Revise :-

1. KNN as a Classification & Regression Algo.
2. Templates use for Classification & Reg' Algo.
3. Evaluation Metrics of Classification Algo.
   (a). Precision
   (b) Recall
   (c) Accuracy
   (d) Confusion Matrix
   (e) Error Rate
   (f) Classification Report
   (g) f1 score
4. KNN on a Regression data.

10 March 2023 19:42

Agenda :-

1. Regression Metrics in python using scikit learn.
2. K-Means Clustering Algorithm.
3. Implementation of k Mean in python using Sklearn.

10 March 2023 19:44

* Regression Evaluation Metrics

1. Mean Absolute Error
2. Mean Squared Error
3. Root Mean Squared Error

}

4. R^2 Error ✓ ☆
5. Adjusted R^2 Error. ✓ ☆

} Very Imp

10 March 2023 20:25

* Euclidean distance calc from Centroid Point to data points 1st row

$C3 :- (8, 4)$ (x_2, y_2) $\quad (x_1, y_1) :- (2, 4)$ (x_1, y_1)

$$ED :- \sqrt{(8-2)^2 + (4-4)^2}$$

$$= \sqrt{6^2+0^2} = \sqrt{36} = 6$$

10 March 2023 20:34

C1 :- $\underset{(x_1,y_1)}{(2,4)}$ $\underset{(x_2,y_2)}{(2,6)}$ $\underset{(x_3,y_3)}{(4,7)}$

CP :- $\left(\frac{x_1+x_2+x_3+\dots}{n}, \frac{y_1+y_2+y_3+\dots}{n}\right)$

$= \left(\frac{2+2+4}{3}, \frac{4+6+7}{3}\right)$

C1 (CP) = (2.67, 5.67)

Revision :-

1. Dataset looks in a Unsupervised ML Algo (Target is absent)
2. Solved a problem using k-Means Clustering.
3. Define a k-Means Algorithm.
4. k in k-Means represent the number of clusters.
5. Implentation of k-Means in python.

Agenda:-

1. Implementation of K-Means using sklearn in python.
2. Ideal method to select k (no. of clusters)
3. Elbow method.
5. Data preprocessing techniques.

Elbow Method :-

$K = 1 \rightarrow C1$

$K = 2 \rightarrow C1, C2$

$K = 3 \rightarrow C1, C2, C3$

$K = 4 \rightarrow C1, \ldots, C4$

$K = 5 \rightarrow C1, \ldots, C5$

$\vdots$

$K = 20 \rightarrow C1, \ldots, C20$

WCSS :- Within Cluster Sum of Squares

:- used for judging the optimal k.

Since we have check for each value of k between 1 to 20,

(1) We have use a for loop

(2) wcss = [ ]

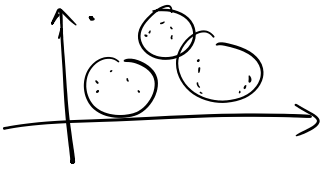
(3) loop running

calc wcss and that will be appended to wcss empty list

(4) We create a visual to identify the optimal k.

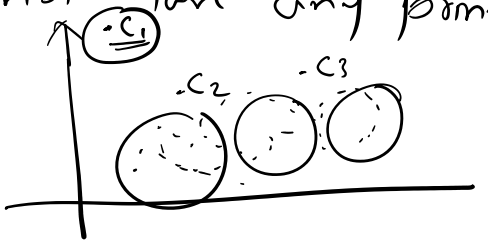
11 March 2023 20:52

init = K-means++

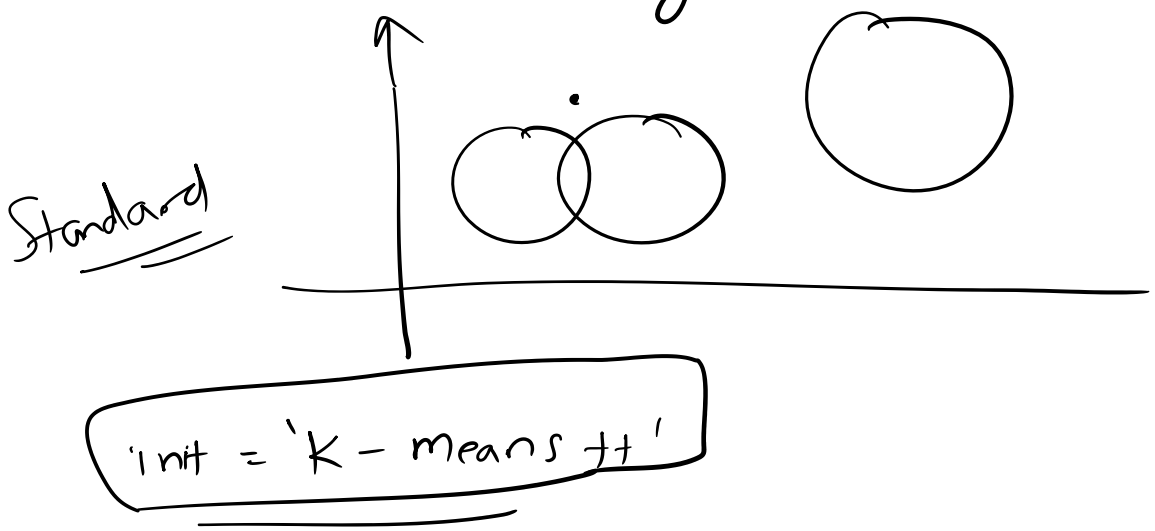


Standard KMeans Algorithm :-

1. Sensitive to the initialization of centroids or the mean points. If the centroid is very far, then that cluster will not have any point associated with it



2. More than 1 cluster might end up linked with single same cluster.



'init = 'k-means++'

K-means++

* K-means++ is the standard K-Means Algorithm coupled with smart initialization of centroid points.

11 March 2023 20:59

* random_state = 42

* Simply used to make the predictions static

k = 1

133 → 132.5 or 133.5 or 134

ML Algorithms :-

1. You cannot supply a missing data
2. If any column (variable) is having higher values then ML Algo will give importance to that.
3. ML algo do not understand text data.

eg :- E-ID (✓) E-Name (✗) Qualification (✓) Age Salary Yrs. Exp. (Numerical)

13 March 2023 19:36

Revision :-

1. Implementation of K-Means Algorithm in python using sklearn library
2. Elbow Method for optimal (k) no of clusters.
3. Data Preprocessing :-
   a. Missing Value Analysis.

13 March 2023 19:38

Agenda:-

1. Missing Value Treatment techniques.
    1. Mean, Median, Mode, Forward, Backward fill.
    2. Sklearn method for Missing Value imputation.
2. Data Encoding Techniques.

13 March 2023 19:40

(1) Missing Value Techniques :-

(1) Dropping the rows :-

data =

Country	Age	Salary	Purchased
France	44	72000	No
Spain	27	48000	Yes
Germany	30	54000	No
Spain	38	61000	No
~~Germany~~	~~40~~	~~Yes~~	
France	35	58000	Yes
~~Spain~~		~~52000~~	~~No~~
France	48	79000	Yes
Germany	50	83000	No
France	37	67000	Yes

data.dropna() ↵

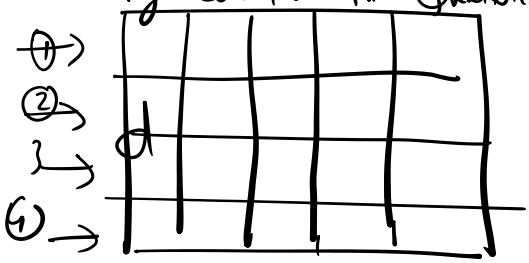
20%. The entire row gets deleted

(2) Dropping the column :-

data.drop('Age', axis=1)

~~Axis~~ : ↓ ↓ ↓ ↓ ↓ Data is 2-D data.

Age Sal Name Gend Qualification.



data.drop('Sal')

data.drop('Sal', axis=1)

data.drop('Gender', axis=1)

Country	Age	Salary	Purchased
France	44	72000	No
Spain	27	48000	Yes
Germany	30	54000	No
Spain	38	61000	No
Germany	40		Yes
France	35	58000	Yes
Spain		52000	No
France	48	79000	Yes
Germany	50	83000	No
France	37	67000	Yes

Age :- Replace all missing with mean of Age.

Salary :- Replace all missing values with mean of Salary

Mode: Most - frequent value in the dataset

eg:- [Apple(1), Orange, Apple(2), Mango, Banana, Apple(3)]

Apple - 3 ⟵ Apple is the mode data

Orange - 1

Mango - 1

Banana - 1

(1) Missing Value Analysis.

(2) It does not accept Categorical (String) data.

↓

Convert Cate data to Num. data

↓

Data Encoding

13 March 2023 21:10

Data Encoding :- Applicable only on Categorical data

- Nominal (No order)
  - ~~Label~~ Encoding
  - One Hot Encoding
- Ordinal (Specific Order)
  - ~~One Hot~~ Encoding
  - Label Encoding

13 March 2023 21:11

Label Encoding :- For ~~Nominal~~ Ordinal data.

eg:- Country Column in the Nominal data

Country	Age	Salary	Purchased
France	44	72000	No
Spain	27	48000	Yes
Germany	30	54000	No
Spain	38	61000	No
Germany	40		Yes
France	35	58000	Yes
Spain		52000	No
France	48	79000	Yes
Germany	50	83000	No
France	37	67000	Yes

Label Enc.

Country	Age	Salary	Purchased
0	44	72000	No
1	27	48000	Yes
2	30	54000	No
1	38	61000	No
2	40		Yes
0	35	58000	Yes
1		52000	No
0	48	79000	Yes
2	50	83000	No
0	37	67000	Yes

France – 0
Spain – 1
Germany – 2

13 March 2023 21:16

One - Hot - Encoding :-

~~Ordinal~~ Nominal data

BS - Bachelor
MS - Master
PHD - Doct.

Country	Education	Age	Salary	Purchased
France	BS	44	72000	No
Spain	MS	27	48000	Yes
Germany	BS	30	54000	No
Spain	MS	38	61000	No
Germany	PHD	40		Yes
France	PHD	35	58000	Yes
Spain	MS		52000	No
France	BS	48	79000	Yes
Germany	PHD	50	83000	No
France	BS	37	67000	Yes

OHE →

Country	Education_Ms	Education_PHD	Age	Salary	Purchased
France	0	0	44	72000	No
Spain	1	0	27	48000	Yes
Germany	0	0	30	54000	No
Spain	1	0	38	61000	No
Germany	0	1	40		Yes
France	0	1	35	58000	Yes
Spain	1	0		52000	No
France	0	0	48	79000	Yes
Germany	0	1	50	83000	No
France	0	0	37	67000	Yes

OHE increases the no. of columns in the data.

Education - BS: 1, 0, 1, 0, 0

Education - MS: 0, 1, 0, 1, 0

Education - PHD: 0, 0, 0, 0, 1

* Whatever we learnt about Label Encoding & One Hot, it is actually is "Wrong".

* It is Wrong, because it introduces intensional.

Label enc

Country	Age	Salary	Purchased
0	44	72000	No
1	27	48000	Yes
2	30	54000	No
1	38	61000	No
2	40		Yes
0	35	58000	Yes
1		52000	No
0	48	79000	Yes
2	50	83000	No
0	37	67000	Yes

OH Encoding

Country	Education_Ms	Education_PHD	Age	Salary	Purchased
France	0	0	44	72000	No
Spain	1	0	27	48000	Yes
Germany	0	0	30	54000	No
Spain	1	0	38	61000	No
Germany	0	1	40		Yes
France	0	1	35	58000	Yes
Spain	1	0		52000	No
France	0	0	48	79000	Yes
Germany	0	1	50	83000	No
France	0	0	37	67000	Yes

Country	Education_Ms	Education_PHD	Age	Salary	Purchased
0	0	0	44	72000	No
1	1	0	27	48000	Yes
2	0	0	30	54000	No
1	1	0	38	61000	No
2	0	1	40		Yes
0	0	1	35	58000	Yes
1	1	0		52000	No
0	0	0	48	79000	Yes
2	0	1	50	83000	No
0	0	0	37	67000	Yes

France - 0
Spain - 1
Germany - 2
X

Ordinal ← Label Enc

Education :- BS - 0, MS - 1, PHD - 2 → 2, 1, 0

Nominal ← One hot Enc

IT Industry

eg :- B.Tech - 2, M.Tech - 1, phd - 0 ← label Encd

Ind Clg

B.Tech - 0
M.Tech - 1
Phd - 2
← Label Enc

New Table after proper One Hot encoding & Label encoding

Original data

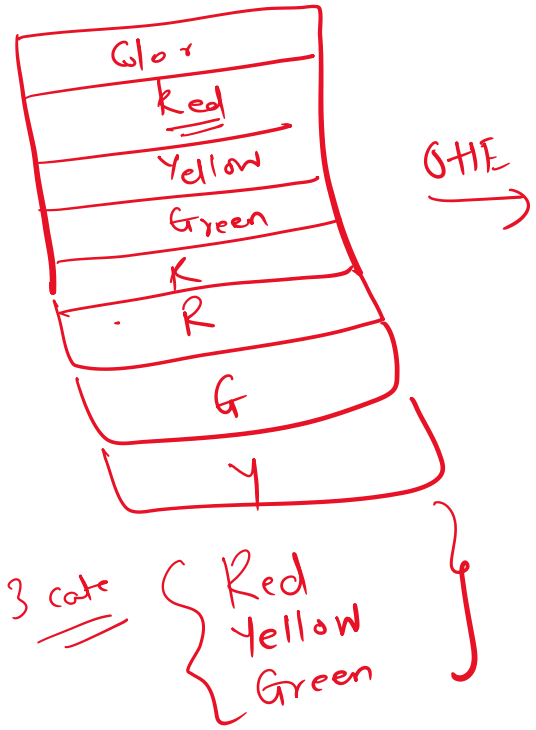
Country	Education	Age	Salary	Purchased
France	BS	44	72000	No
Spain	MS	27	48000	Yes
Germany	BS	30	54000	No
Spain	MS	38	61000	No
Germany	PHD	40		Yes
France	PHD	35	58000	Yes
Spain	MS		52000	No
France	BS	48	79000	Yes
Germany	PHD	50	83000	No
France	BS	37	67000	Yes

For Nominal → OHE

For Ordinal → LE

Final table after OHE & LE

Country_France	Country_Spain	Education	Age	Salary	Purchased
1	0	0	44	72000	No
0	1	1	27	48000	Yes
0	0	0	30	54000	No
0	1	1	38	61000	No
0	0	2	40		Yes
1	0	2	35	58000	Yes
0	1	1		52000	No
1	0	0	48	79000	Yes
0	0	2	50	83000	No
1	0	0	37	67000	Yes



Color_Red	Color_Yellow	Color_Green
1	0	0
0	1	0
0	0	1
0	0	0
1	0	0
0	0	1
0	1	0

14 March 2023 20:41

* Feature Scaling :-

Standardization → [-3, +3]

	Age (A)	Salary (B) ✓ ← Give importance
R_1	25	55000
R_2	35	65000
R_3	40	75000
R_4	45	80000
R_5	50	90000

Standardization formula

$$= \frac{X - X_{min}}{X_{max} - X_{min}}$$

eg :- C61 A

$$\frac{40 - 25}{50 - 25} = \frac{15}{25} = 0.6$$

$$\frac{X - X_{min}}{X_{max} - X_{min}}$$

For Col B,

$$\frac{80000 - 55000}{9000 - 55000} = 0.55$$

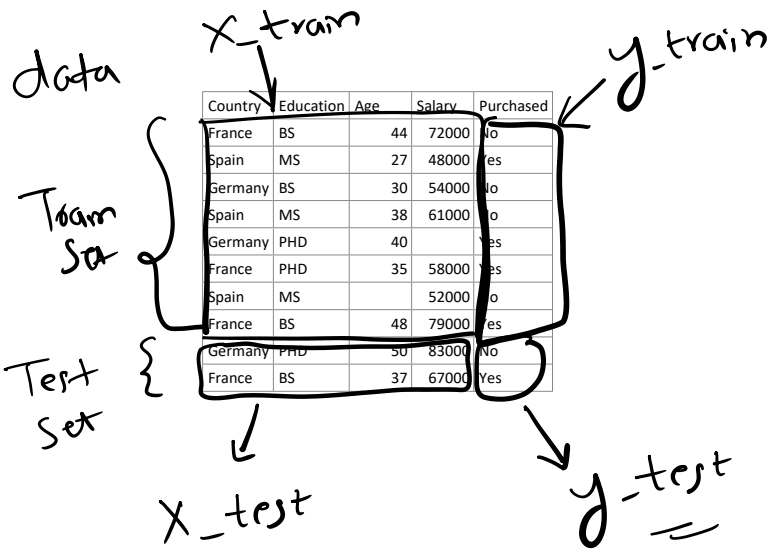
14 March 2023 20:50

In Machine Learning Projects, Data preprocessing steps

1. Missing Value Imputations
2. Data Encoding
3. Divide the data into X & y
4. Split the data using train test split
5. Feature Scaling on X_train & X_test.

14 March 2023 20:56

* Train Test Split :-



Country	Education	Age	Salary	Purchased
France	BS	44	72000	No
Spain	MS	27	48000	Yes
Germany	BS	30	54000	No
Spain	MS	38	61000	No
Germany	PHD	40		Yes
France	PHD	35	58000	Yes
Spain	MS		52000	No
France	BS	48	79000	Yes
Germany	PHD	50	83000	No
France	BS	37	67000	Yes

80 % - Train

20 % - Test

data.shape

(10, 5)

* Outlier Analysis :- Extreme value in any column are outliers.

* We check outliers only in Numerical data.

eg: Salary

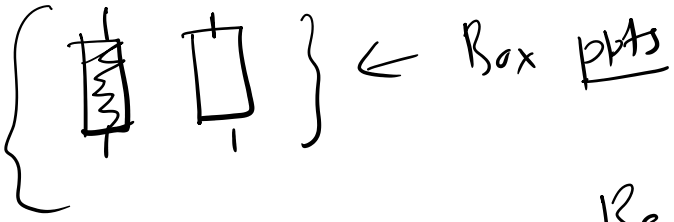
SD, HR { 75000, 80000, 95000 } are in thousands

1.3Cr ← CEO (Crores)

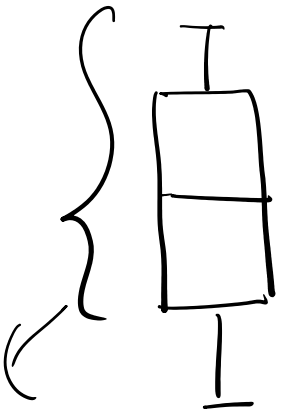
eg. Sachin Tend.

✳: Outlier Check :— Most popular method is

✡ Box plots :-



← Box plots



Reliance Stock ← Chart

14 March 2023 21:28

* Mean of the dataset is affected by outliers

Multicolinearity :- Relation between independent variable in the dataset

Multiple Linear Reg :- ① Many Independent variables
② Single target Variable.

↓

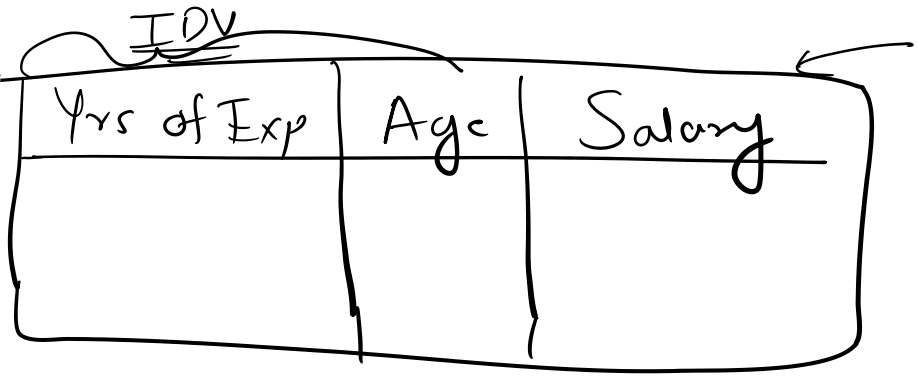
Can be applied when the are satisfied.

① Linear Relation both target & independent variables.

② No Linear Relation both any two independent variables

17 March 2023 19:48

eg:-



MLR

1. Yrs of Exp should have linear relation with Salary
2. Age should have linear relation with Salary

But,

3. Age & Yrs of Exp should have no linear relation
   ↳ Multicollinearity

Correlation :- [-1, 1]

If value = 1, -1, then very high relation

If value = 0.9, -0.9 the very good relation

If value = 0.5, -0.5 then ok. relation

If value = 0.3, -0.3 then temble relation.

If value = 0, the No relation

17 March 2023 20:47

R2 - Score :- [0, 1]

↓

R squared

If 0 Very Bad Model

If 0.5 or 50% then considerably ok

If value = 0.7 - 0.95 / 70% - 95% } Good to Very Good.

If value = 1 or 100% then suspicious.

17 March 2023 20:51

Sales = $b_0 + b_1 * TV + b_2 * Radio + b_3 * facebook + b_4 x$ insta

LR ⇒ $y = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3$

$R^2 = 0.83$ or 83%.

R^2 = either = 83% or < 0.83 after adding fb to the sales eqn.

R^2 score $\downarrow \propto \frac{1}{\text{no. of IDV} \uparrow}$

$\downarrow$ Penalize this

or In order to not have an effect of adding more independent variables we calculate adjusted R^2 value.

$$\text{Adj. } R^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1}$$

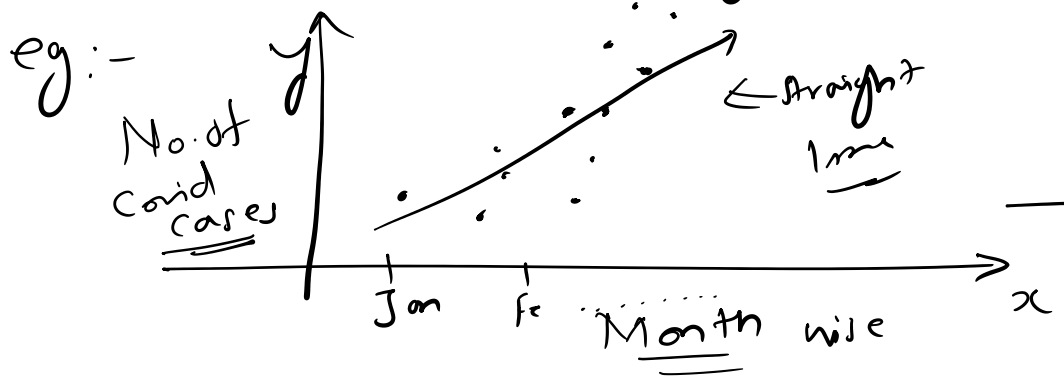
n = no. of rows in the data

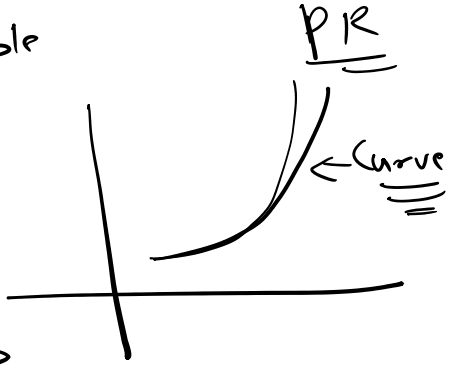
p = no. of columns in the data

Polynomial Regression :-

* Used for non-linear relationships between independent & target variable

eg:-





17 March 2023 21:11