Applied Machine learning  Date: 29 08

MACHINE LEARNING: putting human capabilities in machine

what are human capabilities?

↳ senses, speech, learning. Humans start learning from day 1.
learn from the environment

* Actionable learning

- Machines, like humans also learn from the environment
- Learning is not just knowledge gain, but also improvement in performance.
- Different context has different learning. If the task is different, the learning should be different
- learning sahi jaye jese jese, wese performance improve hoti jaye — task pe learning, performance improvement
- There must be a program capable of learning
- ML → set of algorithms (learning algo) used on different data
- Learning algo is a high level function which builds another function ⟹ this is a hypothesis

  eg. learning algo : k-means, decision tree, KNN
- Every ~~learn~~ algo has different learning rules but we can make one hypothesis out of all
- Performance Evaluation is of the hypothesis — how strong is your hypothesis

  ↳ How?

  Initially jo knowledge thi usi se train kiya aur ab usi se testing karaingay

Teacher's Signature _____

Date: D D M M Y Y

- Highest accuracy doesn't mean good hypothesis
- Good hypothesis is generalized

generalized vs specialized

↓ learning

↓ fix, particular, ratta

(The model must be good on unseen data)

- Model shouldn't be 100% accurate.

⇒ Robustness vs Generalized vs Specialized:

Robustness:

- able to handle anomalies, har situation pe sahi ho
- har environment pe sahi chalay e.g. winter pe train kiya aur summer pe fail ho jayen
- testing error has to be consistent with the training error
- depends on data
- Once you build the model, now it's fixed. Now you change the data
- distribution change in data but hypothesis remains good
- if the result is consistent over every type of distribution, it is the model's robustness.

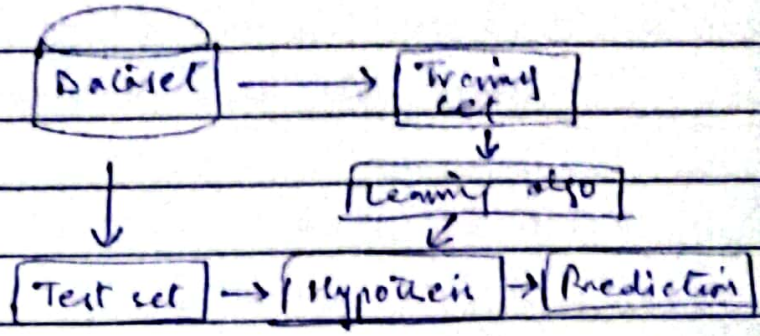
CamScanner

Date: D D M M Y Y

⇒ Types of Machine Learning:

1) Supervised
   - Regression
   - classification
2) Unsupervised
3) Reinforcement

Pipeline:




One-class classification problem:
↳ anomalies detection

- Recommender Systems can be supervised and unsupervised
- if real-value input, then it is a regression problem
- Reinforcement ; Feedback learning
  ↳ rewards

→ Prediction vs Interpretation:

- it is always a trade-off
  ↳ if the goal is high accuracy (e.g. spam filter) then we do not care, why and how the model reaches it
  e.g. future sales (Prediction)
  ↳ if the goal is interpretability (e.g. in bio, SNPs which predict a certain disease risk) then we care why & how
  - reasoning (why this disease occured) (Interpretation)

Teacher's Signature _____

Date: D D M M Y Y

→ Supervised learning:
- model is trained on labeled data to make predictions or classify new data
- eg. spam filtering (spam, non-spam) — new email messages will then be classified as spam or non-spam

→ Unsupervised learning:
- model is trained on unlabeled data to find patterns or relationships in the data
- eg. fraud detection, social networks

→ Reinforcement learning:
- model is trained by learning from the consequences of its actions in an environment
- eg: robotics, gaming, autonomous vehicles.

* Classification/Regression problem:
↳ Accuracy high, but precision, recall low — reason?
* Most important problem is data, without data, we cannot perform.

foundation of ML problem

comprises training examples used to teach models

has features and ~~targets~~ labels

* features (input variables), ~~th~~ labels (target/output variables)

Teacher's Signature _____

HUSSAIN

Date: D D M M Y Y

⇒ Datasets :

→ Kaggle Dataset
- good resource
- published on some paper
- 2024 main upcoming problems are the hot topics for ML problems.

→ UCI ML Repository
- derived from some published paper

MNIST dataset → benchmark dataset (standard dataset)
↳ you can claim that you have the best algorithm

→ DDJB (DNA Databank of Japan)
EBI (European Bioinformatics)
DNA is an information carrier. Eye color, hair color
if we unbind DNA chain, it's way too lengthy

⇒ Model
↳ mathematical representation of a system that learns patterns from data
↳ decision tree, neural network, support vector machine etc

⇒ Training :
↳ process of teaching a model to make predictions/decisions by learning patterns & relationships from data

Teacher's Signature _____

Date: D D M M Y Y

⇒ Testing / Evaluation :
↳ assessing a model's performance on unseen data to gauge its predictive accuracy.

⇒ Evaluation Metrics:

→ classification (predict category):
- Accuracy
- Precision
- Recall
- F Measure
- F beta Measure
- ROC (and AUC)

→ Regression (predict value):
- Sum of Squared Error
- Mean Absolute Error
- RMSE

Multiclass classification → can be multiple outcomes
eg: ek patient ko heart disease hai ho aur sugar bhi high ho

| | regression | classification |
|---|---|---|
| Multi-regression → | 2.5, 7.6 | A, B |

→ ROC curve : Analyzing the strength / predictive power of a classifier
The closer your model is to top-left, the better it is
In marketing, FP rate is not good at all

$$\sum_{i=1}^{k} (y_i' - y_i)^2 \rightarrow \text{Sum of Squared Error}$$

Teacher's Signature _____

HUSSAIN
CamScanner

Date: D D M M Y Y

$$\sqrt{\frac{\sum_{i=1}^{k} (y_i' - y_i)^2}{k}}$$ → Mean Absolute Error

* if error is large, we use root — wrong range main kaunsay k liye

⇒ Confusion Matrix :
- for classification
- to evaluate model (whether model is as we want), performance check
- diagonal should be maximum

→ Accuracy : accuracy of correct prediction / quality of your model
↳ correct predictions out of all predictions
* Model is a function → dependent / independent variables

→ Recall : concerned with FN
↳ should be high
↳ here its costly than FP
↳ it shouldn't be present

$$R = \frac{TP}{TP + FN}$$
↳ minimize

→ Precision : concerned with FP
↳ costly than FN

$$P = \frac{TP}{TP + FP}$$

* High recall, high precision → trade-off

Teacher's Signature _____

Date: D D M M Y Y

* Which system is better depends on the problem

→ Fbeta Measures:

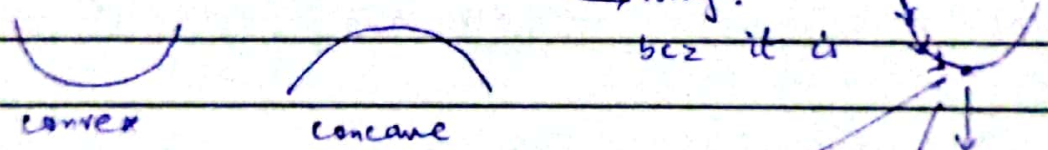
$\beta = 0.5$,

1) $\frac{(1+0.25)(0.7\times 0.6)}{0.25(0.7+0.6)} = \frac{(1.25)(0.42)}{(0.25)(1.3)} = \frac{0.525}{0.325} = 1.6$

2) $\frac{(1.25)(0.8\times 0.5)}{0.25(0.8+0.5)} = \frac{(1.25)(0.4)}{(0.25)(1.3)} = \frac{0.5}{0.325} = 1.5$

Higher accuracy, higher recall, higher precision ⇒ good model

→ In regression problem, we use RMSE → why? bcz it is




error function plot

only one global min

error will reduce, then increase

* Learning rate (LR)

increase LR ← left ← min → right → decrease LR

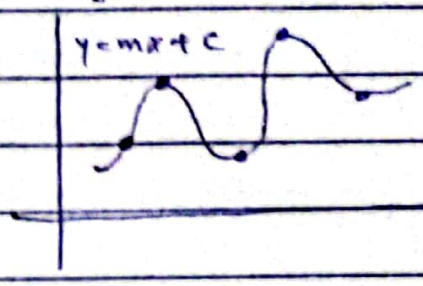
Teacher's Signature _____

Date: D D M M Y Y

→ RMSE :

- we can easily locate global minima
  convergence → model is generalized
  ↳ na overfit, na underfit
  ↳ error k minimum pe

if plot is this (model)




* this was test error

* train errors :
  ↳ using redistribution
    ↳ training samples again

it means error is 0.
this is overfitting

⇒ Cross validation :

- to create different variants for train and test
- Methods :
  - Redistribution (N; N) — training error
  - Holdout (70 train, 30 test) (2N/3 ; N/3)
  - k-fold cross-validation (N − N/k ; N/k)
  - leave-one out

→ underfitting pe error nahi ayega
  ↳ biased model (data ko bohot kam dekhna) → biasness — not looking the patterns in the data

in context of model

→ overfitting
  ↳ data ko bohot dekhna
  ↳ model ko bohot learn kawa diya

Teacher's Signature _____

Date: D D M M Y Y

* Highly biased → underfitting
* Highly variant → overfitting

- bias and variance ~~and~~ are two errors :

[ Training accuracy = 30% ]
[ Testing accuracy = 20% ] underfitting
sign of underfitting

→ model is highly biased

- model complexity : no. of layers, parameters, features, samples,
↳ it can also be low due to less iterations

- Data main masla hosakta hai — 30 images augmented to 30,000 → too many outliers

- Overfitting : you learn the wrong things too
↳ too many outliers — wrong data
  - High learning — overfitting
  - No learning — underfitting
  - low bias, low variance — good balance

* Good model : low bias, low variance, high accuracy, high recall, high precision

Teacher's Signature

Date: D D

- leave-one out (99:1) N = k
- k-fold (90:10)
- redistribution
  - ↳ parameter tuning
- hold out
  - ↳ not reliability
  - ↳ ~~not~~ randomly / manually disjoint
    - ↳ X

LAB Date: D D M M Y Y 3

⇒ Exploratory Data Analysis
↳ understand data

Industry data is in raw form. we have perform some steps

EDA
↳ 1) Analysis
2) Feature Engineering

→ Analysis:

Dimensions, values, samples, distribution, skewness, outlier or no

→ Feature Engineering:
(Transformation steps
↳ drop/keep features based on their contribution and use
- dropping useless features
- Filter & Wrapper ⇒ approaches of feature selection
- Wrapper method is costly and timeconsuming than Filter method
- Filter method is ~~less~~ less costly and less time is required
- Wrapper is trustworthy
- Filter: model se koi talluq nahi hoga.
↳ independent/dependent variables ka correlation/variance
- Wrapper: Randomly subset uthaya aur usey model pe train kiya. Is tarah iterations chalayin different subsets pe aur usi main se best ko select karlay hain

* subset of features
* objective is accuracy

* Entropy: measure of uncertainty
eg: Gender and Salary
M ↙ ↘ F → 'usse salary uncertain hai
50% 50%

Teacher's Signature _____

Date: D D M M Y Y

- Agar saaray mile honge toh ~~sahar~~ certain hoga.
- 50/50 pe decision lena mushkil hoga

This is feature selection based on Filter method.

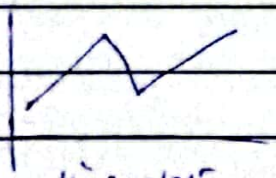
→ Why EDA ?

- detection of mistakes in the data
- Tools : ~~pandas~~ plots, graphs, summary of statistics
- Libraries : pandas, numpy, seaborn, matplot
- Types of analysis

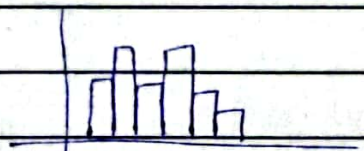
non-graphical (statistical) — univariant / multivariate

graphical — univariate / multivariate

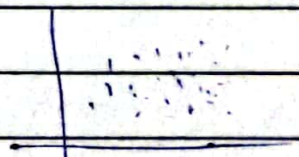
univariate → single variable



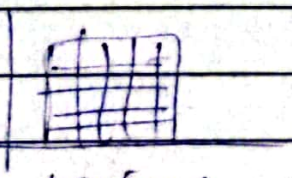
line plots



histogram



scatter plots



heatmap

contour plots → interpretation is important

[saddle point]

concave aur convex saath banta hai

global minima aur global maxima

ek jagah pe bana rahta hotaa hai

Saddle point aanay se pehle 'early stopping' karo

forcely stop kra k saddle point pe na chala jaye

Teacher's Signature _____

Date: D D M M Y Y

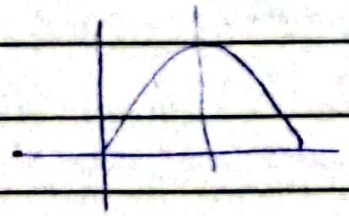
- patience ↙ patience = 5
  e.g. agar 5 iterations tak koi change nahi aaraha tou stop kardo.
- side-by-side box plot
  ↳ comparison of multiple groups

→ Categorical data
  One-hot encoding → convert categorical to numerical
  ↳ iss data pe learning algorithm kaam nahi karta
- feature matrix

GNN

⇒ Quantitative Analysis:
  central tendency → mean, median, mode

→ Symmetric / Asymmetric Distributions
  ↓
  many distributions
  ↳ distribution is a function



symmetric

→ PMF (Categorical)
→ PDF (Continuous)

→ Pareto Principle (80-20 rule)
  eg: 20% fault, 80% loss nor fault ki wajah se small fault, but more loss

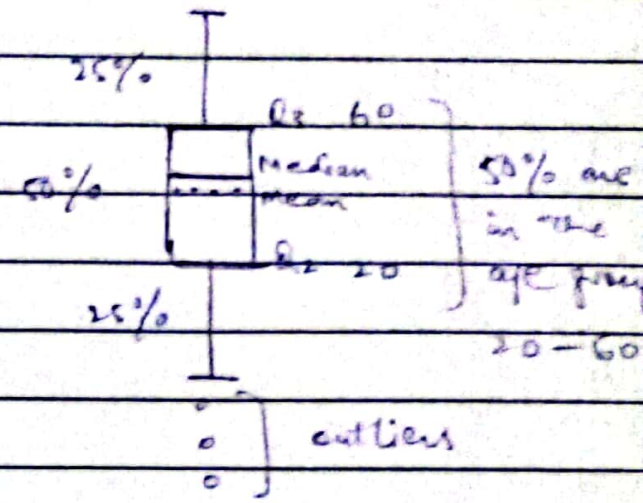
HUSSAIN

CamScanner

Box plot

Date: D D M M Y Y

- NaN → garbage / null




—— Median

- - - - - Mean

outside 25% —— outliers

Lab Activities : all Tasks — Monday

- black : kaunay nain
- red : important
- blue : s

with interpretation

EDA – continued

Date: 10 / 10 / 21 / 2 / 3

→ Spread :

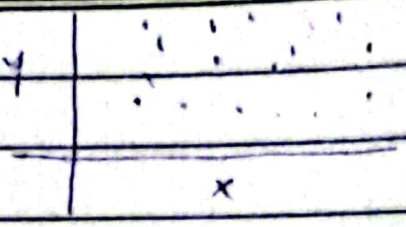
↳ measure the spread of the data

→ measured by variance and SD

- mean / median / mode → central tendency

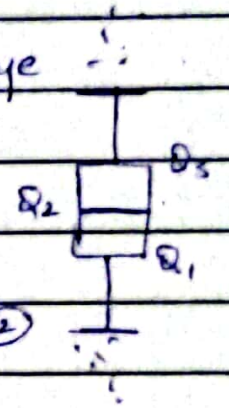
↳ robust measure

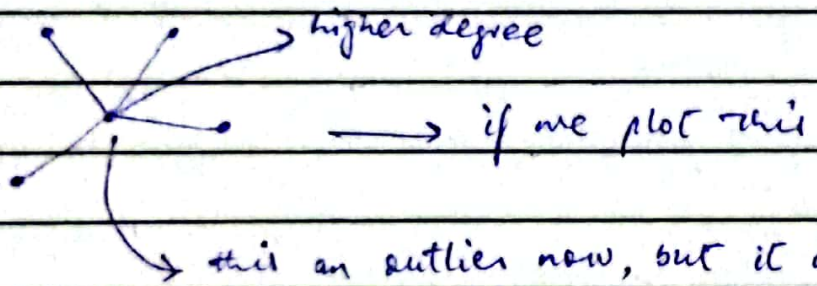
↳ data change bhi hojaye tou inconsistency na aye



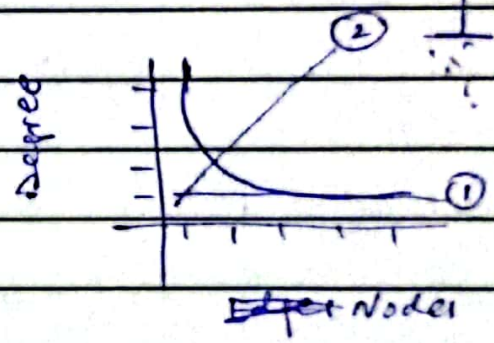

High Spread

- median is very important in central tendency
- spread ziada hogi tou learning acchi nahi hogi
- e.g. graph




→ if we plot this




* agar main node pe hypothesis build karaniyay tou (2) hogi lekin uss main error k chances ziada hain
* agar baki node ki basis pe hypothesis build karaniyay tou (1) hoga

Solutions :

↳ don't take linear hypothesis — make it ~~curt~~ curvy linear using polynomial function

- we usually drop outliers but here we can't

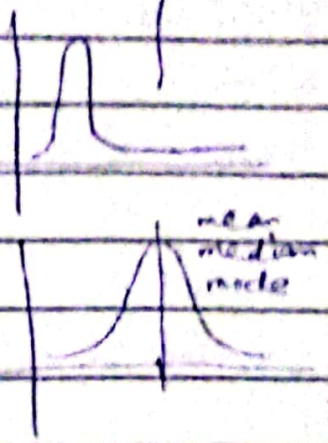
Teacher's Signature _____

this is left skewed

Date: D D M M Y Y

→ log - Transformation

we cannot establish a relation here




- we apply log transformation to balance/reduce skewness — make less skewed
- Skewness main relation pata nahi chalta
- Koi variable hoga which is highly skewed
- We cannot drop that variable

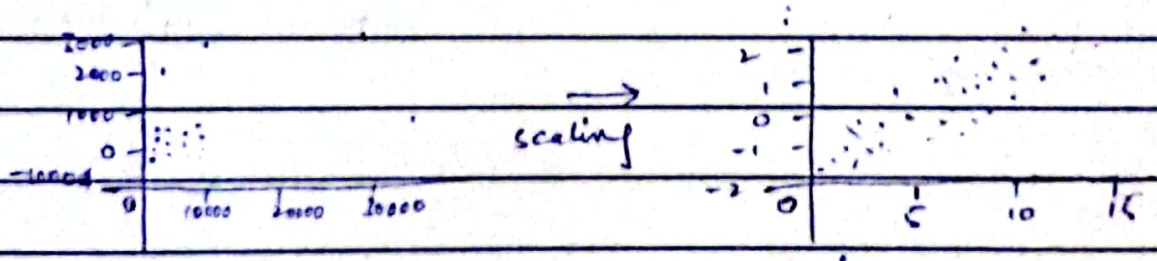
this is the normal trend

Gaussian Distribution/ Normal Distribution

skewed = 0

- We do this by scaling labels

Scale Invariant → pooling (CNN)




here we can build a relationship

* this is a personalized advertisement problem

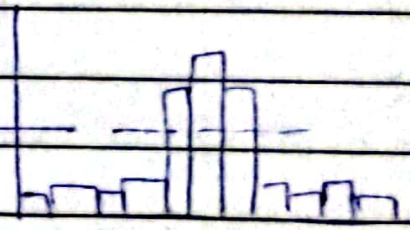
close to 0 → normal distribution

Teacher's Signature

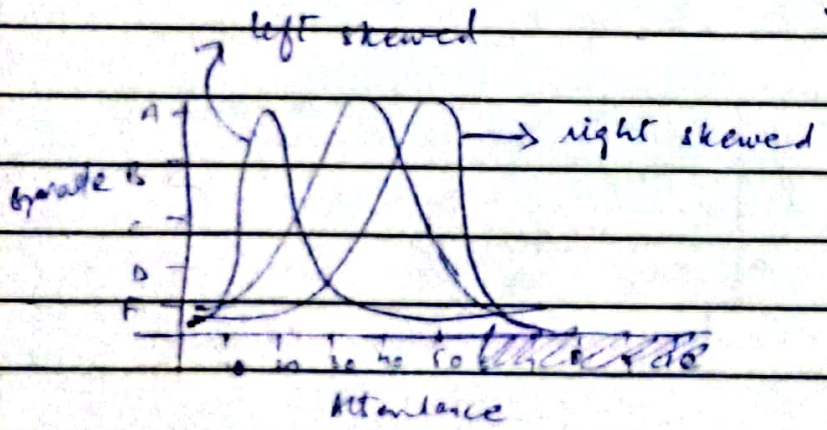
Date: D D M M Y Y

- If we ignore the outliers, it will affect the learning so we cannot ignore them
- Boxplot is one of the best way to discover the outliers

→ Kurtosis : Measure of the peakness
- There can be abnormalities in Gaussian distribution too.
- Bohot ziada peak is a positive kurtosis
- Bohot kam peak is a negative kurtosis
- Those abnormalities can deviate the learning function
- Log transformation can be applied here
- Another way is to drop the values.



yahan tak drop kardo




Teacher's Signature _____

functions → Prediction
functions → Data Compression

# DECISION TREES

Date: D D M M Y Y

- For classification problems, decision tree is the best
- it is a learning algorithm
- DT is a non-linear ML technique
- split from root to leaf
- no. of paths from root to leaf are finite — finite rules to recognize your samples
- test on attributes
- if samples are identical in terms of their attributes, but different labels — how do we classify them ?
  → we drop them → this is a limitation (outlier)
- one tree is a hypothesis
- Big tree OR Small tree ? (independent on the data)
- there may be multiple trees for one data, but one hypothesis will be the best one
  → overfit
  nodes are test cases on attributes
  ↓
  column
  saaray nodes ko test karaingay tou overfit hojayega aur kam nodes ko karaingay tou underfit

→ Data Compression:
  → filter or wrapper
- we do each feature's analysis → filter method
- we can call it hybrid too

Teacher's Signature _____

*calculator for mid

Date: D D M M Y Y

- Each branch corresponds to classification rule
- How is it like if its a regression problem?

* Disadvantage of DT:
- if a problem is ~~linearly~~ linearly solvable, don't go for DT

→ Algorithms:
- kis node ko top pe rakhna hai

* entropy — b/w 0 and 1 (measure of impurities)
pure class has entropy 0

→ Computing GINI

$\leq 97$: $1 - \left(\frac{3}{6}\right)^2 - \left(\frac{3}{6}\right)^2$

YES NO

Total classes = 6
Yes classes = 3
No classes = 3

$= 1 - 0.25 - 0.25 = 0.5$

$> 97$: $1 - \left(\frac{0}{4}\right)^2 - \left(\frac{4}{4}\right)^2$

$= 1 - 0 - 1 = \cancel{-1}\ 0$

$\frac{6}{10} \times 0.5 + \frac{4}{10} \times 0 = 0.3$

Total no. of samples $(6+4)$

Teacher's Signature _____