09-02-2021 (50) KANAV BANSAL

LECTURE - 46

ADVANTAGES OF LINEAR REGRESSION:

1. Linear Regression is simple to implement and easier to interpret the output.

2. When we know the relationship between the independent and dependent variables, they have a linear relationship. This is the best algorithm to use because of it's less complexity to compared to other algorithms.

3. Linear Regression is susceptible to over-fitting but it can be avoided using some dimensionality reduction techniques, regularization techniques and cross-validation.

DISADVANTAGES OF LINEAR REGRESSION:

1. In linear Regression technique, the outliers can have huge effects on the regression and boundaries are linear.

2. Diversly, linear regression assumes a linear relationship between dependent and independent variables. That means it assumes that there is a straight-line relationship between them. It assumes independence between attributes.

3. But then linear regression also looks at a relationship between the mean of the dependent variables and independent variables. Just as the mean is not a complete description of a single variable, linear regression is not a complete description of relationships among variables

X			y
X_1	X_2	X_3	
1	100	2	
2	200	4	
3	300	6	
4	400	8	

From the table,

↳ $X_2 = X_1 * 100$ → ①

$X_3 = X_1 * 2$ → ②

From the above ① & ② eqn's we can say that they are totally dependent.

↳ So, it's better to drop the X_3 column.

↳ These type of problem is called as "Multicollinearity".

MULTICOLLINEARITY:

It is the occurance of high intercorrelations among two or more independent variables in a multiple regression model.

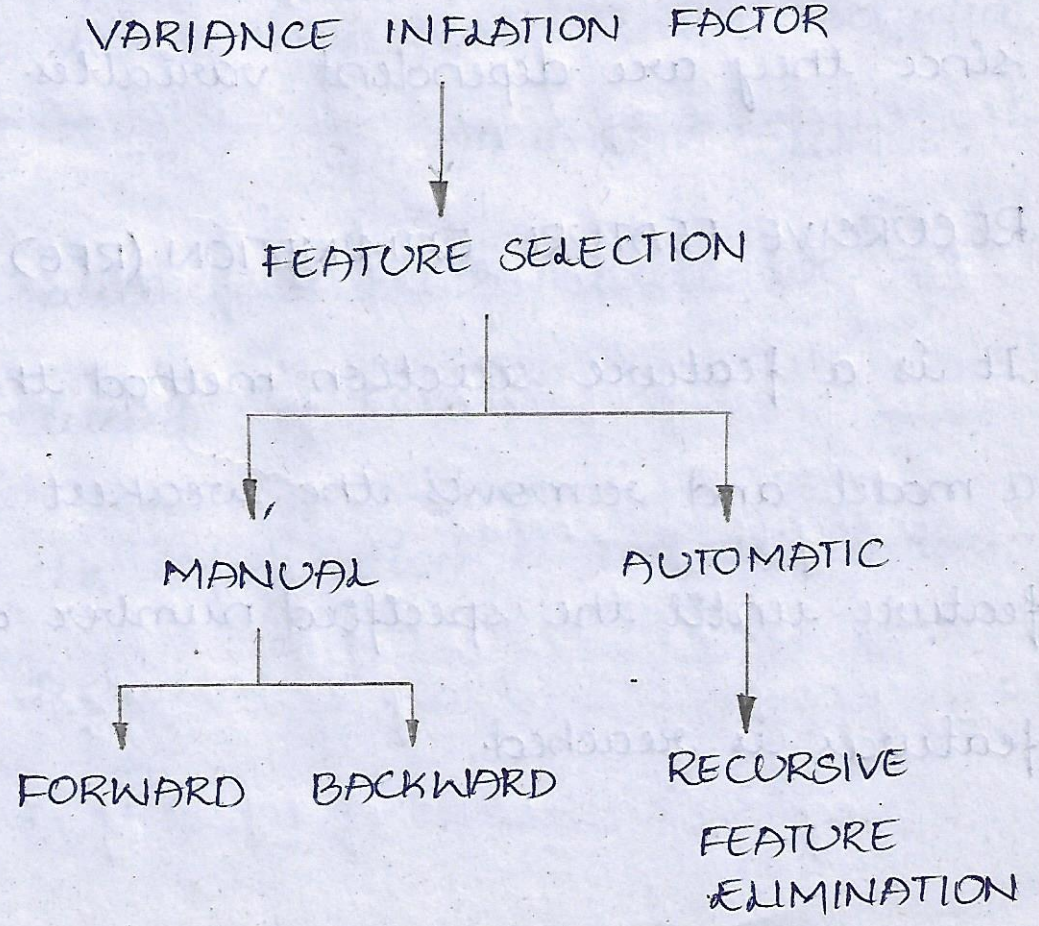
→ It occurs when our model includes multiple factors that are correlated not just to our response variable, but also to

each other.

→ In other words, it results when we have factors that are a bit redundant.

→ It is a problem because it undermines the statistical significance of an independent variable.

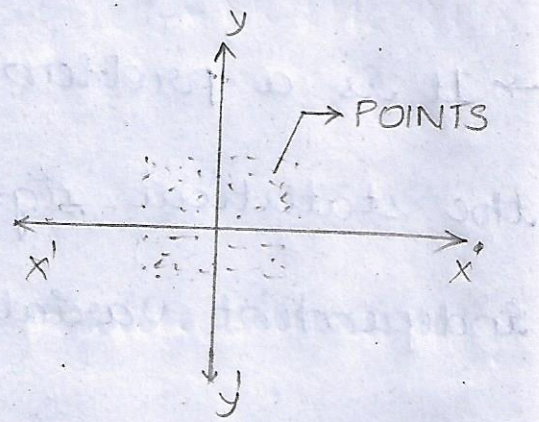
In order to solve multicollinearity Problem we us VIF (VARIANCE INFLATION FACTOR)



NOTE:

As the dimensionality increases, we can't visualize. So, we use math to solve them.

X				y
X_1	X_2	X_3	X_4	
1	2	8	24	
2	4	9	27	
3	6	10	30	



From the above table, we can say that

→ $X_2 = 2 * X_1$ & $X_4 = 3 * X_3$

So, now its better to drop X_2 & X_4 columns since they are dependent variables.

RECURSIVE FEATURE ELIMINATION (RFE):

It is a feature selection method that fits a model and removes the weakest feature until the specified number of features is reached.

RANKING :

↳ rfe.ranking → True → 1 (1st rank)

↳ False → 2, 3.... (ranks)

↳ TRUE – It gives the 1st rank to the data.

↳ FALSE – It gives the rank 2, 3.. and so on to the data.

INCLUDED :

↳ rfe.~~included~~

↳ rfe.supported → this is most important.

* It shows which features are included

i.e., TRUE – 3, 4, 5, 7, 10 → IN COLUMNS

* It also shows which features are not included i.e., FALSE.

* ↳ If the information is lost, the errors get increased.

→ As the number of features increase in the independent variables, then we use feature selection technique → MULTICOLLINEARITY.

↳ So, if we can remove multicollinearity, the model is going to be interpretable.

DIMENSIONALITY REDUCTION:

It refers to the technique that reduce the number of input variables in a dataset.

↳ Large number of input features can cause poor performance for machine learning algorithms.

↳ In general, this study is concerned with reducing the number of input features.

WHY DIMENSIONALITY REDUCTION IS IMPORTANT?

It has several advantages from a machine learning point of view.

↳ Since our model has fewer degrees of freedom, the likelihood of overfitting is lower.

↳ the model will generalize more easily to new data.

↳ If we are using feature selection, the reduction will promote the important variables.

↳ * this is a better technique.

We use this in unsupervised learning i.e., PCA (PRINCIPAL COMPONENT ANALYSIS).

CLASSIFICATION:

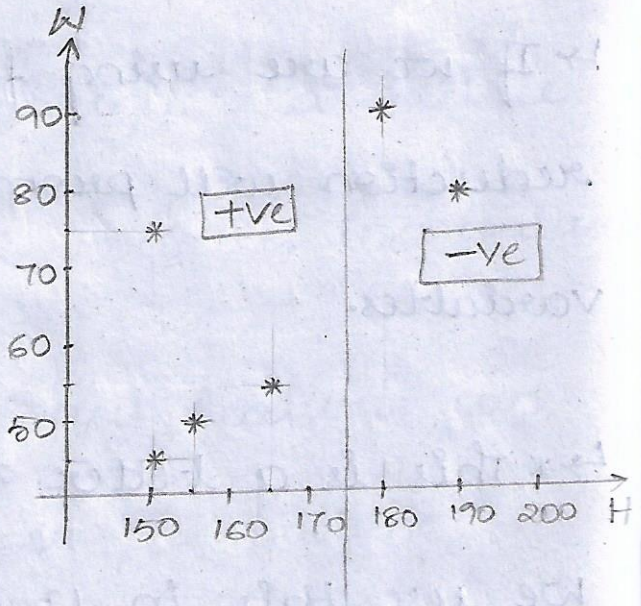
It is about predicting a label.

↳ The classification is the problem of Predicting a discrete class label output.

↳ In simple,

classification model predicts the categorical class labels.

H	W	G
150	45	F
180	90	M
190	80	M
165	65	F
155	50	F
150	75	M



TASK: To find a line that seperates the Females (+ve) class from Males (-ve)

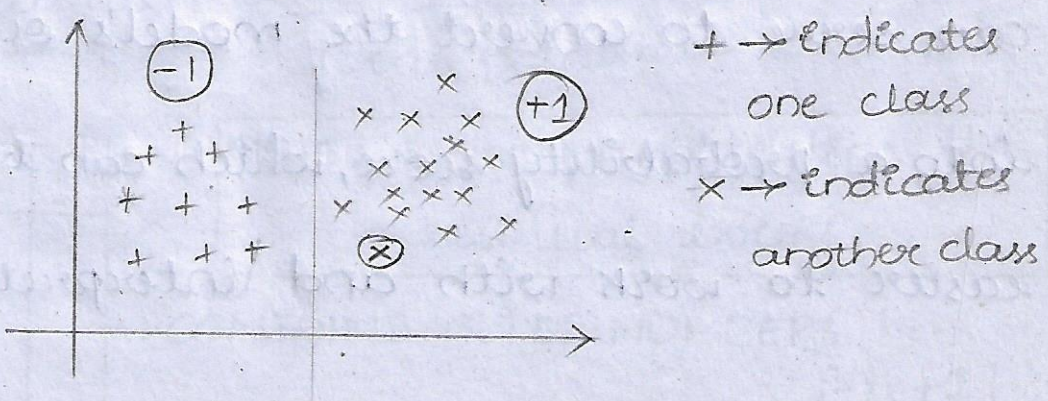
LOGISTIC REGRESSION:

→ It is also called as LOGIT REGRESSION (or) LOGIT MODEL.

↳ It is used to predict the probability of a target variable.

↳ It works with the binary data, where either the event happens or the event doesn't happens.

↳ So, the dependent variable is binary in nature having data coded as either '1' (stands for success/yes) or '0' (stands for failure/NO).



In logistic Regression, the task is:

TASK - Find a line i.e., $f(m,c)$ that best seperates +ve class from -ve class.

BEST SEPERATION - It depends upon the seperation on a hyperplane.

SIGMOID FUNCTION:

It is normally used to refer specifically to the logistic function, also called as the LOGISTIC SIGMOID FUNCTION.

↳ A sigmoid function is placed as the last layer of a machine learning model which can serve to convert the model's output into a probability score, which can be easier to work with and interpret.

So, we use a function $y_{act} * y_{pred}$

We know that,

$y_{pred} = mx + c \rightarrow w^T \cdot x$ (Also be written)

where $[w_1 \; w_0] \begin{bmatrix} x_1 \\ 1 \end{bmatrix}$

* A single dot product gives what type of classification it is.

* ↳ The procedure and steps for classification is same but in the evaluation process,

we use Accuracy, confusion matrix, Precision and recall, F1 score, ROC-AOC. respectively.

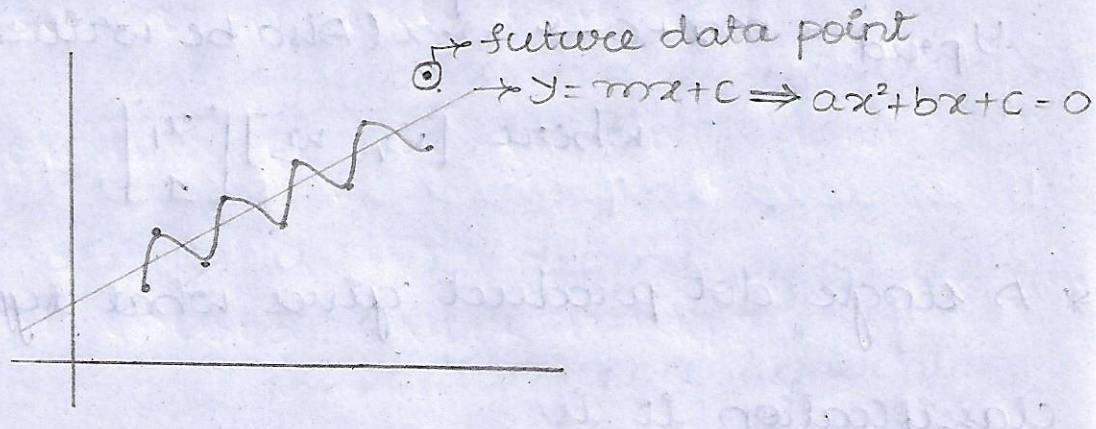
↳ $\text{Accuracy} = \dfrac{\text{correct class}}{\text{Total points.}}$

NOTE: We need not define the residual errors.

X	y

→ CONTAINS SET OF NUMBERS i.e., $[-1, +1]$

For suppose, if the graph looks like the below mentioned, then.....



If the data points have a curved pattern instead of a best fit line, then it is very complicated.

i.e., $ax^6 + bx^5 + cx^4 + dx^3 + ex^2 + fx + g = 0$ which is very complicated.

↳ The training data is preparing for a test.

So, In this case, Overfitting and underfitting comes into existance.

OVERFITTING:

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

↳ this means noise or random fluctuations in the training data is picked up and learned as the concepts by the model.

WHAT CAN CAUSE OVERFITTING?

As we know that, Overfitting happens when a model learns the detail and noise in training data to the extent that it negatively impacts the performance of the model on new data.

UNDERFITTING:

It refers to the model that can neither model the training data nor generalize to new data.

↳ An underfit machine learning model is not a suitable model and will be obvious as it will have a poor Performance on the training data.

WHAT CAUSES UNDERFITTING?

It occurs when a model is too simple. i.e., informed by too few features or regularized too much - which makes it inflexible in learning from the dataset.

↳ Simple learners tend to have less variance in their predictions but more bias towards wrong outcomes.