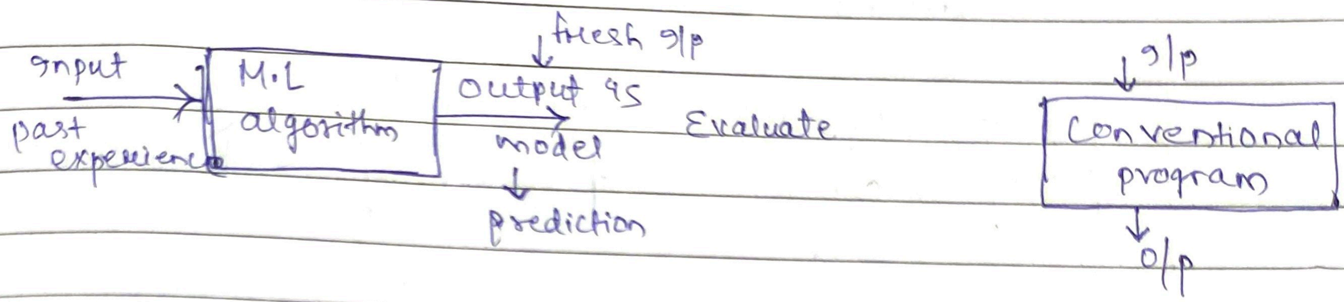
Machine Learning Handwritten Notes

5th January, 2023

MACHINE LEARNING

(TOM Mitchell)

Defn: It is the field of study that gives computers the ability to learn without being explicitly program.



Learning: A computer program is said to learn from experience 'E' w.r.t some class of tasks 'T' and performance measure 'P', if it's performance at tasks 'T' as measured by 'P' improves with experience 'E'.

Ex: handwritting Recognization

Task = recognizing and classifying handwritten words.

performance = percentage of words correctly classified. {Accuracy}

Experience = A dataset of handwritten words with given classification

Ex: Chess learning program/problem.

Task = playing chess

performance = performance of games won against the opponents.

Experience = no. of times, playing practice games against itself.

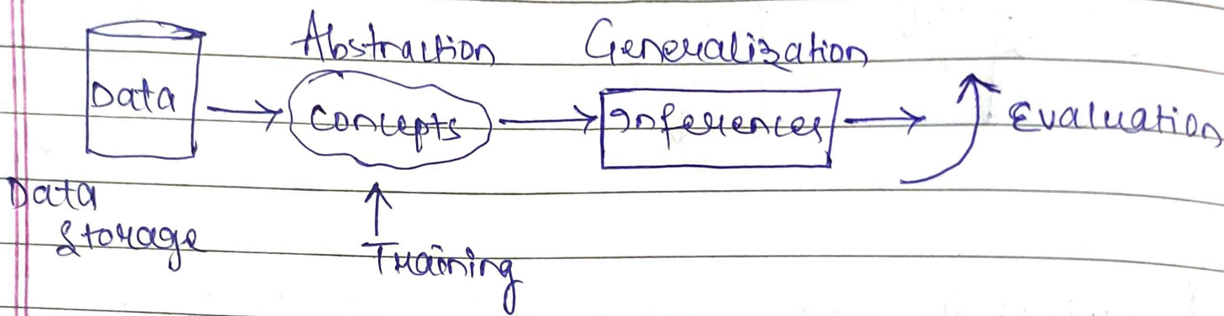
Ex3: Self driving car

Task = driving the car

Performance = average distance travelled before an error.

Experience = Sequence of images and steering commands recorded while observing a human driver.

How Machine learn:
The learning process weather by human or by machines can be divided into four components.



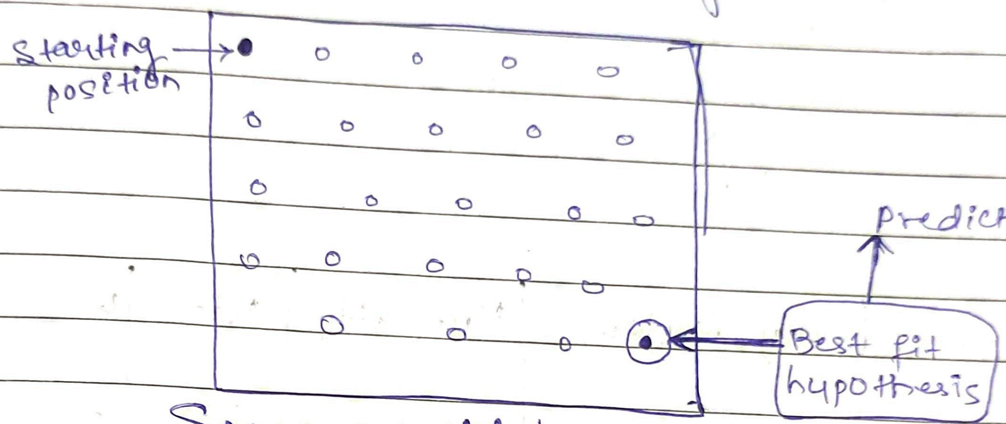
Data Storage: Humans and computers utilize data storage as a foundation for advanced reasoning.

Abstraction: The process of extraction of knowledge about stored data, while it also involves creating general concepts about the data as a whole.

Generalization: It describes the process of turning the knowledge about stored data into a form that can be utilized for future action.

Concept Learning in ML:

- Concept learning can be stated as the problem of inducing general functions from specific training examples. Concept learning can be formulated as a problem of searching through a pre-defined space of potential hypothesis for the hypothesis that best fits the training example.



Space potential hypothesis.

Ex $f_1 = A, B, \phi, ?$

$f_2 = X, Y$

Instance Space :- (A,X), (A,Y), (B,X), (B,Y)

Hypothesis space :- 16

Semantically distinct hypothesis :-

Instance Space → $[3 \times 2 \times 2 \times 2 \times 2 \times 2] = 96$

Hypothesis Space → $[5 \times 4 \times 4 \times 4 \times 4 \times 4] = 5120$

Semantically distinct hypothesis → $[4 \times 3 \times 3 \times 3 \times 3 \times 3] + 1 = 973$

Consistent Hypothesis: A consistency hypothesis is one which truely satisfy to all my training examples.

12th Jan, 2022

- Concept Learning is an inferring a boolean-valued func^n.
- Learning means acquiring general concepts from training examples.

Inductive Learning:

Any hypothesis found to approximate the target func^n well over a sufficiently large set of training examples will also approximate the target func^n well over other unobserved examples.

Target func^n,

$$\boxed{\underset{\text{hypothesis}}{h(x)} = \underset{\text{concept}}{c(x)}}$$

Example	sky	Airtemp	Humidity	Wind	Water	forecast	Enj sp
1	Sunny	Warm	Normal	Strong	warm	same	Ye
2	sunny	Warm	High	Strong	warm	Same	Ye
3	Rainy	cold	High	strong	warm	Change	No
4	Sunny	Warm	High	strong	cool	change	Ye

find-S Algo:

1) Initialize h to the most specific hypothesis in H.

$$h_0 = \{\overset{a_1}{\phi}, \overset{a_2}{\phi}, \overset{a_3}{\phi}, \overset{a_4}{\phi}, \overset{a_5}{\phi}, \overset{a_6}{\phi}\}$$

2) for each +ve training instance x
   - for each attribute constraint a_i in h.
     If the constraint a_i is satisfied by x
     Then do nothing
     Else replace a_i in h by the next more general constraint that is satisfied by x.

3) output hypothesis h

$$h_0 = \{ \overset{a_1}{\phi}, \overset{a_2}{\phi}, \overset{a_3}{\phi}, \overset{a_4}{\phi}, \overset{a_5}{\phi}, \overset{a_6}{\phi} \}$$

Training example — x_1 = <sunny, warm, Normal, strong, warm, same>

hypothesis — h_1 = <sunny, warm, Normal, strong, warm, same>
a_1 a_2 a_3 a_4 a_5 a_6

x_2 = <sunny, warm, high, strong, warm, same>

h_2 = <sunny, warm, ?, strong, warm, same>

x_3 = <Rainy, cold, High, strong, warm, change>

* $h_3 = h_2$

x_4 = <sunny, warm, High, strong, cool, change>

h_4 = <sunny, warm, ?, strong, ?, ?>

Maximally specific = h_4.

= Supervised learning - labeled data

unsupervised learning - unlabelled data

Reinforcement " - feedback & Reward.

- Classification (categorical data)
- Regression (Numerical data)
- Clustering.
- Anomaly detection (outlier detection)
  fraud detection

unsupervised learning.

= Distance measures:- mean, median, mode {mean is biased to outliers so we need median, the mean shifts towards the outliers if present}.

= Spread measures:- variance, Standard Deviation

16th Jan 2023

Example	a_1 Citation	a_2 Size	a_3 In-Library	a_4 Price	a_5 Editions	Buy
1	Some	Small	NO	Affordable	many	No
2	many	Big	NO	Expensi	one	Yes
3	Some	Big	always	Exp	few	No
4	many	medium	NO	Exp	many	Yes
5	many	Small	NO	Aff.	many	Yes

Q1 How many concepts are possible for this instance space?

Q2 How many hypothesis can be expressed by the hypothesis language?

Q3 Apply find S algo. on the given training set.

Soln 1) $2 \times 3 \times 2 \times 2 \times 3 = 72$

2) $4 \times 5 \times 4 \times 4 \times 5 = 1600$

Conceptually distinct hypo. $= (3 \times 4 \times 3 \times 3 \times 4) + 1$
$= 433$

$h_0 = \langle \phi, \phi, \phi, \phi, \phi \rangle$ (a_1, a_2, a_3, a_4, a_5)

3) $h_1 = \langle$ some, small, NO, Aff, many $\rangle$

$h_1 = \langle$ ~~some, small~~ $\phi, \phi, \phi, \phi, \phi \rangle$ { -ve }

$h_2 = \langle$ many, big, NO, Exp, one $\rangle$

$h_3 = \langle$ many, big, NO, Exp, one $\rangle$

Date
Page No.

$h_4 = \langle$ Many, ?, NO, Exp, ? $\rangle$

$h_5 = \langle$ many, ?, NO, ?, ? $\rangle$

∴ Maximally specific hypothesis = h_5

$h_1 \rightarrow$ Most Specific
h_2, h_3
h_4
$h_5 \rightarrow$ Most General.

= hypo. → is predicting.
↳ if all matches then o/p ~~now~~ can be +ve.
↳ if does not match then o/p can be -ve

An hypo. is said to be consistent if ~~the~~ the hypothesis ~~sat~~ satisfies / matches with all the actual ~~values~~ class levels.

Version Space:

Version Space "$VS_{H,D}$" is the subset

H → Hypo. Space, D → Training dataset

Of the hypotheeises from 'H' consistent with the training example in 'D'.

$$VS_{H,D} = \{ h \in H \mid \text{consistent}(h, D) \}$$

List - Then - Eliminate Algorithm:

Step-1: Initialize version space to the list containing every hypo. in 'H'

Step-2: For each training example which is represented as $\{x, c(x)\}$, remove from the version space any hypo. 'H' for which $h(x) \neq c(x)$

Step-3: Output the list ~~VS~~ of hypo. in VS.

Exmp: f_1: A, B
f_2: X, Y

Training instances:

f_1	f_2	Target
A	X	Yes
A	Y	Yes

Soln: $H = \{(A X), (A Y), (A ?), (B X), (B Y), (B ?), (? X), (? Y), (? ?), (\phi, \phi)\}$

1. VS = H
2. Eliminate:

$$VS_{HD} = \{(A ?), (? ?)\}$$

Previous algo. is for 'Rules', as outcome is hypothesis.

18th Jan 2023

Date
Page No.

Candidate Elimination Algorithm:

- We define a boundary space within the entire hypothesis space by shifting most general towards most specific hypo. i.e moving most specific towards most general hypo.

Ex:

Eg	Sky	Airtemp	Humidity	wind	H_2O	forecast	sport
1	Sunny	Warm	Normal	Strong	warm	same	yes
2	Sunny	Warm	High	Strong	warm	same	yes
3	Rainy	cold	High	Strong	warm	Change	No
4	Sunny	Warm	High	Strong	cool	Change	yes.

$S_0 = \{\phi, \phi, \phi, \phi, \phi, \phi\}$

$S_1 = \{$sunny, warm, Normal, strong, warm, same$\}$

{not matched ∴ change}

$S_2 = \{$sunny, warm, ?, strong, warm same$\}$

$S_3 = \{$~~sunny, warm,~~ S_2 {-ve eg. no change}

$S_4 = \{$sunny, warm, ?, strong, ?, ?$\}$

G_4

G_3

G_2

G_1

$G_0 = \{?, ?, ?, ?, ?, ?\}$

$G_1 = G_0$ {Matched}

$G_2 = G_1$

* Only one '?' can be replaced at a time.

$G_3 = \{sunny, ?, ?, ?, ?, ?\};$

~~G4~~ ~~&~~ ~~G5~~ $\{?, warm, ?, ?, ?, ?\};$

~~G~~ $\{?, ?, \cancel{low}\ Normal, ?, ?, ?\}; \{?, ?, ?, ?, \cancel{cool}\ ?\};$

$\{?, ?, ?, ?, ?, same\}$

$G_3 =$ ~~G~~ $\{sunny, ?, ?, ?, ?, ?\};$
$\{?, warm, ?, ?, ?, ?\};$
~~{?, ?, ?, ?, ?, cool}~~
$\{?, ?, ?, ?, ?, same\}$

It should also accomodate the earlier two training examples.

$G_4 = \{Sunny, ?, ?, ?, ?, ?\};$

~~G~~ $\{?, warm, ?, ?, ?, ?\}$

= After redefining we get 3 hypo., But these are not the only hypo. in version space, so, we need to explore.

$S_4 = \{sunny, warm, ?, strong, ?, ?\}$

= $\{sunny, warm, ?, ?, ?, ?\}; \{sunny, ?, ?, strong, ?, ?\};$
$\{?, warm, ?, strong, ?, ?\}$

$G_4 = \{sunny, ?, ?, ?, ?, ?\}; \{?, warm, ?, ?, ?, ?\}$

Above we are generating a specific hypo. by Comparing general hypo., and whenever there is a mismatch, we replace one '?' with one value.

19th Jan, 2023

	Citation	Size	In-Library	Price	Editions	Buy
1	Some	Small	No	Affordable	~~many~~ one	No
2	many	Big	No	Expensive	~~one~~ many	Yes
~~3~~	~~Some~~	~~Big~~	~~Always~~	~~Exp.~~	~~few~~	~~No~~
3 ~~4~~	many	medium	No	Exp	~~many~~ few	Yes
4 ~~5~~	many	Small	No	Aff.	many	Yes

$S_0 = \{\phi, \phi, \phi, \phi, \phi\}$

$S_1 = \{\phi, \phi, \phi, \phi, \phi\}$

$S_2 = \{$many, Big, NO, Expensive, many$\}$

$S_3 = \{$many, ?, NO, Exp, ?$\}$

$S_4 = \{$many, ?, NO, ?, ?$\}$

$G_4 = \{$many, ?, ?, ?, ?, ?$\}$;

$G_3 = \{$many, ?, ?, ?, ?$\}$; $\{$?, ?, ?, Exp, ?$\}$;

$G_2 = \{$many, ?, ?, ?, ?$\}$; $\{$?, big, ?, ?, ?$\}$; $\{$?, ?, ?, Exp, ?$\}$; $\{$?, ?, ?, ?, many$\}$
$\{$?, ?, ?, ?, few$\}$; $\{$?, ?, ?, ?, many$\}$; $\{$?, medium, ?, ?, ?$\}$;

$G_1 = \{$~~some~~ many, ?, ?, ?, ?$\}$; $\{$?, big, ?, ?, ?$\}$; $\{$?, ?, ?, Exp, ?$\}$;

$G_0 = \{$?, ?, ?, ?, ?$\}$

$S_4 = \{many, ?, NO, ?, ?\}$

↓

~~{many, ?, NO, ?, ?}, {many, ?, ?,~~

↑

$G_4 = \{many, ?, ?, ?, ?\}$

Ex	Size	Color	Shape	Class
1	Big	Red	Circle	NO
2	Small	Red	triangle	NO
3	Small	Red	Circle	Yes
4	Big	Blue	Circle	NO
5	Small	Blue	Circle	Yes.

$S_0 = \{\phi, \phi, \phi\}$

$S_1 = \{\phi, \phi, \phi\}$

$S_2 = \{\phi, \phi, \phi\}$

$S_3 = \{$ Small, Red, Circle $\}$

$S_4 = \{$ Small, Red, Circle $\}$

$S_5 = \{$ small, ?, Circle $\}$.

$G_5 = \{$Small, ?, Circle$\}$ ~~{Small, Blue, Circle}~~.

$G_4 = \{$ Small, ?, circle $\}$; ~~{Small, red, Circle}; {Small, Blue, Circle}~~

$G_3 = \{$Small, ?, Circle$\}$

{?, Blue, ?}; ~~{?, Blue, Circle}~~; {Big, ?, Δle}; {?, Blue, Δle}

$G_2 = \{$Small, Blue, ?$\}$; $\{$ small, ?, Circle$\}$; ~~{Big, Blue, ?}~~;

$G_1 = \{$Small, ?, ?$\}$; $\{$?, Blue, ?$\}$; $\{$?, ?, Δle$\}$

$G_0 = \{$?, ?, ? $\}$

23rd Jan, 2023

Date | Page No.

Effect of Noise:

- leads to complicacy in model building.
-

Learning Multiple classes:

- There are 2 ways of dealing with multiple classes:
  1) One-Against-All
  2) one-Against-one

1) One-Against-All: Here, we view a 'K' class classification prblm as 'K' no. of 2 class prblm. In the ith 2 class prblm, the training examples belonging to 'Ci' class are taken as +ve examples and examples of all other classes are taken as -ve examples.

$$h_i(x) = \begin{cases} 1 & , \text{if } x \text{ is in class } C_i \\ 0 & , \text{otherwise} \end{cases}$$

2) One-Against-One: Here, a classifier is constructed for each pair of class. If there are 'K' diff. class levels, a total of '$K(K-1)/2$' classifiers are constructed. An unknown instance is classified with the class getting the most votes. Ties are broken arbitrally.

30th Jan, 2023

Model Selection:

- Model selection is understood as some mathematical expression or equation or some mathematical structures such as graphs and trees or a set of logical rules.
- Inorder to formulate a hypothesis for a problm we have to chose some model and hence the term model selection has been used.
- Model selection can also be defined as the process of chosing one particular approach from among several diff. approaches.
- This may include chosing:

 1) An appropriate algo. from a selection of possible algorithms.

 2) Chosing the set of features to be used for i/p.

 3) Chosing initial ~~par~~ values for certain parameters.

 4) Picking a particular mathematical model from among diff. mathematical models.

- The whole process is also described as process of chosing the right inductive bias.

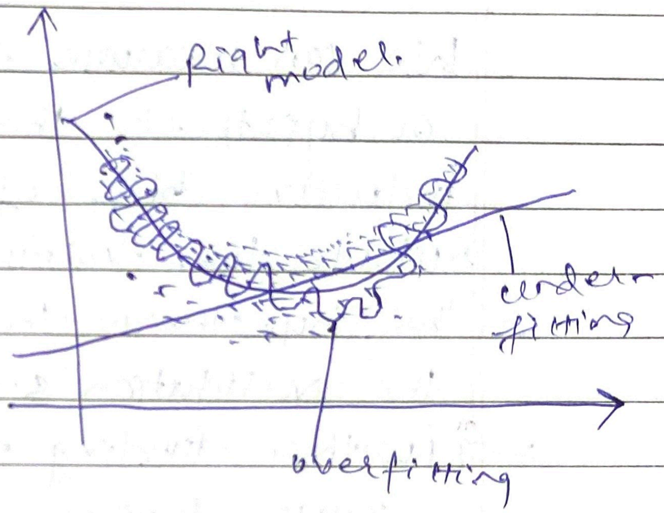
Inductive Bias: ~~An inductive bias is~~ A set of assumptions we make to have learning possible is called inductive bias of the learning algo.

E.g: In Regression assuming a linear funcn is an inductive bias.

Date | Page No.

Advantages of a Simpler Model:

1) A simple model is easy to use.
2) A simple model is easy to train.
3) A simple model is easy to explain. {visualization}
4) A simple model would generalize better than a complex model.



1st feb, 2023

Generalization: Generalization refers to how well the concepts learned by a machine learning model apply to specific examples not seen by the model when it was learning.

- The goal of a good ML model is to generalize well from the training data to any data from the problem domain.
- Overfitting and underfitting are 2 biggest causes of poor performance of the ML algo.
- Underfitting is the predn of the ML model that is not complex enough to accurately capture the r/n ship between a dataset features and the target variables.

- **Overfitting**: is the prodⁿ of a ML model which corresponds too closely or exactly to a particular set of data, and therefore fail to predict future observations reliably.

Cross Validation {Testing Generalization}

We can measure the generalization ability of a hypothesis that is the quality of its inductive bias if we have access to data outside the training set.
The hypothesis that is the most accurate on the validation set is the best one.
= {further dividing the training data set into 2 parts - training and validation is called cross validation}.

Dimensionality Reduction:

- In Statistical ML, dimensionality reduction or dimension redⁿ is the process of reducing the no. of variables under consideration by obtaining a smaller set of principle variables.
- Dim. Redⁿ may be implemented in 2 ways:
  1. Feature Selection in which we are interested in finding 'K' of total 'n' features that gives us the most info. and we discard the other 'n-K' features.
  2. feature Extraction in which we are interested in finding a new set of 'K' features that are ~~they~~ a combination of original 'n' features.

This process of feature extraction may be supervised or unsupervised depending on whether they use the o/p info.

Ex: principle component Analysis (PCA) which is an unsupervised model and Linear discriminate Analysis (LDA) which is a supervised method.

Measures of Error:

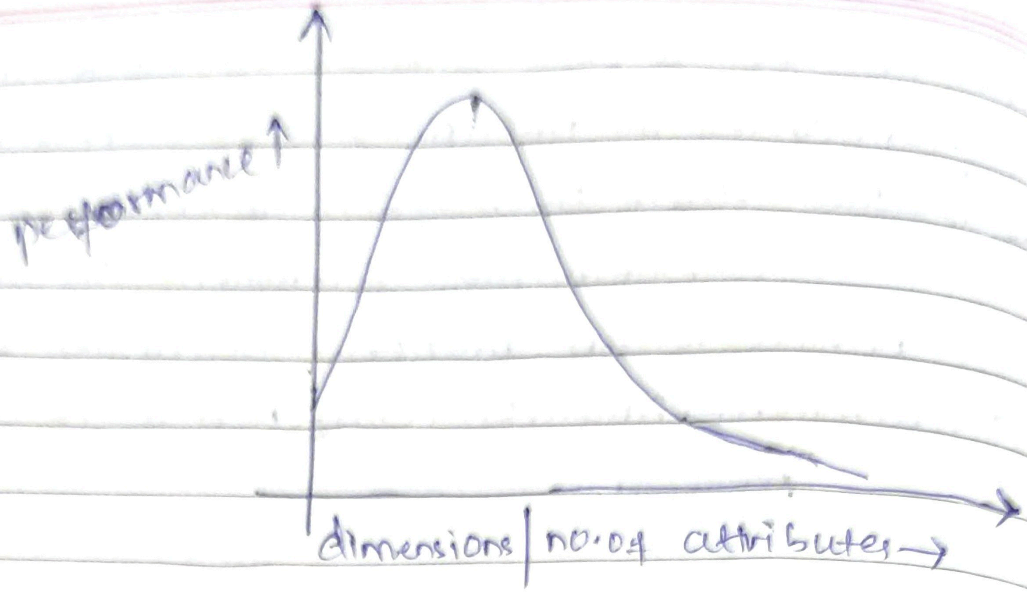
- In both the methods of dim. Redn, whether feature selection or feature extraction we need a measure of error in the model.
- Measures of error can be:
  1. MSE (Mean squared error)
  2. RMSE (Root mean squared error)

These can be for regression prblm(s).

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (\underbrace{y_i}_{\text{observed value}} - \underbrace{\bar{y}_i}_{\text{predicted value}})^2$$

Why Dim. Redn?

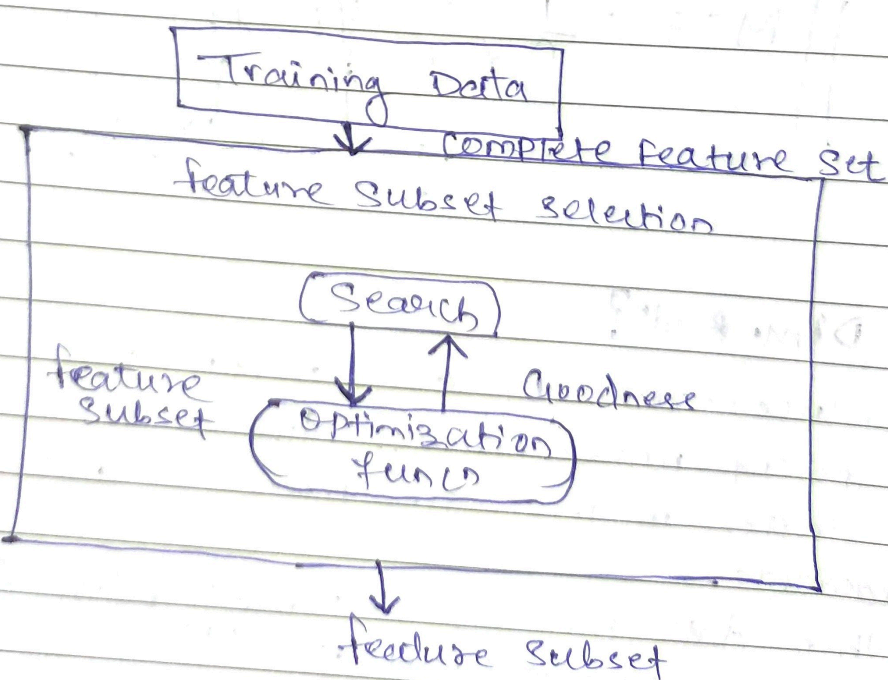
1. Reduced memory and computation.
2. Ignoring unnecessary i/p(s)
3. Simpler models are more robust on small datasets.
4. When data can be explained with few features
   ... continue at your end !!!
5. When data can be represented in few dimension without loss of info.



Assumption: The no. of training instances are constant.
↳ means observations.

8th feb, 2023 Feature Selection
↳ forward Selection
↳ Backward Selection

Basic flowchart for feature Selection Steps:



Measures of Error

for Regression approach:) MSE
) RMSE

for classification approach:
) Misclassification Rate)

$$= \frac{\text{No. of samples misclassified}}{\text{Tot. no. of samples}}$$

Both are optimising criteria.

Forward Selection Algo: Greedy Approach

- Forward subset selⁿ is a greedy approach:
  1. Start with an empty feature set.
  2. Try each remaining feature.
  3. Estimate classification | Regression error for adding each feature.
  4. Select feature that gives maximum improvement.
  5. Stop when there is no significant improvement.

* In forward selⁿ algo. → Requirement: Once a feature is decided to be included in the subset, it is never removed.

Backward Selection Algo: {also called Backward elimination}

1. Backward selection algo. starts with full feature set.
2. Try removing features.
3. Drop the features with smallest impact on error.
4. In Backward selection once a feature is dropped,

it's never going to be considered again.

9th feb 2023

Feature Extraction:

PCA {Principle Component Analysis}

→ which feature is more discriminative
→ which one has most impact on dataset.
→ If the spread across an axis is more i.e., range is more it is more discriminative.

→ It is unsupervised.

Ex:

	Ex_1	Ex_2	Ex_3	Ex_4
X_1	4	8	13	7
X_2	11	4	5	14

$8 = \overline{X_1}$

$\overline{X_2} = 8.5$

Step 1: No. of dimension / feature = 2
No. of datapoints / example = 4 (N)

Step 2: find $\overline{X_1}$ and $\overline{X_2}$.

Step 3: find covariance matrix:
$$\begin{bmatrix} (X_1 X_1) & (X_1 X_2) \\ (X_2 X_1) & (X_2 X_2) \end{bmatrix}$$

$$\text{cov}(X_1\ X_2) = \frac{1}{N-1} \sum_{i=1}^{N} (X_{iK} - \overline{X_i})(X_{jK} - \overline{X_j})$$

$$\Rightarrow \text{cov}(X_1 X_1) = \frac{1}{3} \{ (4-8)^2 + (8-8)^2 + (13-8)^2 + (7-8)^2 \}$$

$$= \frac{1}{3} \{ 16 + 0 + 25 + 1 \}$$

$$= \frac{1}{3} \times 42 = \underline{14}$$

$$\mathrm{cov}(x_1\ x_2) = \frac{1}{3}\{(4-8)(11-8.5) + (8-8)(4-8.5) + (13-8)(5-8.5) + (7-8)(14-8.5)\}$$

$$= \frac{1}{3}\{-10 + 0 - 17.5 - 5.5\}$$

$$= \frac{1}{3}\{-27.5 - 5.5\} = \frac{1}{3}\{-33.0\}$$

$$= -11$$

$$\mathrm{cov}(x_2\ x_1) = \mathrm{cov}(x_1\ x_2) = -11$$

$$\mathrm{cov}(x_2\ x_2) = \frac{1}{3}\{(2.5)^2 + (-4.5)^2 + (-3.5)^2 + (5.5)^2\}$$

$$= \frac{1}{3}\{6.25 + 20.25 + 12.25 + 30.25\}$$

$$= \frac{1}{3}\{69\} = 23$$

$$\therefore \mathrm{cov}(x_1\ x_2) = \begin{bmatrix} 14 & -11 \\ -11 & 23 \end{bmatrix}$$

Date | Page No.

Step 4: find Eigen values for Cov-Matrix.

$$\det(S - \lambda I) = 0$$

Here, det = determinant

S = cov matrix

λ = eigen values

I = Identity matrix.

$$\left| \begin{bmatrix} 14 & -11 \\ -11 & 23 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right| = 0$$

$$\begin{vmatrix} 14-\lambda & -11 \\ -11 & 23-\lambda \end{vmatrix} = 0$$

$$(14-\lambda)(23-\lambda) - 121 = 0$$

$$322 - 14\lambda - 23\lambda + \lambda^2 - 121 = 0$$

$$201 - 37\lambda + \lambda^2 = 0$$

$$\lambda^2 - 37\lambda + 201 = 0$$

$$\therefore \lambda = 30.38 \;||\; 6.62$$

Step 5: ~~Generate from λ_1 the first eigen vector.~~

Computation of eigen vectors:

$$U = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$$

$$\begin{bmatrix} 0 \\ 0 \end{bmatrix} = (S - \lambda I)\, U$$

First eigen vector from first eigen value.

$$\begin{bmatrix} -16.38 & -11 \\ -11 & -7.38 \end{bmatrix} U = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\begin{bmatrix} -16.38 & -11 \\ -11 & -7.38 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\begin{bmatrix} -16.38\, u_1 - 11\, u_2 \\ -11\, u_1 - 7.38\, u_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$\cancel{16.38 u_1 + 11 u_2 = 0}$ — (1) $\qquad (14 - \lambda_1) u_1 - 11 u_2 = 0$ — (1)

$\cancel{11 u_1 + 7.38 u_2 = 0}$ — (2) $\qquad -11 u_1 + (23 - \lambda_1) u_2 = 0$ — (2)

$11 u_1 = (23 - \lambda) u_2 = t$

$\dfrac{u_1}{23 - \lambda_1} = \dfrac{u_2}{11} = t$

$u_1 = 0.5574$

$u_2 = -0.8303$

$u_1 = 23 - \lambda_1$ so, u_2

$23 - 30.38$

$= -6.62$

$u_2 = 11$

$u_2 = 11$

$u_1 = -6.62$

$$\begin{bmatrix} 0.5574 \\ -0.8303 \end{bmatrix}$$

$$\begin{bmatrix} -6.62 \\ 11 \end{bmatrix}$$

2nd eigen vector

$$\begin{bmatrix} 7.38 & -11 \\ -11 & 16.38 \end{bmatrix} \begin{bmatrix} U_1 \\ U_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\begin{bmatrix} 7.38U_1 - 11U_2 \\ -11U_1 + 16.38U_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$7.38U_1 - 11U_2 = 0$ —①

$-11U_1 + 16.38U_2 = 0$ —②

$$U_2 = \begin{bmatrix} 0.8303 \\ 0.5574 \end{bmatrix}$$

Step 6: Computation of principle Components.

from eqⁿ ①

$$\frac{U_1}{11} = \frac{U_2}{14-\lambda_1} = t \quad \{\text{Assume } t=1\}$$

$U_1 = 11; \ U_2 = 14 - \lambda_1$

Hence, eigen vector: $U_1 = \begin{bmatrix} 11 \\ -16.38 \end{bmatrix}$

By normalizing: $\|U_1\| = \sqrt{11^2 + (-16.38)^2} = \underline{\underline{19.73}}$

$$\therefore \ e_1 = \begin{bmatrix} 11/\|U_1\| \\ -16.38/\|U_1\| \end{bmatrix} = \begin{bmatrix} 0.55 \\ -0.83 \end{bmatrix}$$

Similarly

2nd eigen vector

$$\begin{bmatrix} 14-\lambda_2 & -11 \\ -11 & 23-\lambda_2 \end{bmatrix} \begin{bmatrix} U_1 \\ U_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$(14-\lambda_2) U_2 - 11 U_1 = 0$ —— (1)

$-11 U_1 + (23-\lambda_2) U_2 = 0$ —— (2)

from eq (2)

$\frac{U_1}{14-\lambda_2} = \frac{U_2}{11} = t$ {Assume $t = 1$}

$\Rightarrow$ ~~$U_1 = 0.83$~~

~~$U_2 = 0.55$~~

$U_2 = \begin{bmatrix} 7.38 \\ 11 \end{bmatrix}$

~~$\therefore U_2 = \begin{bmatrix} 0.83 \\ 0.55 \end{bmatrix}$~~

By normalizing $|U_2| = \sqrt{11^2 + (7.38)^2} = 13.3$

$\therefore e_2 = \begin{pmatrix} 7.38/13.3 \\ 11/13.3 \end{pmatrix} = \begin{bmatrix} 0.83 \\ 0.55 \end{bmatrix}$

Step 6: Derive New Data Set :

Principle Component PC-1	P_{11}	P_{12}	P_{13}	P_{14}

$$P_{11} = e_1^T \begin{bmatrix} 4-8 \\ 11-8.5 \end{bmatrix}$$

$$\begin{bmatrix} 0.55 & -0.83 \end{bmatrix} \begin{bmatrix} -4 \\ 2.5 \end{bmatrix}$$

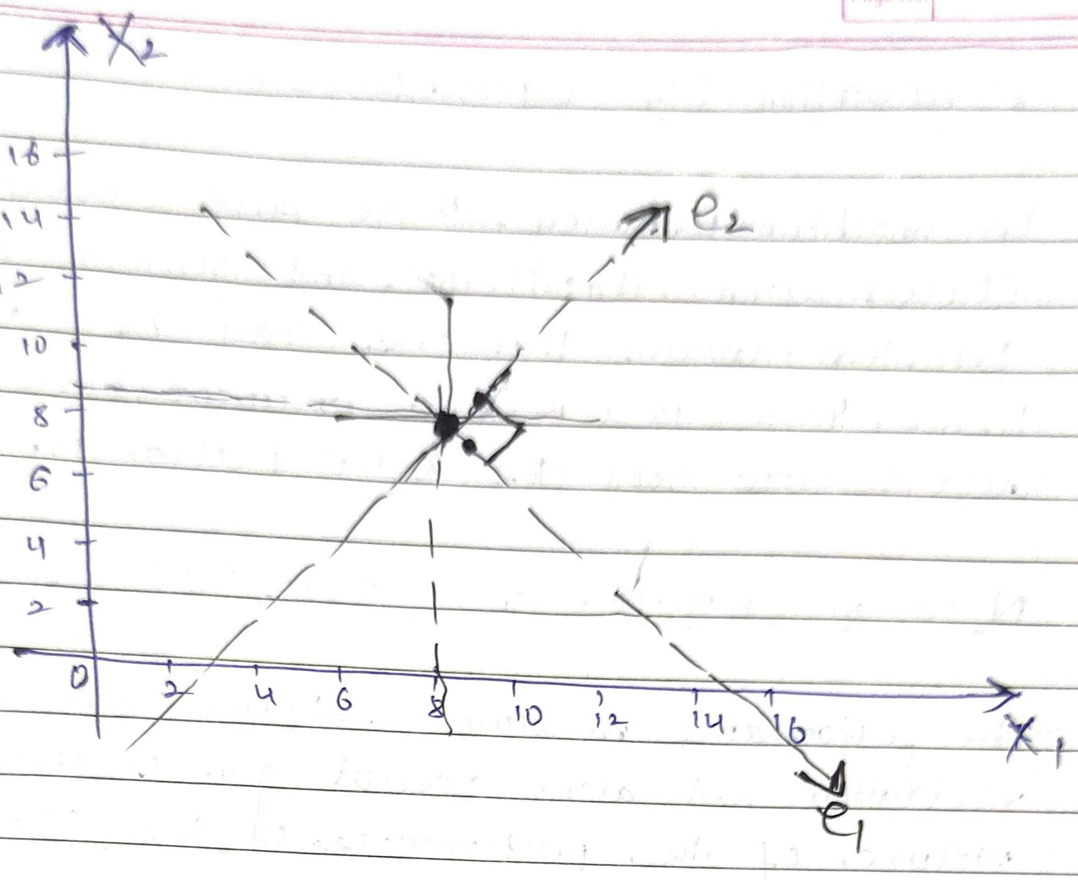
$$= \underline{\underline{-4.30}}$$

$$P_{12} = e_1^T \begin{bmatrix} 8-8 \\ 4-8.5 \end{bmatrix} = 3.73$$

$P_{13} = 5.69$; $P_{14} = -5.12$

$\therefore$

PC-1	-4.30	3.73	5.69	-5.12



X_2
16
14
12
10
8
6
4
2
0
2
4
6
8
10
12
14
16
X_1
e_2
e_1

12th Feb, 2023 UNIT-2

Evaluation Of Classifiers:

In machine Learning, there are several Classification algorithms, and given a certain problem, more than one may be applicable. Hence, there is a need to examine, how we can access how good the selected algo. is.

Need for multiple validation Sets?

- The performance measure obtained by a single validation set alone doesnot give a true picture of the performance of the classifier.
- Also, this measures alone cannot be meaningfully used to compare 2 algo.

Hence, we need to have diff. validation Sets. Diff. methods available for ~~providing~~ generating multiple training validation Sets, from a given dataset are Cross validation in general and K-fold Cross validation in particular.

- Classification algo.(s) can be compared based on not only error rates but several other criterias like:
  1. Training time and space complexity
  2. Testing time and space complexity
  3. Interpreterability
  4. Easy programmability

K-fold cross validation:

- Cross validation is a technique to evaluate predictive models by partitioning the original sample ~~by~~ into training set to train the model and test set to validate it.
- In K-fold cross validation, a dataset is divided randomly into 'K' equal size parts. To generate each pair we keep one of the 'K' parts out as the validation set and combine the remaining 'K-1' parts to form the training set. This process is carried out 'K' times.

Demerits: {limitations}

① ~~First~~ We keep the training set large. Validation sample is small.

② The training sets overlapp considerably

③ Typically 'K' is set to 10 or 30. As 'K' increases, the %age of training instances increases, and we get more robust estimaters but the validation set becomes smaller.

④ There cost of training the classifier 'K' times which increases as the 'K' is increase

15th feb, 2023

5x2 cross validation:

→ X

$$\left[\ \frac{X_1}{T} \qquad \frac{X_2}{V}\right.$$

$$\left[\ \frac{X_2}{T} \qquad \frac{X_1}{V}\right.$$

~~Tee~~ - Sampling is done.

Boot Strapping:

~~Sampling~~ - sampling with replacement.

- Boot strapping in ML is the process of computing performance measures, using several randomly selected training and test datasets, which are selected through a process of sampling with replacement.
- Sample datasets are selected multiple times.

Measuring Error: True positive (TP)
False Negative (FN)
True Negative (TN)
False Positive (FP)

Confusion matrix:

	Actual label of x is C	Actual label of x is ¬C
Predicted label of x is True	TP	FP
x is false	FN	TN

Performance Measure:

1. Accuracy : defined as → $\frac{TP+TN}{TP+FP+FN+TN}$

2. Error Rate : 1 − Accuracy.

3. Sensitivity : $\frac{TP}{TP+FN}$

4. Specificity : $\frac{TN}{TN+FP}$

5. F-Measure : $\frac{2\times TP}{2\times TP+FP+FN}$

16th Feb, 2023

Precision (P) : $\dfrac{TP}{TP+FP}$

Recall r(R) : $\dfrac{TP}{TP+FN}$

Q. Suppose a computer program for recognising dogs in photograph identifies 8 dogs in a picture containing 12 dogs and some cats. Of the 8 dogs identified, 5 actually are dogs while the rest are cats. Compute the precision and recall of the computer program.

Actual →

prediction ↑	C (dogs)	¬C (cats)
dogs (C)	TP 5	FP 3
¬C (cats)	FN ~~4~~ ~~3~~ 7	TN

$\therefore$ Precision $= \dfrac{TP}{TP+FP} = \dfrac{5}{5+3} = \dfrac{5}{8}$ //

$r(R) = \dfrac{TP}{TP+FN} = \dfrac{5}{5+7} = \dfrac{5}{12}$ //

Problm stat. 2

Let there be 10 balls (6 white, 4 red) in a box, and let it be required to pick red balls from them. Suppose we pickup 7 balls as the red balls of which only 2 are actually red balls. find values of precision & recall in picking up red ball.

	C (Red) (4)	¬C (White) (6)
C (Red)	TP 2	FP 5
¬C (White)	FN 2	TN 1

$$\therefore \text{Precision} = \frac{TP}{TP+FP} = \frac{2}{2+5} = \frac{2}{7}$$

$$\text{Recall} = \frac{TP}{TP+FN} = \frac{2}{2+2} = \frac{2}{4}$$

ROC(AUC) {Receiver operating characteris

> TPR : True +ve Rate
> FPR : false +ve Rate

- ~~Receivi~~ ROC stands for Receiver. Operating characteristics.
- ROC curve is plotted taking into account 2 parameters namely: TPR & FPR.

$$TPR = TP/(TP+FN) = \text{Sensitivity}$$

$$FPR = FP/(FP+TN) = 1-\text{sensitivity}$$

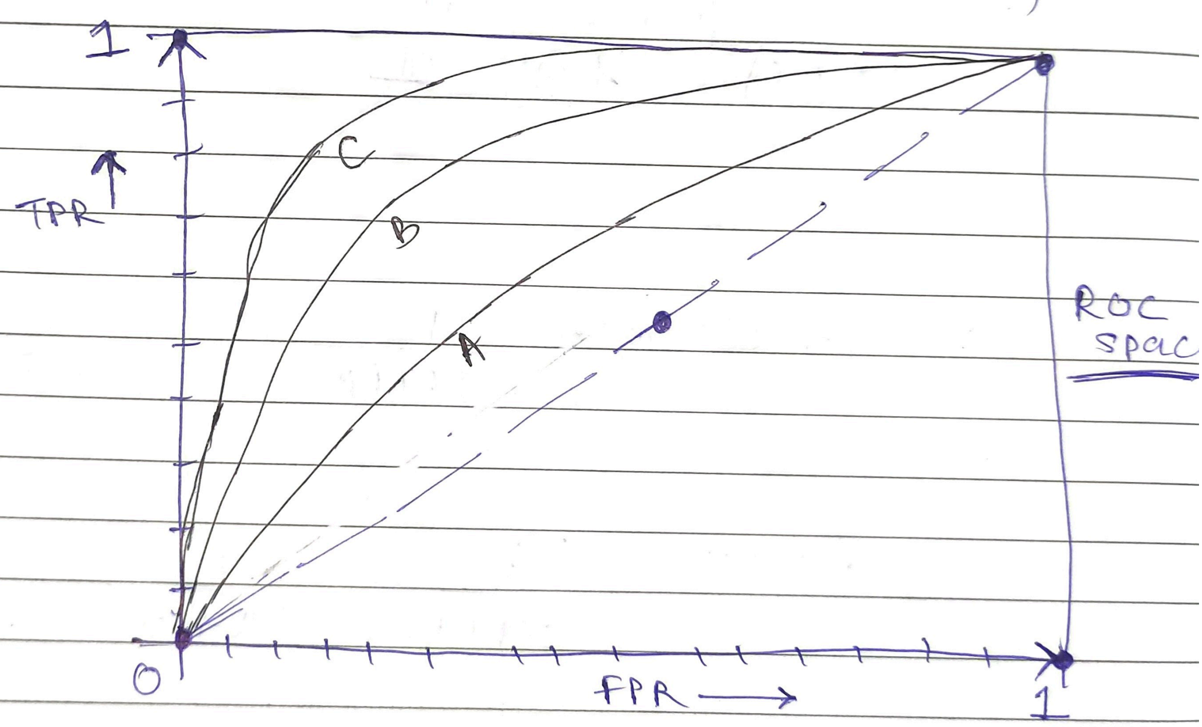


fig. ROC curve

> In above figure C is the best classifier. {e.g.}

More closer to perfect prediction {0,1}

Date
Page No.

- In case of certain classification algorithms the classifier may depend on a parameter. Different values of the parameter will give different classifier, and this in turn will give diff. values to TPR and FPR.

- The ROC curve is the curve determined by plotting the points (TPR, FPR) in the ROC space.

1: The left bottom corner (0,0): This interprets always negative prediction.
2: The right ~~bottom~~ top corner (1,1): Always +ve prediction
3: The left top corner (0,1): Perfect prediction.
4: The right bottom corner (1,0): no significance.
5: Any point on the diagonal: represents random prediction {unpredictable}

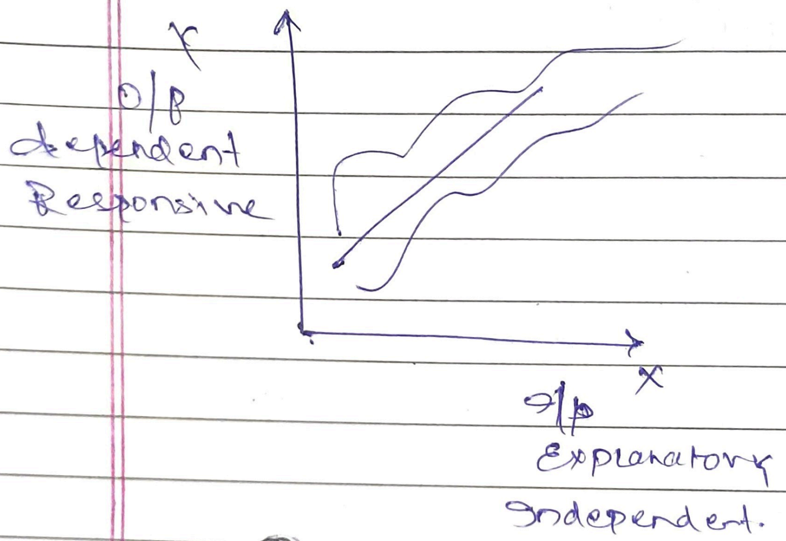
01/03/2023

REGRESSION ANALYSIS:

+ In Regression Analysis, we study about the r/nship b/w variables which are continous / Numerical.
+ It is basically:
  - Linear v/s Non-Linear regression
  - Simple v/s multiple regression

Simple: Only one independent variable
Multiple: multiple independent "

Linear: The ~~ind~~dependent variable changes linearly with independent variable
Non-Linear: The dep. var. changes non-linearly with ind. var.



Goals

- Prediction
- Interpretation.

Example of Regression Analysis:
1) How do wages of the employees get affected with experience, education, promotion, et
{cross-sectional data}

Date			
Page No			

2) How do current price of the stock depend on its past values. { Time Series data }

{ For cross sectional data, there is no time strain attached, to the value of variables, but if time has got influence on values of variable, then the r/nshp b/w such variables is called Time Series data. }

3) How does sales revenue get affected as a func^n of advertising expenses, competitor advertisement, etc.

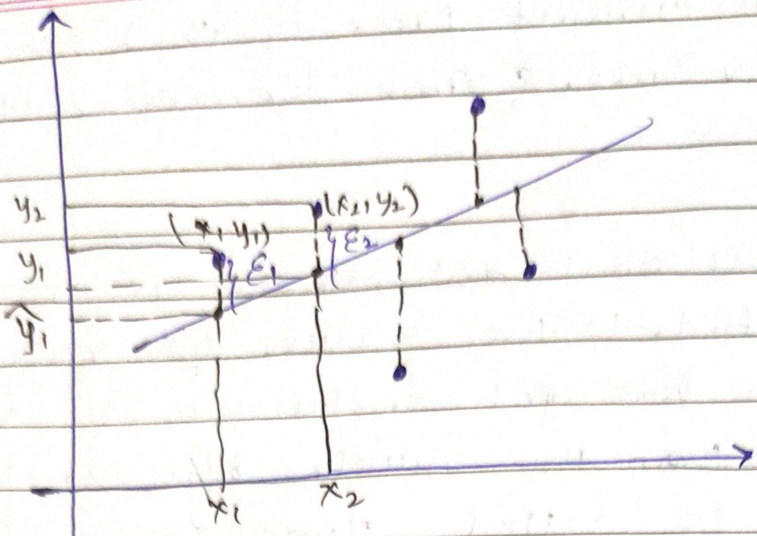
4) R/nshp b/w Speed and fuel efficiency of a car.

5) How does the price of a house gets affected by no. of bedrooms, square footage, etc.

2nd March, 2023

Simple Linear Regression:

↳ Ordinary Least Squares (OLS) { optimization techniques }



$y_1, y_2 \cdots$ = actual o/p
$\hat{y}_1, \hat{y}_2 \cdots$ = predicted o/p
$x_1, x_2 \cdots$ = given i/p
$\varepsilon_1, \varepsilon_2 \cdots$ = Error {Actual - predicted}

$$\boxed{y_i = \hat{y}_i + \varepsilon_i}$$ {Actual = predicted + Error}

$y = mx + c$

$\Rightarrow \boxed{y_i = b_0 + b_1 x} + \varepsilon_i$

$\Rightarrow \varepsilon_i = y_i - [b_0 + b_1 x]$

$\Rightarrow \| \varepsilon_i = y_i - b_0 - b_1 x \|$

Since, we are trying to minimize the square of errors, So, we need to find an expression for square of error. So, we find SSE i.e, Sum of square of errors :

$$\| SSE = \sum_{i=1}^{n} (\varepsilon_i)^2 \|$$

$$SSE = \sum_{i=1}^{N} (y_i - b_0 - b_1 x_i)^2$$

Goal : To minimize SSE

So, $$\frac{\partial SSE}{\partial b_0} = \frac{\partial}{\partial b_0}\left[\sum_{i=1}^{N} (y_i - b_0 - \underset{\substack{\uparrow \\ \text{constant}}}{b_1 x_i})^2\right] \quad \text{—} (1)$$

$$\frac{\partial SSE}{\partial b_1} = \frac{\partial}{\partial b_1}\left[\sum_{i=1}^{N} (y_i - \underset{\substack{\uparrow \\ \text{Constant}}}{b_0} - b_1 x_i)^2\right] \quad \text{—} (2)$$

{We have 2 unknown b_0 and b_1 So, we can minimize it by double partial differentiation. first by keeping one constant and then keeping the other one constant.}

3rd March 2023

Date | Page No.

$$b_0 = \bar{y} - b_1 \bar{x}$$

$$b_1 = \frac{\overset{\text{covariance}}{cov(x, y)}}{\underset{\text{variance}}{var(x)}}$$

$$var(x) = \frac{1}{n-1} \sum (x_i - \bar{x}_i)^2$$

$$cov(x, y) = \frac{1}{n-1} \sum (x_i - \bar{x})(y_i - \bar{y})$$

- b_0, b_1 are also called as regression coefficients.

Q

X	1.0	2.0	3.0	4.0	5.0
Y	1.00	2.0	1.30	3.75	2.25

Obtain a linear regression for data given in the table assuming 'Y' is the ~~in~~ dependent variable.

$$\bar{X} = \frac{15}{5} = 3$$

$$\bar{Y} = \frac{10.30}{5} = 2.06$$

$n = 5$ {no. of data points}

$$y = b_0 + b_1 x$$

$\cancel{b_0 = \bar{y}}$.

$$b_1 = \frac{cov(x,y)}{var(x)}$$

$$cov(x,y) = \frac{1}{4}\left\{\cancel{-2-1+0+1+2}\right\}\left\{\cancel{\ldots}\right\}$$

$$= 0$$

$\therefore b_1 = 0$

So, $b_0 = 2.06$

$\therefore y = 2.06$

$$cov(x,y) = \frac{1}{4}\left\{-2.(-1.06) + (-1)(-0.06) + 0 + (1)(1.69) + (2)(0.19)\right\}$$

$$= \frac{1}{4}\left\{4.25\right\} = 1.0625$$

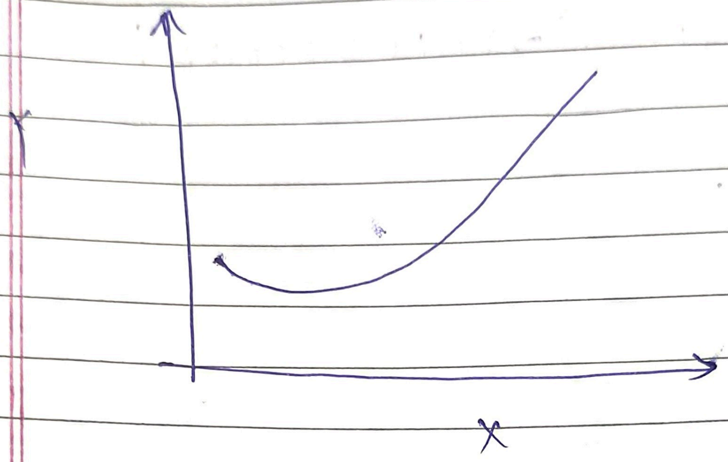
$$var(x) = \frac{1}{4}\left\{4+1+1+4\right\} = 2.5$$

$$\therefore b_1 = \frac{1.0625}{2.5} = 0.425$$

$$\text{So, } b_0 = 2.06 - 0.425(3)$$
$$= 0.785$$

So, $\| y = 0.785 + 0.425x \|$ Reg. Linear Regression

Polynomial Regression:



→ We will not get a linear model as it will not correctly represent the datapoints.

Actual Eqⁿ: $$\| y = b_0 + b_1 x + b_2 x^2 + b_3 x^3 + \cdots + b_k x^k \|$$

Error Eqⁿ: $$E = \sum_{i=1}^{n} \Big[\underbrace{y_i}_{\text{Actual}} - \underbrace{(b_0 + b_1 x + b_2 x^2 + \cdots + b_k x^k)}_{\text{predicted}} \Big]^2$$

System of Non-linear Eqⁿ:

$$\boxed{\frac{\partial E}{\partial b_i} = 0} \quad \forall\ i = 0 \text{ to } K \qquad (1)$$

for optimization

~~Soln~~ Simplifying eqn (1) we get system of non-linear eqn:

$$\Sigma y_i = b_0 n + b_1(\Sigma x_i) + \dots\dots + b_k(\Sigma x_i^k)$$

$$\Sigma y_i x_i = b_0(\Sigma x_i) + b_1(\Sigma x_i^2) + \dots\dots + b_k(\Sigma x_i^{(k+1)})$$

$$\Sigma y_i x_i^2 = b_0(\Sigma x_i^2) + b_1(\Sigma x_i^3) + \dots + b_k(\Sigma x_i^{(k+2)})$$

⋮

~~and so on~~

$$\Sigma y_i x_i^k = b_0(\Sigma x_i^k) + b_1(\Sigma x_i^{k+1}) + \dots + b_k(\Sigma x_i^{2k})$$

Q

X	3	4	5	6	7
Y	2.5	3.2	3.8	6.5	~~4.5~~ 11.5

Find a quadratic regression model for above data

Ans model eqn will be of form:

$$\| y = b_0 + b_1 x + b_2 x^2 \|$$

~~1st eqn~~ 1st eqn

$$\Sigma y_i = b_0 n + b_1(\Sigma x_i) + b_2(\Sigma x_i^2)$$

$$\Rightarrow 27.5 = 5b_0 + 25b_1 + 135b_2 \quad \text{—— (1)}$$

Similarly

$$158.8 = 25b_0 + 135b_1 + 775b_2 \quad \text{--- (2)}$$

$$996.2 = 135b_0 + 775b_1 + 4659b_2 \quad \text{--- (3)}$$

Solve eqn (1), (2), (3) for getting values of b_0, b_1, b_2.

$b_0 = 61.71$; $b_1 = -26.94$; $b_2 = 2.90$

$$\therefore \quad \boxed{y = 61.71 - 26.94x + 2.90x^2}$$