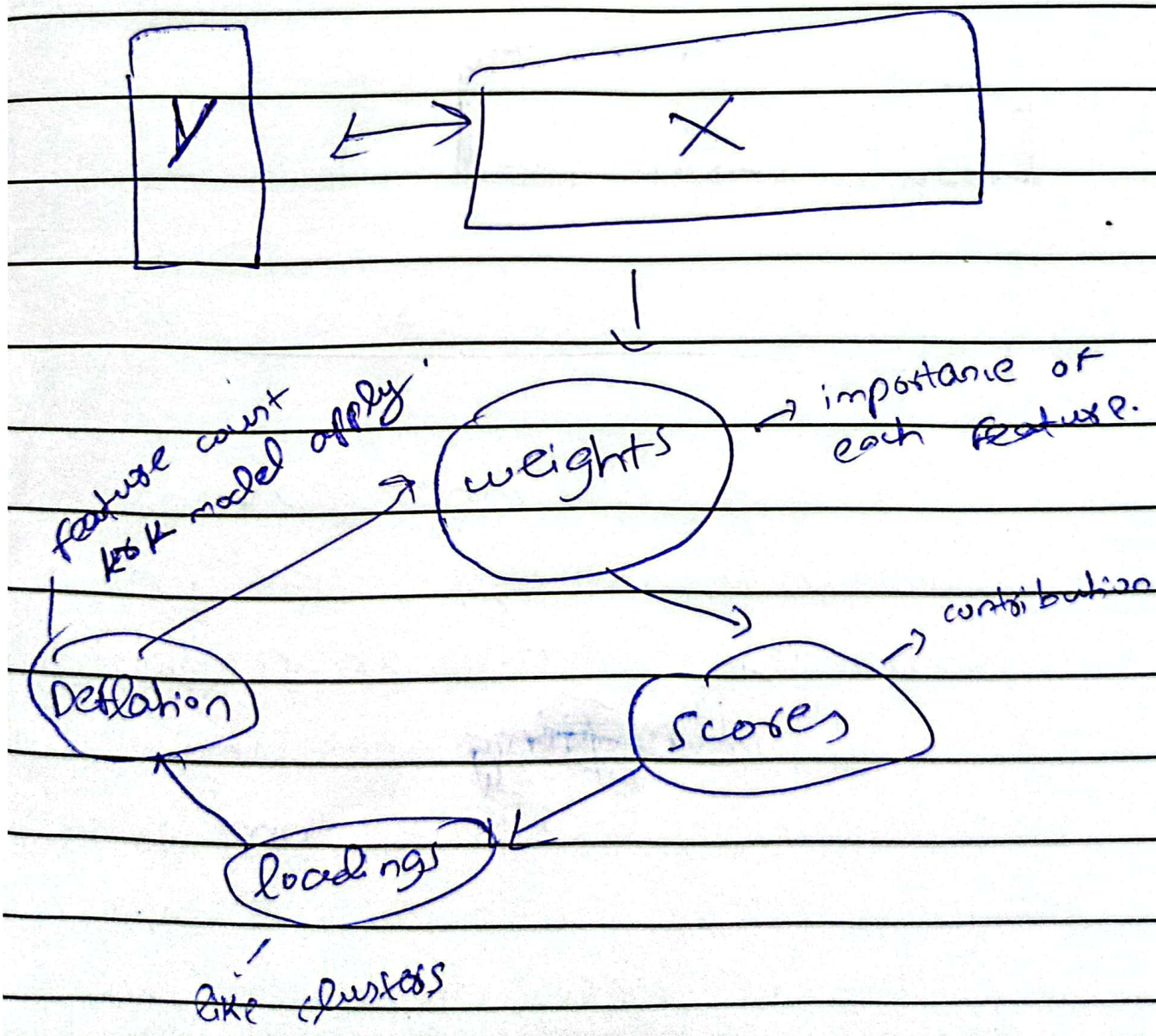
→ It also help to reduce multi-colinearity

⇒ Target variable ko scale ni krtay

⇒ unimportant features ko jb hm drop krdetay hain to automatically ~~feature select~~ dimension reduction hojati hai.

Partial least squares regression (PLS)

purpose same as LASSO, Ridge, but it is used for high dimensional datasets.



pta chal jata hai kitnay components
koo MSE kam hai. means utnay
components pe maximum information
mil rae hai.

-1 PLS generates principal components, hence reduces dimensions.

Classification

1) Logistic regression

supervised machine learning algorithm widely used for binary classification tasks. such as identifying whether an email is spam or not, presence and absence of diseases, but also be used for multi-class.

→ It can be used for image classification.

Linear discrimate Analysis

→ Classification algorithm and also reduces dimensions.

→ supervised

→ PCA reduces dimensions on the basis of linear co-relations.

→ LDA reduces dimensions on the basis of class labels — ~~co-relations both~~.

→ LDA is for that type of tasks whose linear relationship exists means jis ko straight line se separate kr sken. For non-linear QDA/RDA is used.

PCA pre-processing mein apply hota hai and classification ki data and unsupervised learning mein ata hai. LDA is supervised.

→ ROC curve tells us about accuracy.

→ Suitable for low-dimensional data.

Support vector Machine SVM

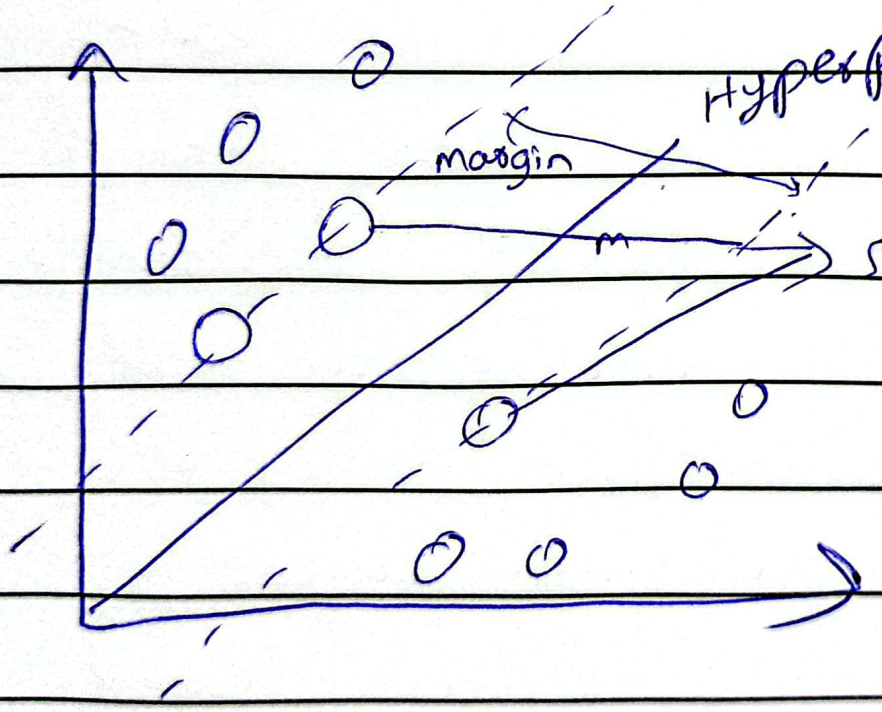
→ classification algorithm.

→ SVR for regression.

→ Low dimensional data ko project krta hai high-dimensional space pe

Kernel trick se non-linear decision boundaries ko handle kartay hain.

- Binary and multi-class classification dono ko deal karta hai.



Common Kernel Trick

Linear, polynomial, radial basis function (RBF) and sigmoid.

Decision Function

→ Classifies new data points based on their position relative to hyperplane.

KNN → Supervised learning

- stores all available data and classifies new data point based on similarity. It can be used for regression as well as classification but mostly it is used for classification purpose.
- It is also called lazy learner algorithm because it does not learn from training set immediately
- instead it stores the dataset and at the time of classification, it performs an action on dataset

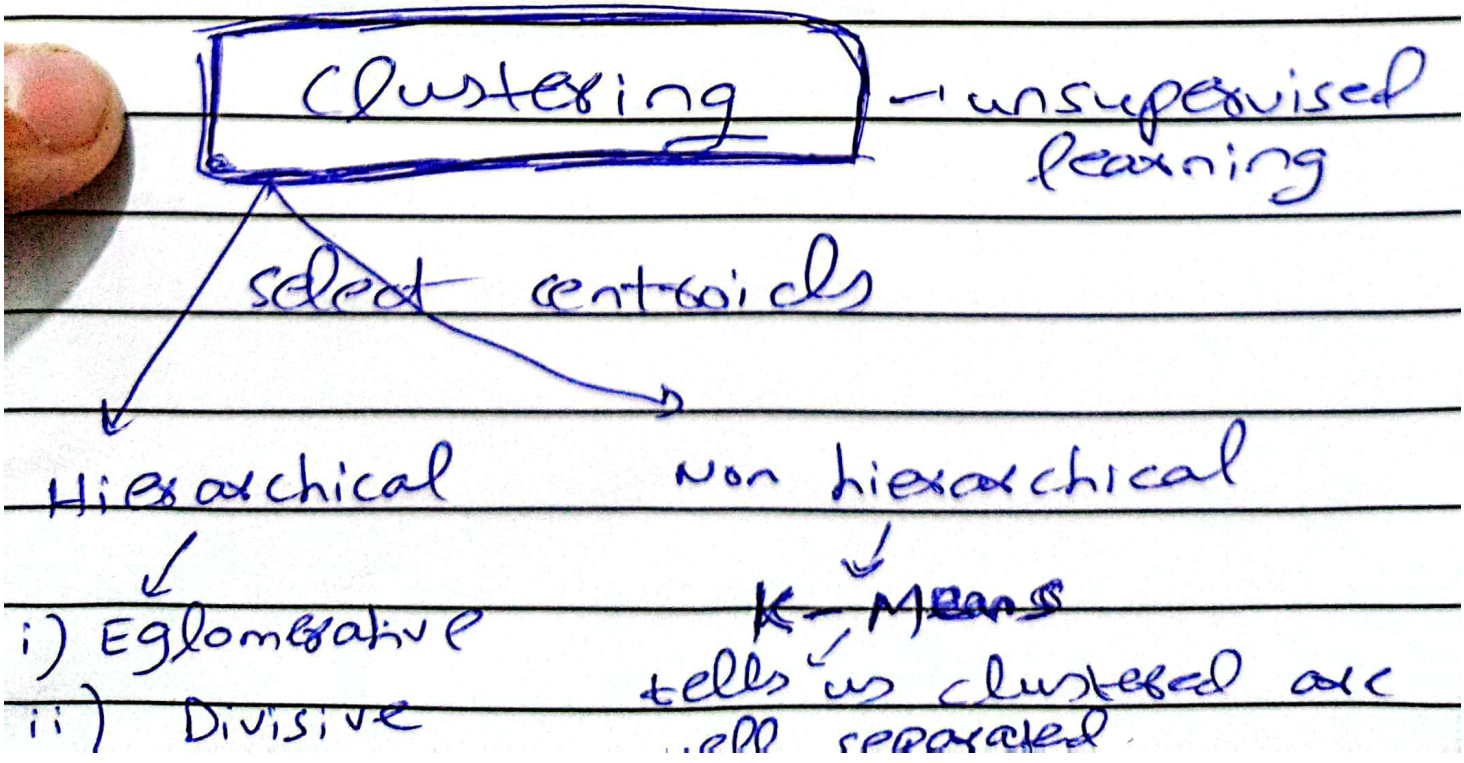
Step 1 Select the number K of neighbours

Step 2 Calculate Euclidean distance of K-number of neighbours

Step 3 Take the K-nearest neighbour as per calculated Euclidean distance.

Step 4 Among these K-neighbours, count number of data points in each category.

Step 5 Assign new data points to category for which the number of neighbours is maximum.



Dendogram

Hierarchical

Divisive

Top to bottom study

Agglomerative

Bottom to top study

Neural Networks → Deep learning

both for regression and classification

→ inspired by human brain

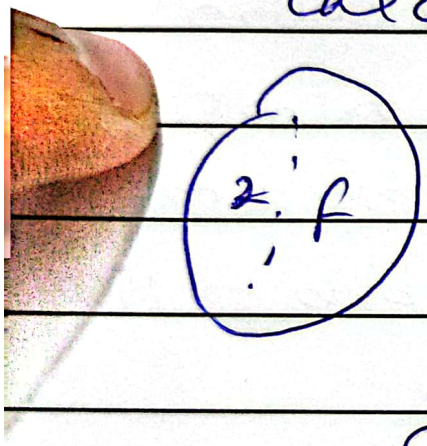
TPU's → Tensor processing unit advanced form of GPU.

RNN → when time oriented image data

CNN → image datasets, video datasets, classification tasks.

→ Deep learning model pre-processing usual workflow hain.

→ Perceptron main probabilities calculate.



sum of probabilities se decide

Sigmoid and ReLU most popular activation functions.

Activation function in context of neural networks is a mathematical function applied to output of neuron. The purpose of activation function is to introduce non-linearity into

the model. allowing it to learn and represent complex patterns in data. without non linearity, Neural network will behave like a linear regression model, regardless of number of layers.

→ The activation function decides whether a neuron should be activated or not by calculating weighted sum and further adding bias to it.

→ In neural network, we would update the weights and biases of neurons based of the error at the output. This is known as back-propagation Activation functions make back-propagation possible.

→ Collab or kaggle notebook for free if our RAM is low

→ Base theorm.

Public dataset

i) Image → Microsoft coco
ii) Text → Squad
iii) Video → ~~Youtube~~ Youtube 8M
iv) Audio → Google Audioset.

Libraries

i) Sklearn → Machine learning
ii) Tensor Flow → Google.
iii) Pytorch → facebook 2016
iv) Keras → 2019 most popular.

Existing architecture

i) Image classification → Resnet.
ii) Image segmentation → Unet
iii) Text classification → BeRT (Transforma
iv) Image transformation → Pix 2 pix
v) object detection → YoLO
vi) Speech generation → wavenet.

→ Radial kernel → if polynomial structure
→ RBF → For complex tasks.

→ Designed for quantitative datasets but can handle image datasets
→ Need large data to train well.

Date

Hierarchical

Divisive

Top to bottom study

Agglomerative

Bottom to top study

Convolutional Neural Networks (CNN)

computer vision algorithm.

Computer vision ke mostly tasks main use hota hai CNN,

→ Image recognition.
→ object detection.
→ Transfer learning.

Multiclass → Softmax.
Binary class → Sigmoid.

Loss - entropy

ek node se kitni information aa rahi hai and vo kitna important hai

overcome computational complexity

Date

1D main usually ki structure ni path like ANN.

2D main data le leta hai.

→ ~~RCNN~~

→ RCNN mostly use usay hain

Hamstash

i) padding layer.

ii) Filters → edge detection (3×3) mostly

iii) pooling layers.

max, Min, average.

$(n-m+1)\ (n-m+1)$ 6×6 - original image

$(6-3+1)\ (6-3+1)$ n, m → filters = 3

4 × 4

so filters se dimension kam ho rahe hain.

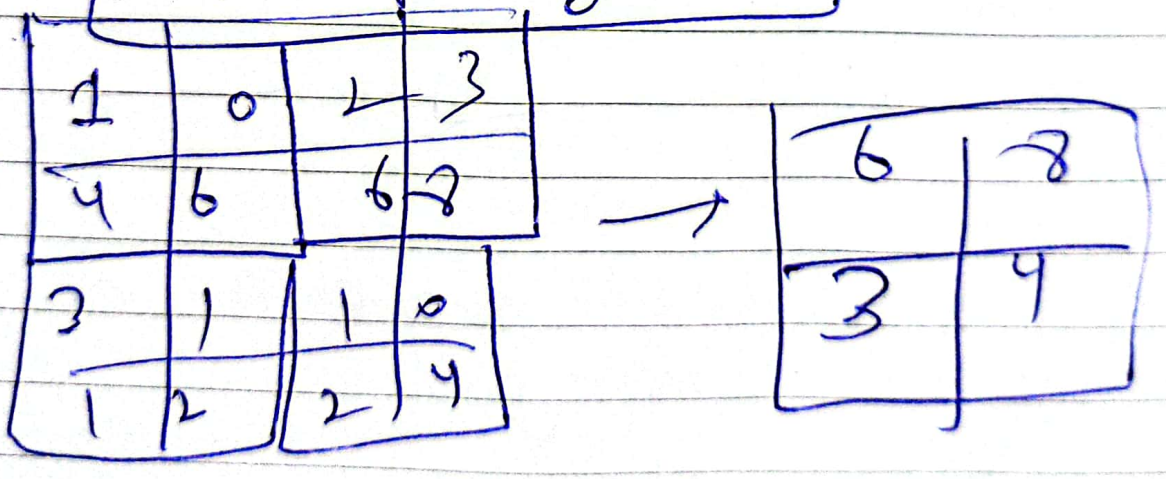
Date

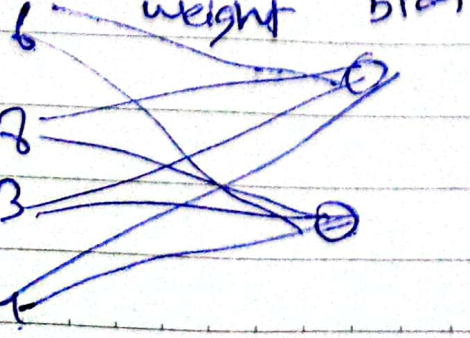
side wali values sirf one time ahi hai so usko cater karnay k liye extra layer lgaey hai that is known as padding layer pooling main usay hain.

Strides :- Determine how many squares or pixels our filter skip when they move across the image.

move kitna krna hai, jump kitna krna hai.

max pooling





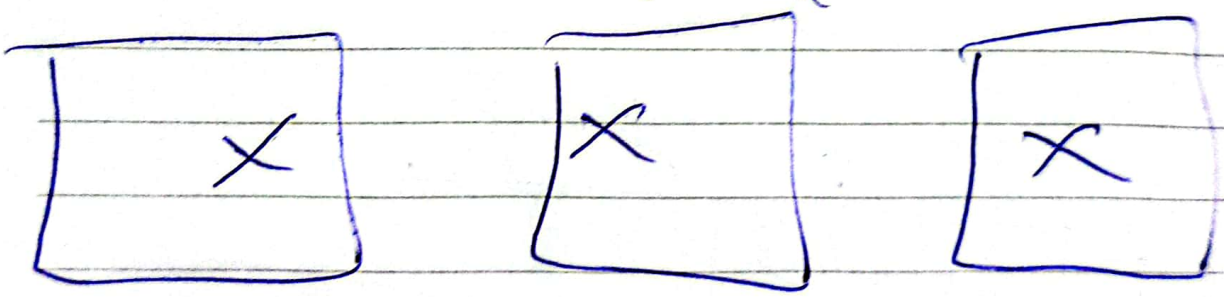
Date

CNN

four layered concepts.

1) Convolution → Filters. Feature map.
2) ReLU → Activation function.
3) Pooling.
4) Fully connected layer.

Pooling layers add because not on same grid



→ MNIST dataset available in Python library.

→ Automatic driving cars
→ Facial recognition

Hidden in ANN

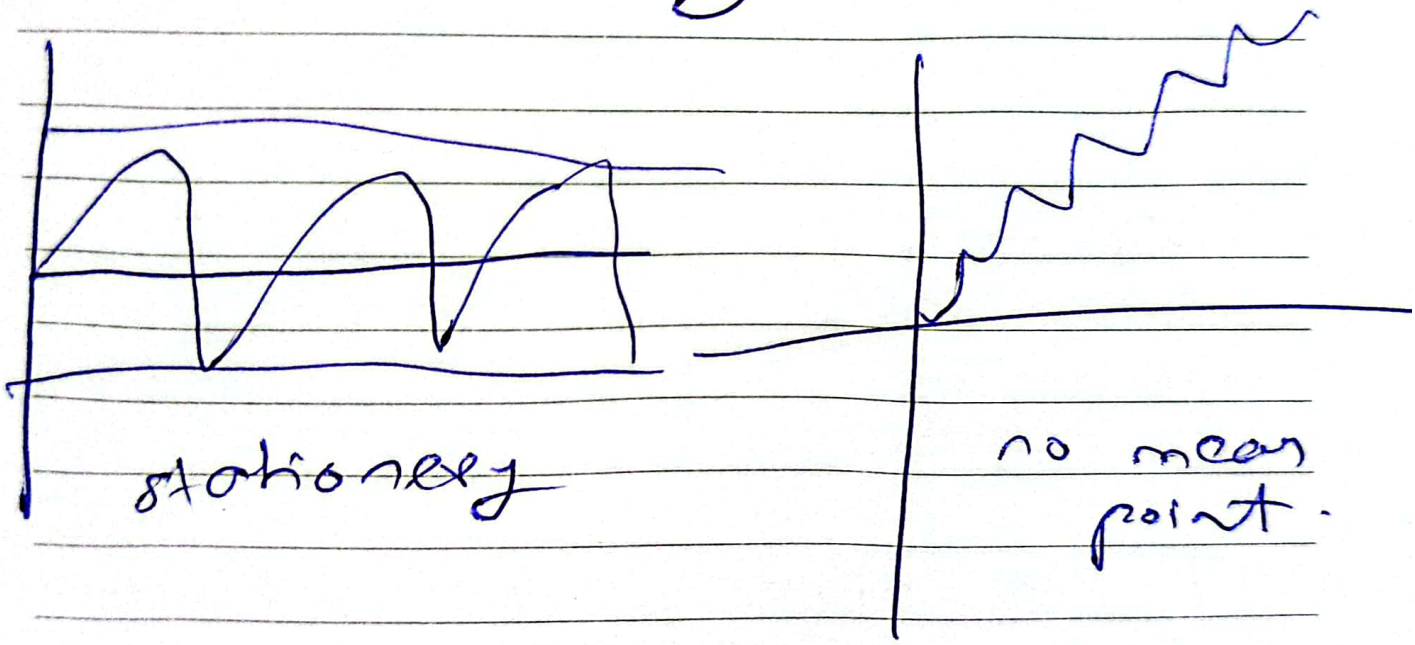
Date

Dense → Filter layers in CNN

Dropout → To reduce dimension.

Time-series analysis.

For TSA, we need to make data stationery.



Trend, seasonality, cycle, irregular in sb ko remove kro 'k aagay chalna hai means data stationery.

interpolation and extrapolation.

within existing values

TSA main line plot e bnatay hain.

Date

Date ko stationary karny k liye smoothing techniques.

→ Holt's linear → use when data have trend.

→ Holt's winters seasonal → when both trend and seasonal.

AR → p
I → d
MA + q

10 seasons ka.

Recurrent Neural network

Time series data both image and numeric.

when data is sequential.

Gmail main be RNN work kr rha hai

lstm e use krty hain.