Sudesh Pawar
BE Comp A
Roll No. 24

A1-LR

Q.1) What are the applications of Linear Regression?

Ans If a company's sales have increased steadily every month for the past few years, by conducting a linear analysis on the sales data with monthly sales, the company could forecast sales in future months.

- If a company wants to know if the funds that they have invested in marketing a particular brand has given them substantial return on investment they can use linear regression
- Linear Regression can also be used to assess risk in financial ~~sources~~ services or insurance domain.
- In credit card industry, financial company may be interested in minimizing the risk portfolios & wants to understand the top five factors that cause customer to default.
- Based on the results of the company could implement specific EMI options so as to minimize default among risky ~~so~~ customers

Q.2) What are important functions used for linear regression while program implementation & explain their purpose:

Ans.: i) LinearRegression(): Fits a linear model with coefficients $w = (w_1, w_2, \ldots, w_p)$ to minimize the residual sum of squares between the observed targets in the dataset & the targets predicted by the linear approximation.

ii) fit(x,y): Fit linear model

iii) predict(x): Predict using linear model

iv) score(x,y): Returns coefficient of determination R^2 of the prediction

Q.3) Solve problem for given dataset in problem statement & find b_0 and b_1 in equation.

Ans.

X	Y	x^2	X.Y.
10	95	100	950
9	80	81	720
2	10	4	20
15	50	225	600
10	45	100	450
16	98	256	1488
11	38	121	418
16	93	256	1568
$\Sigma X = 73$	$\Sigma Y = 509$	$\Sigma x^2 = 887$	$\Sigma XY = 4876$

$$m = \frac{n.\Sigma xy - \Sigma x.\Sigma y}{n.\Sigma x^2 - (\Sigma x^2)^2} = \frac{8 \times 4876 - 73 \times 509}{8 \times 887 - 5329} = 4.58.$$

$$c = (\Sigma y - b \times \Sigma x) \div n = 509 - 4.58 \times 73 / 8$$
$$= 12.5846$$

Equation of Line is $y = mx + c$
$$= 4.58x + 12.5846.$$

Sudesh Pawar
BE Comp A
Roll No: 24

A2 - DT

Questions

1) Explain following terminologies related to decision tree building.

Ans. a) Impurity:

Impurity defines how well each classes are seperated. In general, the impurity measure should satisfy the most when data are split evenly for attribute values,

$$P_i = \frac{1}{\text{No. of classes}}$$

Impurity should be 0 when all data belong to same class.

b) Entropy:

The entropy of a random variable X is defined by,

$$\text{Entropy } (H(x)) = -\sum p(x) \log p(x)$$

- The entropy measures expected uncertainity in X. It has following properties:

- $H(x) \geq 0$, entropy is always non-negative.

$H(x) = 0$, if & only if x is deterministic

c) Information Gain:

The expected information needed to classify a tupple in P is given by:

$$\text{Info}(P) = \sum_{i=1}^{n} P_i \log_2(P_i)$$

$$\text{Info}_A(P) = \sum_{i=1}^{n} \frac{|P_i|}{|P|} \times \text{Info}(P_i)$$

$$\text{Gain}(A) = \text{Info}(P) - \text{Info}_A(P)$$

Q.2) What is Gini Index? Explain with formula.

Ans. The Gini Index is used in CART, of P, a data partition or set of training tuples as

$$Gini(P) = 1 - \sum_{i=1}^{n} P_i^2$$

- The attribute that maximizes the reduction in impurity (or equivalent, has the min. Gini Index 1 is selected as selected as splitting attribute.
- Gini Index is a metric to measure how often a randomly chosen element would be incorrectly identified.

Q.3) Solve the problem for given dataset in problem statement to explain how to find root node using entropy & information gain.

→ Solution:

Step 1: Finding the entropy

$$Entropy(E(s)) = -\Sigma p(x) . \log_2 p(x)$$

∴ for given dataset, entropy is,

$$Entropy = -P(yes) . \log_2(P(yes)) - P(No) . \log_2(P(No))$$

$$= -\left(\frac{9}{14}\right) \log_2\left(\frac{9}{14}\right) - \left(\frac{5}{14}\right) \log_2\left(\frac{5}{14}\right)$$

$$Entropy = 0.94$$

Step 2: Finding Information Gain of each attribute

Age					
< 21		21-35		> 35	
Yes	No	Yes	No	Yes	No
2	3	4	0	3	2

$$H(age_{<21}) = -\frac{2}{5}\log_2\left(\frac{2}{5}\right) - \frac{3}{5}\log_2\left(\frac{3}{5}\right) = 0.971$$

$$H(age_{21-35}) = -\frac{4}{4}\log_2\left(\frac{4}{4}\right) - \frac{0}{4}\log_2\left(\frac{0}{4}\right) = 0.$$

$$H(age_{>35}) = -\frac{3}{5}\log_2\left(\frac{3}{5}\right) - \left(\frac{2}{5}\right)\log_2\left(\frac{2}{5}\right) = 0.971$$

$$\therefore \text{Information Gain (age)} = -\frac{5}{14} \times H(age_{<21}) + \frac{4}{14} \times H(age_{21-35}) + \frac{5}{14} \cdot H(age_{>35})$$

$$= \frac{5}{14} \times 0.971 + 0 + \frac{5}{14} \times 0.971$$

$$\text{Info (age)} = 0.693$$

$$\text{Gain(Age)} = E(S) - \text{Info(age)} = 0.94 - 0.693 = 0.247$$

Income					
High		Low		Medium	
Yes	No.	Yes	No	Yes	No.
2	2	3	1	4	2

Similarly,

$$\text{Info (Income}_{high}) = -\frac{2}{4}\log_2\left(\frac{2}{4}\right) - \frac{2}{4}\log_2\left(\frac{2}{4}\right) = 1$$

Similarly, $Info(Income_{Low}) = 0.811$, $Info(Income_{Medium}) = 0.918$,
$Info(Income) = 0.911$.

$Gain(Income) = E(S) - Info(Income) = 0.94 - 0.911 = 0.0291$

Gender			
Male		Female	
Yes	No	Yes	No
3	4	6	1

$\therefore Info(Gender_{male}) = 0.985$, $Info(Gender_{female}) = 0.592$, $Info(Gender) = 0.788$

$Gain(Gender) = E(S) - Info(Gender) = 0.94 - 0.7885 = 0.1515$.

Martial Status			
Single		Married	
Yes	No	Yes	No
5	2	4	3

$\therefore Info(M.S._{single}) = 0.863$, $Info(M.S._{marital}) = 0.985$

$Info(M.S.) = 0.924$

$Gain(Martial\ Status) = E(S) - Info(M.S.) = 0.94 - 0.924$
$= 0.016$.

Attributes	Information Gain
Age	0.247
Income	0.0291
Gender	0.1515
Martial Status	0.016

Since, the attribute "Age" has the highest Information Gain, it is selected as root node.

Sudesh Pawar
BE Comp A
24

A3 K Means

1) Why does K-Means clustering algorithm use only Euclidean Distance metric?

Ans: i) The K-means uses a vector quantization method often used as a clustering method that doesn't explicitly use pairwise distance data points at all.

ii) It amounts to repeatedly assigning pts. to closest centroid thereby using Euclidean distance from data pts. to a centroid.

iii) Euclidean distance is used ~~to~~ because the sum of squared deviations from centroid is equal to sum of pairwise squared euclidean distances divided by the no. of points.

iv) The term "centroid" is itself from euclidean geometry.

Q.2) Explain K-means algorithm with example.

Ans. Eg: Given,

$P_1 = [0.1, 0.6]$, $P_2 = [0.15, 0.71]$, $P_3 = [0.03, 0.9]$, $P_4 = [0.16, 0.85]$
$P_5 = [0.2, 0.3]$, $P_6 = [0.25, 0.5]$, $P_7 = [0.24, 0.1]$, $P_8 = [0.3, 0.2]$

M_1 (centroid) $\Rightarrow [0.1, 0.6]$ (C_1 centroid)
M_2 $\Rightarrow [0.3, 0.2]$ (C_2 centroid)
C_1 & C_2 are cluster.

x	y	$\sqrt{(x-0.1)^2+(y-0.6)^2}$	$\sqrt{(x-0.3)^2+(y-0.2)^2}$	Cluster.
0.1	0.6	0	0.45	C_1
0.15	0.71	0.12	0.53	C_1
0.08	0.9	0.30	0.73	C_1
0.16	0.85	0.25	0.66	C_1
0.2	0.3	0.31	0.14	C_2
0.25	0.5	0.18	0.30	C_1
0.24	0.1	0.51	0.11	C_2
0.3	0.2	0.44	0	C_2

$C_1 = [P_1, P_2, P_3, P_4, P_6]$

$= \{(0.1, 0.6), (0.08, 0.9), (0.16, 0.85), (0.25, 0.5)\}$

$C_2 = [P_5, P_7, P_8]$

$= \{(0.2, 0.3), (0.24, 0.1), (0.3, 0.2)\}$

$M_1 = \left(\frac{0.1+0.08+0.16+0.25}{4}, \frac{0.6+0.9+0.85+0.5}{4}\right)$

$= (0.15, 0.71)$

$M_2 = \left(\frac{0.2+0.24+0.3}{3}, \frac{0.3+0.1+0.2}{3}\right)$

$= (0.25, 0.2)$

X	Y	$\sqrt{(X-0.15)^2+(Y-0.71)^2}$	$\sqrt{(X-0.25)^2+(Y-0.2)^2}$	Cluster
0.1	0.6	0.12	0.43	C_1
0.15	0.71	0	0.52	C_1
0.08	0.9	0.20	0.72	C_1
0.16	0.85	0.14	0.66	C_1
0.2	0.3	0.41	0.11	C_2
0.25	0.5	0.23	0.3	C_1
0.24	0.1	0.62	0.10	C_2
0.3	0.2	0.53	0.05	C_2

$\therefore C_1 = \{P_1, P_2, P_3, P_4, P_6\}$ (Pts. in Cluster 1)

$C_2 = \{P_5, P_7, P_8\}$ (Pts in Cluster 2)

Q.3) Write down any 2 applications of K-means in details:

Ans. K-means application are:

1) Vector quantization:

- K-means originated from signal processing & still finds use in the domain. For example, in computer graphics, color quantization is task of reducing color palette of an image to a fixed no. of color K.
- The K-means algorithm can be easily used for this tasks & produces competitive results.
- A use case for this approach is image segmentation.
- Others uses of vector quantization include non-random sampling, as K-means can easily be used to choose K different but prototypical objects from large data sector for further analysis.

2) Feature learning:

- K-Means clustering has been used as a feature learning step for either supervised or unsupervised learning.
- The basic approach is first to train a K-means clustering representation, using training data

Sudesh Pawar
BE Comp A
24

A4 - KNN

Q.1) What are the applications of KNN?

Ans: Classification and Interpretation
- Legal, Medical, News, Banking

- Problem Solving:
  - Planning, Pronunciation
- Function Learning:
  - Dynamic Control
- Teaching & Aiding:
  - Help Desk, User Training
- Used to get missing values & in pattern recognition.

Q.2) What is sklearn neighbours & its functions?

A. Sklearn neighbours, provides functionality for unsupervised & supervised neighbors based learning methods.

- The principle behind nearest neighbour method is to find a predefined no. of training samples closest in distance to the new point & predict the label from these
- The no. of samples can be ~~user~~ user-defined constant or vary based on local density of points
- Neighbors based methods are known as non-~~generating~~ generalizing ML methods since they simply "remember all of its training data".
- Functions of sklearn neighbours:

a) fit(x,y): Fit model using X as training data & y as target values.

b) get_params(deep=true): get parameters for estimator.

c) kneighbors(x, n_neighbors=None, return_distance=True)
Finds the k-neighbours of a point.
Return distance.

d) Predict(x): Predict class labels for provided data.

e) Score(x,y): Returns mean accuracy on given test data & labels.

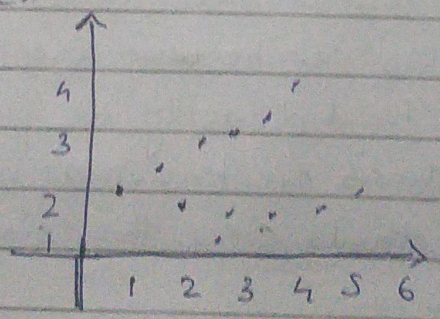
f) Set-param(**params): set parameter of estimator.

Q.3) Explain Distance weighted K-NN with example.

A. weighted KNN is a modified version of KNN. One of many issues that affect performance of KNN algorithm is the choice of hyperparameters

- This method follows rule of taking majority note but this can be a problem if nearest neighbors vary widely in their distance & closest neighbors more reliably indicate the class of project
- The idea is to weight the contribution of each of the k-neighbors according to their distance to the query. So, the closer the neighbor the more important it is

Example:



The shaded labels indicate the class o pts. & the lined labels indicate class points