Assignment 1

1) Explain the need for data normalization in K-NN. What are the benefits and limitations of K-NN algorithm. Explain the difference b/w
i) regression and correlation.
ii) regression and causation
iii) linearity and non-linearity relationships.

→ Data normalization in K-NN several reasons.

1) Equalization of Features :- In K-NN, the algorithm calculates distances b/w data points to determine similarity. if the features have different scales, those with larger scales will dominate the distance calculation.

2) Improved Convergence :- Normalizing the data helps in faster convergence during training. since the scale of features is consistent, the algorithm can converge more efficiently towards a solution

3) Avoiding Bias :- without normalization, features with larger scale might bias the model.

4) Efficient Computation :- Normalization can lead to faster computation times since the algorithm doesn't have to deal with varying scales. This is especially important in large datasets where computation time can become a bottleneck.

5) Distance metrics :-

many distance metrics, such as Euclidean distance or Manhattan distance, are sensitive to the scale of the features.

(7) Benefits of k-NN algorithms

1) Simplicity :- k-NN is easy to understand and implement. It doesn't require any assumptions about the underlying data distribution

2) No Training phase :- k-NN does not require a training phase. it simply stores the training data and makes predictions based on the similarity of new instances to the existing data.

3) Versatility :- k-NN can be used for both classification and regression tasks. It can handle both numerical and categorical data.

Limitations of K-NN algorithm

- Computationally expensive
- Sensitivity to noise and irrelevant features.
- Need for optimal K.
- Distance metric selection

i) regression and correlation.

Correlation	Regression
• Correlation measures the degree of relationship b/w two variables	• Regression is about how one variable affects the other.
• To find the numerical value that defines and shows the relationship b/w two variables	• To estimate the value of a random variable based on the values of fixed variables
• It doesn't capture causality but the degree of interrelation b/w two variables	• It is based on causality and shows no degree of connection but cause and affect.
• The coefficients are generally relative measures	• The coefficient is generally an absolute measure.

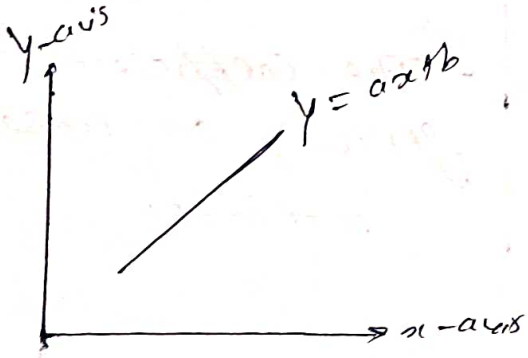
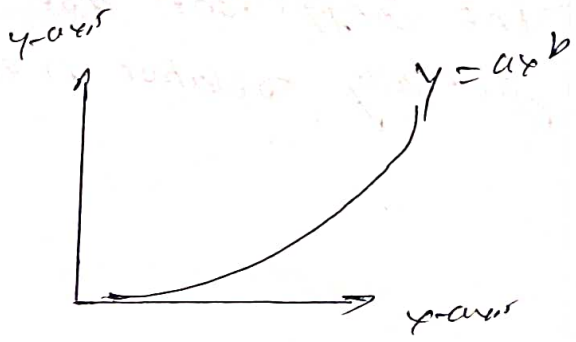
(ii) regression and causation

→ Although regression analysis deals with the dependence of one variable on other variable it does not necessarily imply causation

→ A statistical relationship per se cannot logically imply causation

→ Regression provides only statistical relationship

(iii) linearity and non-linearity relationships

Linearity Relationships	non-linearity relationships
→ The linearity relationships b/w the variables means the relationship b/w the dependent and independent variables can be visualized as a straight line.	→ A non-linearity relationship exists in functions such as exponential function and power function
→ $y = ax + b$	ex: $y = ax^b$, $y = \frac{x}{ax+b}$
	

2) Explain an algorithm for Radial Basis Function Neural network. 7

→ Input : input Vector $(x_1, x_2, \ldots, x_n)$

output : Y_p

Assign random weights for every connection from the Hidden layer the output layer in the network in the range $[-1, +1]$.

Forward phase:

Step 1: Calculate input and output in the input layer:

$$I_i = x_i$$

I_i = input node in input layer

x_i = input received at node i

output at node i 'O_i' in the input layer is

$$\boxed{O_i = I_i}$$

Step 2: For each node j in the Hidden layer, find the centre/receptor C and the variance σ.

Define hidden layer neurons with Gaussian RBF whose output is

$$\mu_j(x) = e^{\frac{-(x-C_j)^2}{\sigma^2}}$$

x → the input

C_j → centre

σ → radius

Step 3:- For each node k in output layer, Compute linear weighted sum of the output of each neuron k from the hidden layer neurons j.

$$F_k(x) = \sum_{j=1}^{m} w_{jk} H_j(x)$$

where,

w_{jk} is the weight in the link from the Hidden layer neuron j to the output layer neuron k.

$H_j(x)$ is the output of a Hidden Layer neuron j for an input vector x.

Backward Phase:

Step 1: Train the Hidden layer using Back propagation.

Step 2: update the weights b/w the Hidden layer and output layer.

3) Consider the training dataset shown below. Discretize the continuous attribute 'Percentage'

9

Sl.No	Percentage	Award	Discretize
1	95	Yes	>72
2	80	Yes	>72
3	72	No	≤72
4	65	Yes	≤72
5	95	Yes	>72
6	32	No	≤72
7	66	No	≤72
8	54	No	≤72
9	89	Yes	>72
10	72	Yes	≤72

→ First sort the values in an ascending order

32	54	65	66	72	72	80	89	95	95

Remove the duplicates and consider only the unique values of the attribute.

32	54	65	66	72	80	89	95

10

Now, compute the Gain for the distinct values of this continuous attribute.

(a) Gain Values for CGPA

	32		84		65		66		72		80		89		95	
Range	≤	>	≤	>	≤	>	≤	>	≤	>	≤	>	≤	>	≤	>
Yes	0	6	0	6	1	5	1	5	2	4	3	3	4	2	6	0
No	1	3	2	2	2	2	3	1	1	0	4	0	4	0	4	0
Entropy	0	0.918														
Entropy-info (S,T)	0.8846		0.6440		0.8796		0.71052		0.55047		0.6846		0.8		0.97095	
Gain	0.14057		0.3219		0.0912		0.2564		0.41946		0.2812		0.17094		0	

$$\text{Entropy-info}(T, \text{Award}) = +\left[\frac{5}{10}\log_2\left(\frac{6}{10}\right) + \frac{4}{10}\log_2\left(\frac{4}{10}\right)\right]$$

$$= 0.97095$$

$$\text{Entropy}(6,3) = +\left[\frac{6}{9}\log_2\left(\frac{6}{9}\right) + \frac{3}{9}\log_2\left(\frac{3}{9}\right)\right]$$

$$= 0.91829.$$

$$\text{Entropy-info}(T, \text{Percentage} \in 32) = \frac{1}{10} \times 0 + \frac{9}{10} \times 0.9182$$

$$= 0.82638$$

$$\text{Gain}(T, \text{Percentage}) = 0.97095 - 0.82638$$

$$= 0.14457$$

(4) Entropy-info (T, P(54)) $= \frac{2}{10} \times$ Entropy (0, 2) $+ \frac{8}{10}$ Entropy (6, 2)

$= 0.6440$

Gain $= 0.97044 - 0.6440 = 0.3219$

(5) Entropy-info (T, P(61)) $= \frac{3}{10} \times$ entropy (1, 2) $+ \frac{7}{10}$ Entropy (5, 2)

$= 0.87967$

Gain $= 0.97044 - 0.87967$

$= 0.0912$

(6) Entropy-info (T, P(66)) $= \frac{4}{10} \times$ Entropy (1, 3) $+ \frac{6}{10}$ Entropy (5, 1)

$= 0.71452$

Gain $= 0.97044 - 0.71452 = 0.2564$

(7) Entropy-info (T, P(70)) $= \frac{6}{10} \times$ Entropy (2, 4) $+ \frac{4}{10} \times$ entropy (4, 0)

$= 0.55047$

Gain $= 0.97044 - 0.55047 = 0.41996$

(8) Entropy-info (T, P(80)) $= \frac{7}{10} \times$ Entropy (3, 4) $+ \frac{3}{10}$ Entropy (3, 0)

$= 0.6846$

Gain $= 0.97044 - 0.6846 = 0.2812$

(9) Entropy-info (T, P(89)) $= \frac{8}{10} \times$ Entropy (4, 4) $+ \frac{2}{10}$ Entropy (2, 0)

$= 0.8$

Gain $= 0.97044 - 0.8 = 0.17044$

(10) Entropy-info (T, P(95)) $= \frac{10}{10} \times$ Entropy (6, 4) $+ 0$

$= 0.47045 - 0.97045$

$= 0$

→ The maximum gain as 0.41446 being percentage 72

It categories with ≤ 72 and > 72

12

4] Construct a regression tree for the following data set.

S.NO	Assessment	Assignment	Project	Result (%)
1.	Good	yes	yes	95
2.	Average	yes	No	70
3.	Good	No	yes	75
4.	Poor	No	No	45
5.	Good	yes	yes	98
6.	Average	No	yes	80
7.	Good	No	No	75
8.	Poor	yes	yes	65
9.	Average	No	No	58
10	Good	yes	yes	89

Step 1: Compute standard deviation for each attribute with respect to the target attribute

$$\text{average} = (45+70+75+45+98+80+75+65+58+89)/10 = 75$$

$$S.D = \sqrt{\frac{(45-75)^2 + (70-75)^2 + (75-75)^2 + (45-75)^2 + (98-75)^2 + (80-75)^2 + (65-75)^2 + (58-75)^2 + (89-75)^2}{10}}$$

$$= 16.55$$

Assessment = Good

S.No	Assessment	assignment	project	Result (%)
1.	Good	Yes	Yes	95
3.	Good	No	Yes	75
5.	Good	Yes	Yes	48
7.	Good	No	No	75
10.	Good	Yes	Yes	89

Average $= (95 + 75 + 48 + 75 + 89) / 5 = 86.4$

$$S.D = \sqrt{\frac{(95-86.4)^2 + (75-86.4)^2 + (48-86.4)^2 + (75-86.4)^2 + (89-86.4)^2}{5}}$$

$= 10.9$

Assessment = Average

S.No	Assessment	Assignment	Project	Result (%)
2.	Avg	Yes	No	70
6.	Avg.	No	Yes	80
9.	Avg	No	No	58

Average $= (70 + 80 + 58) / 3 = 69.3$

$$S.D = \sqrt{\frac{(70-69.3)^2 + (80-69.3)^2 + (58-69.3)^2}{3}}$$

$= 11.01$

14

Assessment = Poor

S.No	Assessment	Assignment	Project	Result (%)
6.	Poor	No	No	46
8.	Poor	Yes	Yes	65

Average $= (46 + 65)/2 = 55$

$$S.D = \sqrt{\frac{(46-55)^2 + (65-55)^2}{2}} = 14.14$$

Standard deviation for assessment

Assessment	S.D	Data instances
Good	10.9	5
Average	11.01	3
Poor	14.14	2

weighted S.D for Assessment $= \left(\frac{5}{10}\right) \times 10.9 + \left(\frac{3}{10}\right) \times 11.01 + \left(\frac{2}{10}\right) \times 14.14$

$= 11.58$

S.D reduction for Assessment $= 16.65 - 11.58 = 4.97$

Assignment = yes

S.No	Assessment	Assignment	Project	Result(%)
1.	Good	yes	yes	95
2.	Average	yes	no	70
5.	Good	yes	yes	98
8.	Poor	yes	yes	65
10.	Good	yes	yes	89

Average = (95 + 70 + 98 + 65 + 89)/5 = 83.4

$$S.D = \sqrt{\frac{(95-83.4)^2 + (70-83.4)^2 + (98-83.4)^2 + (65-83.4)^2 + (89-83.4)^2}{5}}$$

= 14.98

Assignment = No

S.No	Assessment	Assignment	Project	Result(%)
3.	Good	No	yes	75
4.	Poor	No	no	45
6.	Average	No	yes	80
7.	Good	No	no	75
9.	Average	No	no	58

Average = (75 + 45 + 80 + 75 + 58)/5 = 66.6

$$S.D = \sqrt{\frac{(75-66.6)^2 + (45-66.6)^2 + (80-66.6)^2 + (75-66.6)^2 + (58-66.6)^2}{5}}$$

= 14.7

16 Standard Deviation for Assignment

~~Assessment~~

Assignment	S.D	Data instances
Yes	14.98	5
No	14.7	5

weighted S.D for assignment $= \left(\frac{5}{10}\right) \times 14.98 + \left(\frac{5}{10}\right) \times 14.7$

$= 14.84$

S.D reduction for assignment $= 16.54 - 14.84 = 1.71$

Project = Yes

S.No	Assessment	Assignment	Project	Result (%)
1.	Good	Yes	Yes	95
3.	Good	No	Yes	75
5.	Good	Yes	Yes	98
6.	Average	No	Yes	80
8.	Poor	Yes	Yes	65
10.	Good	Yes	Yes	89

Average $= (45 + 75 + 48 + 80 + 65 + 84)/6 = 83.7$

S.D $= 12.6$

(7) Project = No

S.No	Assessment	Assignment	Project	Result (%)
2.	Average	Yes	No	70
4.	Poor	No	No	45
7.	Good	No	No	75
9.	Average	No	No	58

Average = $(70+45+75+58)/4 = 62$

$$S.D = \sqrt{\frac{(70-62)^2 + (45-62)^2 + (75-62)^2 + (58-62)^2}{4}}$$

$= 13.39$

Standard deviation for project

Project	S.D	Data instances
Yes	12.6	6
No	13.39	4

Weighted S.D for project $= \left(\frac{6}{10}\right) \times 12.6 + \left(\frac{4}{10}\right) \times 13.39$

$= 12.92$

S.D reduction for project $= 16.55 - 12.92 = 3.63$

Attributes	S.D Reduction
Assessment	4.47
Assignment	1.71
Project	3.63

Assessment

Good

S.No	Assesment	Assignet	Projet	Result
1.	Good	Yes	Yes	95
3.	Good	No	Yes	75
5.	Good	Yes	Yes	98
7.	Good	No	No	75
10.	Good	Yes	Yes	84

Poor

S.No	Assesment	Assignment	Projet	Re
4.	Poor	No	No	4
8.	Poor	Yes	Yes	6

Average

S.No	Assessment	Assignment	Projet	Result
2.	Average	Yes	No	70
6.	Average	No	Yes	80
9.	Average	No	No	58

Regression Tree with assessment as Root node

5) illustrate Bayes Theorem and maximum posterior hypothesis. The following dataset gives information about stolen vehicles. using Naive Bayes classifier classify the new data (Red, SUV, Domestic)

Color	Type	origin	Stolen
Red	Sports	Domestic	yes
Red	Sports	Imported	yes
Red	SUV	Imported	No
yellow	Sports	Domestic	No
yellow	SUV	Imported	yes
yellow	Sports	Domestic	yes
Red	SUV	Imported	No

(4) Bayes Theorem:-

Bayes theorem is also known as Bayes' rule, Bayes' law or Bayesian reasoning, which determines the probability of an event with uncertain knowledge.

In probability theorem ~~was~~ it relates the conditional probability and marginal probabilities of two random events

20

→ The Bayesian inference is an application of Bayes' theorem which is fundamental to Bayesian statistics.

It is way to calculate the value of $P(B|A)$ with the knowledge of $P(A|B)$

→ Bayes' theorem allows updating the probability of an observing new information of the real world.

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

$P(A|B)$ → Posterior

$P(B|A)$ → likelihood, in which we consider that hypothesis is true, then we calculated the probability of evidence.

$P(A)$ → prior probability.

$P(B)$ → marginal probability

(7) Maximum Posterior hypothesis

→ A maximum a posterior probability (MAP) estimate is an estimate of an unknown quantity, that equals the mode of the posterior distribution.

→ The posterior distribution, $f_{X|Y}(x|y)$ (or $P_{X|Y}(x|y)$), contains all the knowledge about the unknown quantity x. Therefore, we can use the posterior distribution to find point or interval estimates of x.

To find the MAP estimate, we need to find the value of x that maximizes

$$f_{X|Y}(x|y) = \frac{f_{Y|X}(y|x)\, f_X(x)}{f_Y(y)}$$

→ $f_Y(y)$ does not depend on the value of x. Therefore, we can equivalently find the value of x that maximizes.

$$f_{Y|X}(y|x)\, f_X(x).$$

22

(7) Naive Bayes Classifier classify the new data

(Red, SUV, Domestic)

$P(yes) = \frac{4}{7} = 0.57144$

$P(No) = \frac{3}{7} = 0.42857$

Color	Y	N
Red	2/4	2/3
yellow	2/4	1/3

Type	Y	N
Sports	3/4	1/3
SUV	1/4	2/3

origin	Yes	No
Domestic	2/4	1/3
imported	2/4	2/3

$P(stolen/yes) = \frac{4}{7} \times \frac{2}{4} \times \frac{1}{4} \times \frac{2}{4}$

$= 0.035714$

$P(stolen/No) = \frac{3}{7} \times \frac{2}{3} \times \frac{2}{3} \times \frac{1}{3}$

$= 0.06349$

$$P(yes) = \frac{0.0357}{0.0357 + 0.0634} = 0.36024$$

$$P(No) = \frac{0.0634}{0.0634 + 0.0357} = 0.6397$$

$$P(yes) + P(No) = 1$$

$$0.36024 + 0.6397 = 1$$

$$1 = 1$$

Stolen = "NO"

6) Explain the role of probability distribution in ML.

* Probabilistic models play a crucial role in the field of machine learning, providing a framework for understanding the underlying patterns and complexities in massive datasets.

1) Capturing uncertainty :- Probabilistic models recognize that real-world data is inherently uncertain. Instead of making predictions based on absolute values, they operate with probability distributions.

2) Bayesian Inference :-

→ One common approach in probabilistic modeling is Bayesian Inference. It combines prior knowledge with observed data to make predictions.

3) Maximum likelihood estimation (MLE) :-

→ MLE seeks to maximize the likelihood of the observed data given the model parameters. It's particularly useful for parameter estimation.

4) Generative Models :-

→ These models aim to capture the joint distribution of input variables. They generate new data based on the probability distribution of the original dataset.

5) Discriminative Models :-

→ Unlike generative models, discriminative models focus on modelling the conditional distribution of the output variable given the input. They learn decision boundaries that separate different classes. Image recognition, speech recognition, and sentiment analysis often use discriminative models

6) Graphical Models :-

→ These models represent conditional dependencies b/w variables using graphs. They're valuable for tasks like image recognition, natural language processing and causal inference.

7) Naive Bayes algorithm :-

→ A widely used probabilistic approach, the Naive Bayes algorithm, simplifies independence assumptions. it efficiently calculate the probability of a target class given the feature set.