Assignment - 2

Q1) What is Regression?

→ • Regression analysis is a form of predictive modelling technique which investigates the relationship between a dependent variable & independent variable.

• A regression analysis involes graphing line over a set of data points that most closely fit overall shape of the data or regression.

• Regression is a supervised learning technique which helps in finding the correlation between variables & enables us to predict the continuous output variable based on one or more predictor variables.

Q-2) Define linear regression & explain its types.

→ • Linear regression is a method to predict dependent variable (Y) based on values of independent variables (x).

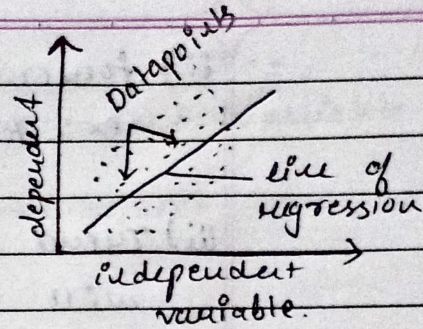
• It can be used for the cases where we want to predict some continuous quantity.

• only one independent variable x.

• Relation between x & Y defined by a linear function.

Date ______
Page ______

- Dependent variable (Y):
who's value need to be predicted.



- Independent variable (X):
used to predict the response variable.

- outliers: is observation which contain either very low value or very high value in comparison to other observed values.

→ Types of Regression Analysis:

1) Simple linear Regression: if a single independent variable is used to predict the value of a numerical dependent variable, ~~the~~ such a linear regression algorithm is called simple linear regression.

2) Multi linear regression: if more than one independent variables is used to predict the value of a numerical dependent variable, then such linear regression algorithm is called multiple linear regression.

Q3) when we can use regression?

→ i) Determining the strength of predictions:
To determine the strength of the effect ~~that~~ the independent variable have on the dependent variable.
ex: relationship between sales & marketing spending.

ii) forecasting an effect:
ex: Relation between age & income.

iii) Trend forecasting:
will I get bunch of rupees spent on marketing.

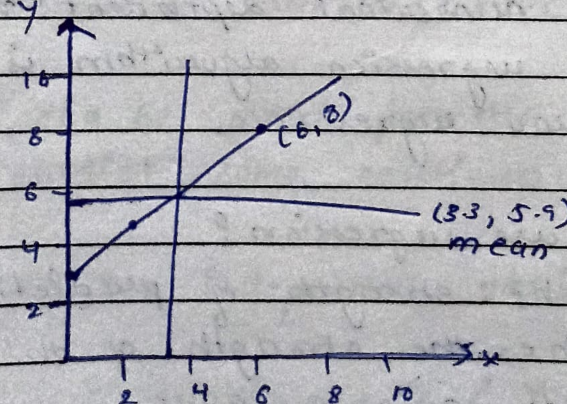
Q4) The value of x & Y given below.

X	Y
0	3
2	4
3	6
4	7
5	7.5
6	8

find least square regression line $y = mx + c$
estimate the value of Y when x is 10.

1) first we have to find mean.
$\bar{x} = 3.3 \quad \bar{y} = 5.91$

2) plot them in a graph.



3) we need to find regression of line.

line of equation $y = mx + c$

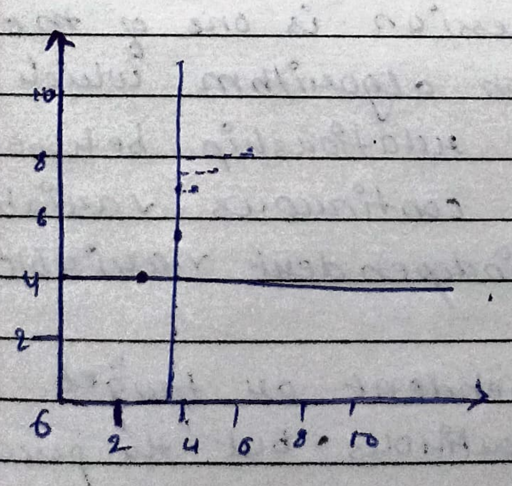
- y → Dependent variable
- x → independent variable
- m → slope
- c → intercept.

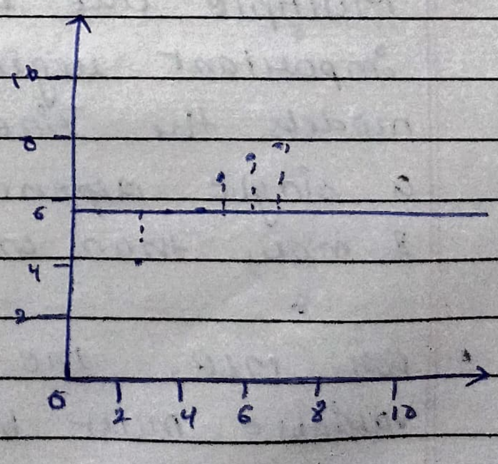
$$m = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(x-x)^2}.$$

	x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})^2$	$(x-\bar{x})\cdot(y-\bar{y})$
	0	3	-3.3	-2.9	10.89	9.57
	2	4	-1.3	-1.9	1.69	2.4
	3	6	-0.3	0.1	0.09	-0.03
	4	7	0.7	1.1	0.49	0.77
	5	7.5	1.7	1.6	2.89	2.72
	6	8	2.7	2.1	7.29	5.67
Mean	3.3	5.9			3.8	3.5

$(x-\bar{x})$ is distance of all the point through the line y equal 3.3 ($y=3.3$)

$(y-\bar{y})$ is distance of all the point through the line x equal 5.9 ($x=5.9$).





Date ______
Page ______

$$m = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(x-\bar{x})^2}$$

$$= \frac{3.5}{3.8}$$

$$m = 0.9$$

so, in eq. $y = mx + c$

where $y = 5.9$

$m = 0.9$

$x = 3.3$

$$5.9 = (0.9)\,3.3 + c$$

$$5.9 = 2.9 + c$$

$$c = 5.9 - 2.9$$

$$c = 3$$

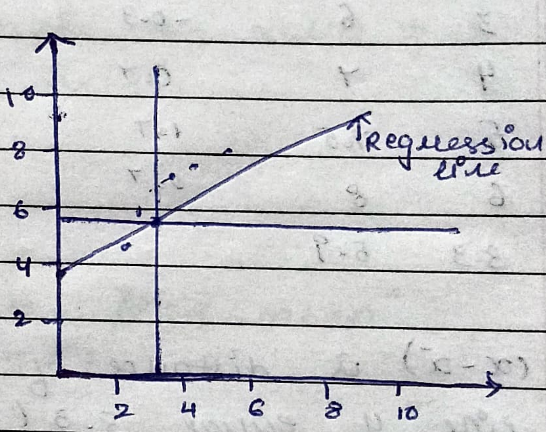
ii) y at $x = 10$

$$y = mx + c$$

$$y = 0.9 \times 10 + 3$$

$$y = 9 + 3$$

$$y = 12.$$



Q5) What is Multiple Regression?

→ Multiple line regression is one of the important regression algorithm which models the linear relationship between a single dependent continuous variable & more than one independent variable

- For MLR, the dependent or target variable must be continous but the predictor or independent variable may be of continuous or categorical form.

- Each feature variable must model the linear relationship with the dependent variable.

- MLR tries to fit a regression line through a multidimensional space of data-points.

Assignment - 3

Q1) What is classification? Explain types of classification.

Classification: classification algorithms are used when the output variable is categorical which means there are two classes such as Yes-No, Male-female, True-false, etc.

- There are 4 types of classification:

1) Binary Classification: It refers to those classification tasks that have two class labels.

Ex: Email Spam detection (Spam or not)

Binary classification tasks involve one class that is ~~it takes~~ the normal state & another class that is the abnormal state.

- Popular classification that can be used for binary classification include:
  - Logistic Regression
  - K-Nearest Neighbors
  - Decision Trees
  - Support vector machine
  - Naive Bayes.

2) Multi-class classification: when we have more than 2 classes for classes/category then it is multi-class classification.

It include algorithm :
- KNN, Random forest
- Decision Trees, Gradient booting
- Naive Bayes.

- Lots model n lots of data ~~read~~ needs to be feeded for accurate result.
- Multi class classification does not have the notion of normal & abnormal outcomes.

Ex:-
1 pic is given with an animal & we need to find that it is from which category like cat, dog, cow etc.

3) Multi-label classification :
It refers to those classification task that have two or more class label, where one or more class labels may be predicted for each example.
Ex: photo classification.
where more than one type of class label object are there
ex: bicycle, apple, etc.

- In binary & multi-class classification single class label is predicted for each example.
- Specialized version of standard, classification algorithm can be used,
  → multi-label Decision tree
  → multi-label Random forests
  → Multi-label Gradient Boosting.

Date
Page

4) Imbalanced classification:

It refers to classification tasks where the no. of examples in each class is unequally distributed.

ex:

Where the majority of ex in the training dataset belong to the normal class & a minority of ex. belong to the abnormal class.

like fraud detection
outlier detection
medical diagnostic tests.

Q2) Explain logistic Regression with ex:

- Logistic Regression is one of the most popular ML algorithm, which comes under the supervised learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables.

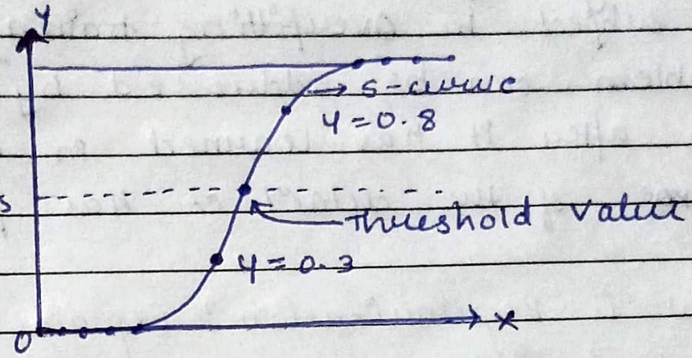
- It is used for solving Regression problem, where as logistic regression is used for solving the classification problems.

- In this instead of fitting a regression line, we fit an "s" shaped logistic function which predicts two max. value (0 or 1)

- It is a significant ML algorithm because it has the ability provide probabilities & classify new data using continuous & discrete datasets.

Date
Page

- It can be used to classify the observation using different types of data & can easily determine the most effective variable used for the classification.



- Logistic Regression uses the concept of predictive modelling as regression, therefore it is called logistic regression, but it used to classify samples, ∴ it falls under the classification algorithm.

- So, mathematically, it is model predicts $P(Y=1)$ as f^n of X. It is one of the simplest ML algorithm that can be used for various classification problems such as spam detection, Diabetes prediction, cancer detection, etc.

Q3) Explain problem of overfitting in logistic regression.

→ Overfitting refers to a model that models the training data too well.

- It happens when a model learns the detail & noise in the training data to the extent that is negatively impacts the performance of the model on new data.
- features are noisy / uncorrelated to concept.
- Modelling process very sensitive (powerful).
- Too much search.

Date
Page

Ex:
decision trees are a nonparametric machine learning algorithm that is very flexible & is subject to overfitting training data. This problem can be addressed by pruning a tree after it has learned in order to remove some of the detail it has picked up.

Q4) What is Regularization? Explain types of Regularization.

- Regularization is any modification we make to a learning algorithm that is intended to reduce its generalization error but not its training error.

- Regularization can be used to train models that generalize better on unseen data, by preventing the algorithm from overfitting the training dataset.

- Regularization Techniques.
There are two types of techniques, Ridge Regression & Lasso Regression.
1) Ridge Regression (L_2 Regularization):
- This regularization technique prefers L_2 regularization. It modifies the RSS by adding the penalty (shrinkage quantity) equivalent to the square of magnitude of coefficients.

- Ridge Regression penalize the size of the regression coefficients based on their l^2 norm:

Date
Page

$$\text{Origin } \beta \sum_i (y_i - \beta' x_i)^2 + \lambda \sum_{x=1} \beta_K^2$$

The tuning parameter serves λ to control the relative impact of those two terms on the regression coefficient estimates.

2) Lasso Regression (L1):
It is a type of linear regression that uses shrinkage. Shrinkage is where data values are shrunk towards a central point, like the mean. The Lasso procedure encourages simple, sparse models.

$$L_{lasso}(\beta) = \sum_{i=1}^{n} (y_i - x_i^T \hat{\beta})^2 + \lambda \sum_{j=1}^{m} |\hat{\beta}_j|$$

Q5) What is support vector machine algorithm? Explain with example.

→ • "Support Vector Machine" is a supervised machine learning algorithm that can be used for both classification or regression challenges.

• In SVM algorithm, we plot each data item as a point in n-dimensional space with the value of each feature being the value of a particular coordinate.

• The SVM classifier is a frontier that best segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future.

Date
Page

Ex: It can used in KNN classifier.

→ Suppose, we see a strange cat that also has some features of dogs, so if we want a model that can accurately identify whether it is cat or dog, so such a model can be created by using the SVM algorithm.

- we will train our model with lots of images of cats & dogs so that it can learn about diff^n features of cats & dogs, & then we test it with this strange creature.

- SVM algorithm can be used for face detection, image detection, text categorization etc.

Q6) Explain KNN algorithm with example.

- • K - nearest neighbours algo. is type of supervised ML algo which can be used for classification as well as regression predictive problems.

- • The KNN algo. assumes that similar things exist in close proximity.

- • KNN is a non-parametric algo, which means it does not make any assumption on underlying data.

- • It is also called lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset & at the time of classification, it performs an action on the dataset.

Date
Page

→ **Algo:**

1) For implementing any algo, we need dataset so during the first step of KNN, we must load the training as well as test data.

2) choose the value of K i.e, the nearest data points. K can be any integer.

3) for each point in the data do →
- calculate the distance between test data & each row of training data with the help of any of method, euclidean, Hamming distance.

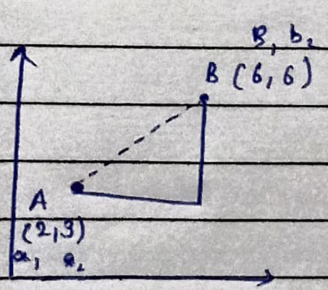
- based on the distance value, sort them ascending order.

- It will choose the top K row from the sorted array.

- Now, it will asign a class to the test point based on most frequent class of these rows.

4) End

Ex:



Euclidean distance (a,b)

$= \sqrt{(a_1-b_1)^2 + (a_2-b_2)^2}$

$= \sqrt{(2-6)^2 + (3-6)^2}$

$= \sqrt{(-4)^2 + (-3)^2}$

$= \sqrt{16+9}$

$= \sqrt{25}$

distance $= 5$.