$y_1 = \begin{bmatrix} 0.91 & 0.39 \end{bmatrix} \begin{bmatrix} 4 \\ 1 \end{bmatrix} =$

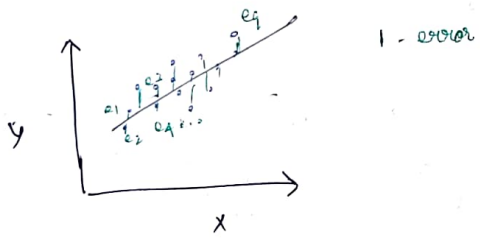
16/3

Regression Analysis

Primary objective : to predict data

Linear regression

x → independant variable
y → dependant variable



1 - error

residue = $e_1 + e_2 + \cdots e_9$

$y = a_0 + a_1 x + e$

a_0 → intercept
a_1 → slope
e → error term

$y_i = a_0 + a_1 x_i + e_i$

$\Rightarrow e_i = y_i - (a_0 + a_1 x_i)$

Sum of errors squared, $E = \sum_{i=1}^{n} e_i$

$= \sum_{i=1}^{n} [y_i - (a_0 + a_1 x_i)]$

Sum of errors squared,

$= \sum_{i=1}^{n} e_i^2$

$= \sum_{i=1}^{n} [y_i - (a_0 + a_1 x_i)]^2$

Minimisation fn.

$J(a_1, a_0) = \sum_{i=1}^{n} [y_i - (a_0 + a_1 x_i)]^2$

$a_1 = \dfrac{\overline{xy} - \bar{x} \cdot \bar{y}}{(\bar{x^2}) - (\bar{x})^2}$

$a_0 = \bar{y} - a_1 \bar{x}$

Q.

x_i (Week)	y_i (Sales in thousands)	xy	x^2
1	1.2	1.2	1
2	1.8	3.6	4
3	2.6	7.8	9
4	3.2	12.8	16
5	3.8	19.0	25

$\bar{x} = 3$ $\quad \overline{xy} = 8.88$ $\quad \bar{y} = 2.52$

$\overline{x^2} = 11$

$$a_1 = \frac{\overline{xy} - \bar{x}\,\bar{y}}{(\overline{x^2}) - (\bar{x})^2} = \frac{8.88 - 3(2.52)}{11 - 9}$$

$$= \frac{1.32}{2} = 0.66$$

$$a_0 = \bar{y} - a_1 \bar{x} = 2.52 - 0.66(3)$$

$$= 0.54$$

$$\Rightarrow y = 0.54 + 0.66x$$

$\Rightarrow$ Sales for 7^{th} week

$$= 0.54 + 0.66(7)$$

$$y = 5.16$$

for 12^{th} week $y = 8.46$

Matrix Representation

$$Y = Xa + e$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \end{bmatrix} + \begin{bmatrix} e_0 \\ e_1 \\ \vdots \\ e_n \end{bmatrix}$$

$$\Rightarrow a = \left((X^T X)^{-1} X^T\right) Y$$

Q. Consider the no. of hours spent in a library. Predict the no. of hours that would be spent in the 7^{th} and 9^{th} week.

Week	Hours Spent
1	12
2	18
3	22
4	28
5	35

$$Y = \begin{bmatrix} 12 \\ 18 \\ 22 \\ 28 \\ 35 \end{bmatrix} \qquad X = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix}$$

x_0	y	x^2	xy
1	12	1	12
2	18	4	36
3	22	9	66
4	28	16	112
5	35	25	175

$\bar{x} = 3, \quad \bar{y} = 23 \quad \overline{x^2} = 11 \quad \overline{xy} = 80.2$

$$a_1 = \frac{\overline{xy} - \bar{x}\,\bar{y}}{\overline{x^2} - \bar{x}^2} = \frac{80.2 - 3(23)}{11 - 3^2}$$

$$= \frac{11.2}{2} = 5.6 \; ; \quad a_0 = \bar{y} - a_1 \bar{x} = 6.2$$

$$\Rightarrow \boxed{y = 6.2 + 5.6x}$$

For 7^{th} week, $y = 45.4$

9^{th} week, $y = 56.6$ hours

21/3 Multiple Linear Regression

$$y = f(x_1, x_2 \cdots x_n)$$

$$= a_0 + a_1 x_1 + a_2 x_2 + \cdots a_n x_n + \epsilon$$

$$\hat{a} = \left((X^T X)^{-1} \cdot X^T \right) y$$

x_1	x_2	y
1	4	6
2	5	8
3	8	12
4	12	12 → 2 ?

$$X = \begin{bmatrix} 1 & 1 & 4 \\ 1 & 2 & 5 \\ 1 & 3 & 8 \\ 1 & 4 & 12 \end{bmatrix} \qquad X^T = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 \\ 4 & 5 & 8 & 12 \end{bmatrix}$$

$$X^T X = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 \\ 4 & 5 & 8 & 12 \end{bmatrix}_{3\times4} \begin{bmatrix} 1 & 1 & 4 \\ 1 & 2 & 5 \\ 1 & 3 & 8 \\ 1 & 4 & 12 \end{bmatrix}_{4\times3}$$

$$= \begin{bmatrix} 4 & 10 & 29 \\ 10 & 30 & 86 \\ 29 & 86 & 249 \end{bmatrix} \qquad \begin{bmatrix} 4 & -10 & 19 \\ -10 & 30 & -46 \\ 19 & -46 & 109 \end{bmatrix}$$

$$(X^T X)^{-1} = \begin{bmatrix} 3.15 & -0.59 & -0.3 \\ -0.59 & 0.20 & 0.016 \\ -0.3 & 0.016 & 0.054 \end{bmatrix}$$

$$(X^T X)^{-1} X^T = \begin{bmatrix} 3.15 & -0.59 & -0.3 \\ -0.59 & 0.2 & 0.016 \\ -0.3 & 0.016 & 0.054 \end{bmatrix}_{3\times3} \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 \\ 4 & 5 & 8 & 12 \end{bmatrix}_{3\times4}$$

$$= \begin{bmatrix} 1.36 & 0.47 & -1.02 & -2.81 \\ 0.32 & 0.1 & 0.155 & 0.402 \\ -0.065 & 0.002 & 1.85 & -0.128 \end{bmatrix}$$

$$(X^T X)^{-1} X^T \cdot y = \begin{bmatrix} 1.36 & 0.47 & 1.02 & -2.81 \\ 0.32 & 0.1 & 0.155 & 0.402 \\ -0.065 & 0.002 & 1.85 & -0.128 \end{bmatrix}_{3\times4} \begin{bmatrix} 1 \\ 6 \\ 8 \\ 2 \end{bmatrix}_{4\times1}$$

$$\hat{a} = \begin{bmatrix} 6.72 \\ 2.964 \\ 47.239 \end{bmatrix} \qquad \begin{bmatrix} -1.69 \\ 3.48 \\ -0.05 \end{bmatrix}$$

$$y = 1.69 + 3.48 x_1 - 0.05 x_2$$

Q. Using multiple regression solve

z → Equity
x → net sales
y → Asset

z is dependant variable

x, y are independant variables

x_i (x_1)	y_i (x_2)	z_i (y)
12	8	4
18	12	6
22	16	7
28	36	8
35	42	11

$$X_i = \begin{pmatrix} 1 & 12 & 8 \\ 1 & 18 & 12 \\ 1 & 22 & 16 \\ 1 & 28 & 36 \\ 1 & 35 & 42 \end{pmatrix} \qquad X^T = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 12 & 18 & 22 & 28 & 35 \\ 8 & 12 & 16 & 36 & 42 \end{pmatrix}$$

$$X^T X = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 12 & 18 & 22 & 28 & 35 \\ 8 & 12 & 16 & 36 & 42 \end{pmatrix}_{3\times 5} \begin{pmatrix} 1 & 12 & 8 \\ 1 & 18 & 12 \\ 1 & 22 & 16 \\ 1 & 28 & 36 \\ 1 & 35 & 42 \end{pmatrix}_{5\times 3}$$

$$= \begin{pmatrix} 5 & 115 & 114 \\ 115 & 2961 & 3142 \\ 114 & 3142 & 3524 \end{pmatrix}$$

Polynomial Regression

$$y = a x^b$$

→ Non - linear

$$\log y = \log (a x^b)$$

$$\Rightarrow \log y = \log_{10} a + b \log_{10} x$$

$$y = a_0 + a_1 x + a_2 x^2 + \dots$$

$$\begin{pmatrix} n & \Sigma x_i & \Sigma x_i^2 \\ \Sigma x_i & \Sigma x_i^2 & \Sigma x_i^3 \\ \Sigma x_i^2 & \Sigma x_i^3 & \Sigma x_i^4 \end{pmatrix}$$

Matrix format of polynomial reg.

$$\begin{pmatrix} n & \Sigma x_i & \Sigma x_i^2 \\ \Sigma x_i & \Sigma x_i^2 & \Sigma x_i^3 \\ \Sigma x_i^2 & \Sigma x_i^3 & \Sigma x_i^4 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} \Sigma y_i \\ \Sigma x_i y_i \\ \Sigma x_i^2 y_i \end{pmatrix}$$

$$a = X^{-1} B$$

Q.

x_i	y_i	$x_i y_i$	x_i^2	x_i^3	x_i^4	$x_i^2 y$
1	1	1	1	1	1	1
2	4	8	4	8	16	16
3	9	27	9	27	81	81
4	15	60	16	64	256	240

$n = 4$, $\Sigma x_i = 10$, $\Sigma x_i y_i = 96$, $\Sigma x_i^2 = 30$,

$\Sigma x_i^3 = 100$ $\quad \Sigma x_i^4 = 354$ $\quad \Sigma y_i = 29$

$$\Rightarrow a = \begin{pmatrix} 4 & 10 & 30 \\ 10 & 30 & 100 \\ 30 & 100 & 354 \end{pmatrix}^{-1} \begin{pmatrix} 29 \\ 96 \\ 338 \end{pmatrix}$$

$$\Rightarrow A = \begin{pmatrix} -123/25 & \cdots & \cdots \\ 123 \cdot 45 & \cdots & \cdots \\ -25 & \cdots & 75 \end{pmatrix}$$

$$a = \begin{pmatrix} -0.75 \\ 0.95 \\ 0.75 \end{pmatrix}$$

$y = -0.75 + 0.95x + 0.75x^2$

Logistic Regression

$$P(x) = a_0 + a_1 x$$

$$y = \frac{1}{1 + \exp(-(a_0 + a_1 x))}$$

Q. Let's assume a binomial logistic regression where the classes are pass and fail. The student dataset has entrance based on historic data of those who are selected & not selected. Based on logistic regression values of learned parameters,

$a_0 = 1$, $a_1 = 8$ assuming marks of $x = 60$,

Compute the resultant value

$$a_0 + a_1 x = 1 + 8(60) = 481$$

$$y = \frac{1}{1 + \exp(-481)} = 1 \quad \text{(0.44?)}$$

> 0.5 selected

else not selected

Validation of regression method

① Mean Absolute Error (MAE)

$$\frac{1}{n} \sum_{i=0}^{n-1} |y_i - \hat{y}_i|$$

y_i = Actual target, $\hat{y}_i$ = predicted output

② Mean Squared Error (MSE)

$$\frac{1}{n} \sum_{i=0}^{n-1} (y_i - \hat{y}_i)^2$$

③ Root Mean Squared Error (RMSE)

$$= \sqrt{MSE}$$

④ Relative MSE

$$\frac{\sum_{i=0}^{n-1} (y_i - \hat{y}_i)^2}{\sum_{i=0}^{n-1} (y_i - \bar{y}_i)^2}$$

Coefficient of variation, $CV = \frac{RMSE}{\bar{y}}$

Standard error: $y_i - \hat{y}$

Q.

	x_i → y_i (Actual Sales)
i1	80
i2	90
i3	100
i4	110
i5	120

Consider 2 fresh items i6 & i7 whose actual values are 80 & 75 respectively. Regression model predicts the values of i6 & i7 as 75 & 85 respectively.

Find MAE, MSE, RMSE, Relative MSE, CV

$n = 2$

$$MAE = \frac{1}{2}\left(|75-80| + |85-75|\right)$$

$$= \frac{1}{2}(5+10) = 7.5$$

$$MSE = \frac{25+100}{2} = \frac{125}{2} = 62.5$$

$$RMSE = \sqrt{62.5} = 7.9$$

$$\text{Relative MSE} = \frac{\sum_{i=0}^{n-1}(y_i - \hat{y}_i)^2}{\sum_{i=0}^{n-1}(y_i - \bar{y})^2}$$

$$= \frac{(80-75)^2 + (75-85)^2}{(80-77.5)^2 + (75-77.5)^2}$$

$$= \frac{25+100}{6.25+6.25} = \frac{125}{12.5} = 10$$

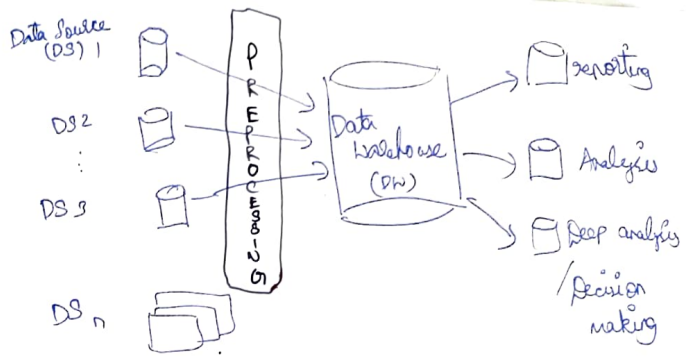
$$CV = \frac{RMSE}{\bar{y}} = \frac{7.9}{77.5} = 0.102$$

23/3

Data Warehouse (DW)

Characteristics

1. Subject-oriented (Sales / Marketing / Financial / etc.)
2. Integrated
3. Non volatile

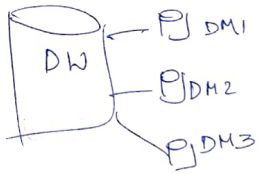


Online Analytical Processing - OLAP

4. Time variant (Date / Month / Years)

Data warehouse can contain
- metadata
- summary of data

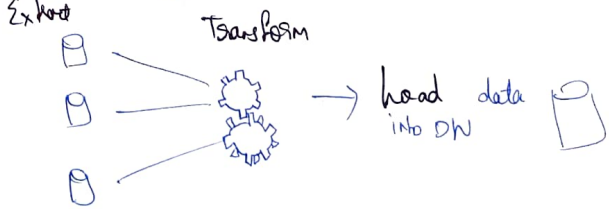
→ Tools are used in DW for analysis & reporting (Business Intelligence)



Data Mart (DM) → Smaller version of DW (for 2 years/subject & so on)

Datawarehouse Operation

ETL (Extract Transform load)

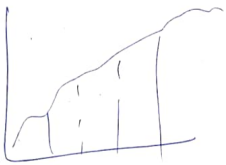


→ Extract from n sources. Transformation to avoid inconsistency among the data.

Transform : Slice, dice operations on data cube

Application of DW

→ Sales ↑ / ↓ in last 3 years



→ SAP software

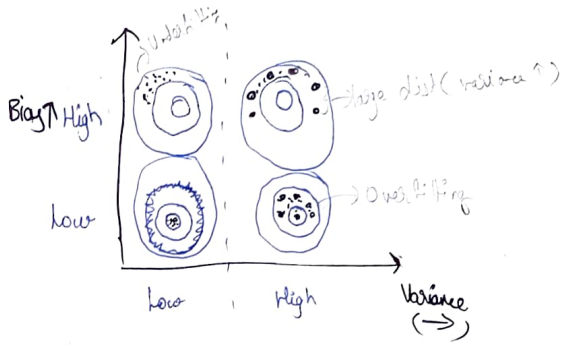
28/3

→ Difference b/w predicted value and actual value is bias error.

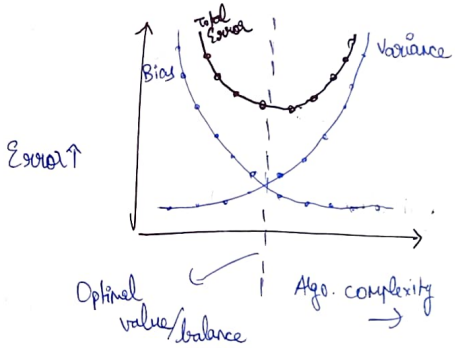
→ High bias, low variance → underfitting

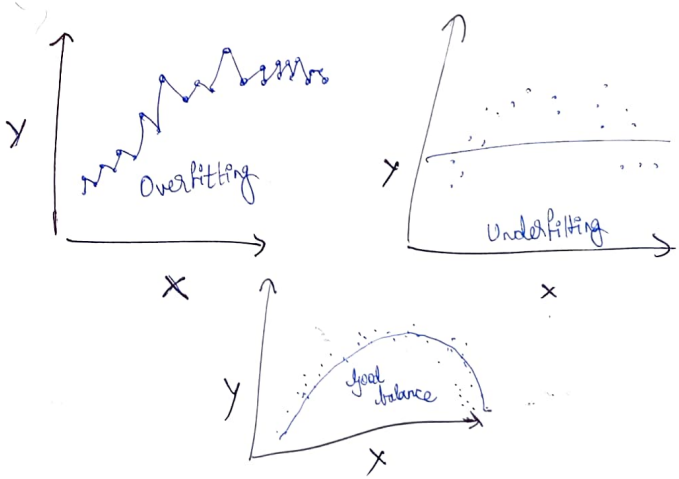
→ low bias, High variance → Overfitting

Bias / Variance trade off



Total Error = Bias² + variance + irreducible error.





→ In supervised learning, underfitting happen when the model is unable to capture underlying pattern of data.

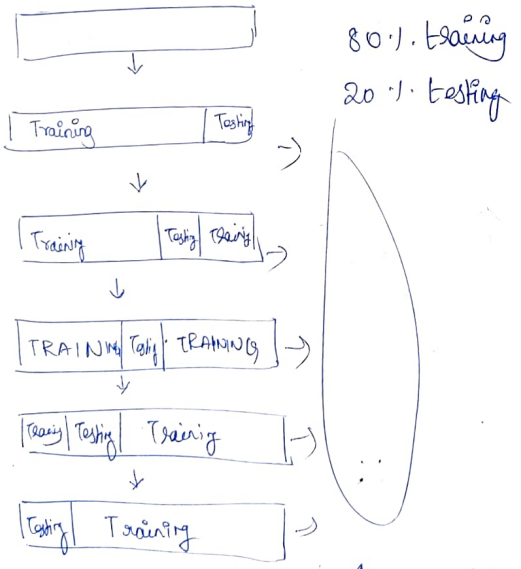
↳ Creating a linear model using non-linear data, underfitting can happen.

→ In supervised learning, overfitting happen when the model captures noise along with underlying pattern in data.

↳ Happens when we train the model lot over a noisy dataset.

K-fold cross validation

$k = 1 \dots 5$



→ LOOCV
→ Stratified

Q. Total instances: 25
K value: 5

⇒ each time 20% taken as test data

20% of 25 = $\frac{20}{100} \times 25 = \frac{20}{4} = 5$.

No. iteration	Training set observation	Testing set observation
1	All remaining values [5 6 7 ... 24]	[0 1 2 3 4]
2	[0 ... 4 10 ... 24]	[5 6 7 8 9]
3	[0 ... 9 15 ... 24]	[10 ... 14]
4	[0 ... 14 20 ... 24]	[15 ... 19]
5	[0 ... 19]	[20 ... 24]

Write the actual values, not ...

You can divide available data into multiple fold or subsets. Using one of this folds or subsets as a validation set & training the models on remaining folds. This process repeated for multiple times. Each time using different fold as a validation set. Finally results from each validation steps are averaged to produce more robust estimate of the models performance.

10 rows for iris dataset with $k=5$

	Sepal length	Sepal width	Petal length	Species
0	1.3	2.0		Virginica
1	2.5	3.7		Virginica
2	6.9	5.5		Versicolor
3	3.7	1.2		Virginica
4	4.1	5.8		Versicolor
5	3.0	6.0		Versicolor
6	5.3	6.9		Versicolor
7	2.0	2.1		Virginia
8	6.6	6.2		Versicolor
9	3.9	2.4		Virginica

20% of 10 = $\frac{20}{100} \times 10 = 2$.

Training data

Iteration	Testing data	Training data
1	[0,1]	[2,...9]
2	[2,3]	[0,1,4...9]
3	[4,5]	[0...3, 6..9]
4	[6,7]	[0...5, 8,9]
5	[8,9]	[0...7]

No.	Serial no.	CGPA	Interactivity	Practical knowledge	Communication skills	Job offer
0	1	>=9	Yes	Very good	Good	Yes
1	2	>=8	No	Good	Moderate	Yes
2	3	>=9	No	Average	Poor	No
3	4	<8	No	Average	Good	No
4	5	>=8	Yes	Good	Moderate	Yes
5	6	>=9	Yes	Good	Moderate	Yes
6	7	<8	Y	Good	Poor	N
7	8	>=9	N	Very good	Good	Y
8	9	>=8	Y	Good	Good	Y
9	10	>=8	Y	Avg.	Good	Y

Apply k-fold

Iteration	Training data	Testing data
1	[0 1]	[2 3 4 5 6 7 8 9]
2	[2 3]	[0 1 4 5 6 7 8 9]
3	[4 5]	[0 1 2 3 6 7 8 9]
4	[6 7]	[0 1 2 3 4 5 8 9]
5	[8 9]	[0 1 2 3 4 5 6 7]

Noisy data

→ Blank values

→ Outlier data

→ Inconsistent data (like working days has 'good', 'very good', 'moderate')

If value is not present,
- → can be removed.
- → Fill mean of the existing values / mode of for non-numeric values

→ Regression can be used to handle noise

→ Replace inconsistent rows / replace with most frequent value.

Bin methods → 0 to n bins, calc avg for each bin

→ Centroid method to classify into classes

30/4

Normalisation

1. Decimal scaling ($\div$ by 10^j) [j = 3 if max val = 700]
2. Min-max normalization
3. Z-score normalization

Decimal Scaling

$$V_i' = \frac{V_i}{10^j}$$ j depends on max. value in dataset

-10, 201, 301, -401, 501, 601, 701

Max abs value = 701 $\Rightarrow j = 3$

$$V_i = \frac{-10}{10^3}, \frac{201}{10^3}, \frac{301}{10^3}, \frac{-401}{10^3}, \frac{501}{10^3}, \frac{601}{10^3}, \frac{701}{10^3}$$

$$= -0.01, 0.201, 0.301, -0.401, +0.501, 0.601, 0.701$$

Min max normalisation

$$\frac{V - min(A)}{max(A) - min(A)} \left(new\_max(A) - new\_min(A)\right) + new\_min(A)$$

Z-Score normalisation

$$V' = \frac{V_i - \bar{A}}{\sigma_A}$$

For a normal dist, $\bar{A} = 0$, $\sigma_A = 1$

Advantages of using normalisation

→ Improve performance of ML algo.

→ Helps to reduce outlier data / better handling of outlier data.

→ Improve interpretability of results.

→ Better generalisation.

Disadvantages

→ Loss of information

→ Impact on outliers, interpretability

→ Additional computational cost

→ Pandas, numpy, sklearn can be used to normalise data.

Normalisation Implementation

```
from sklearn import preprocessing
import pandas as pd
from sklearn.datasets import fetch_housing data
housing = fetch_housing data (as_frame = True)
scaler = preprocessing.minmaxScaler()
d = scaler.fit_transform(housing.dat)
scaled_df = pd.Dataframe(d, colums = housingdata.
                                    colums)
Print (scaled_df)
```

to mention range: `minmaxScaler(feature_range=(6,2))`

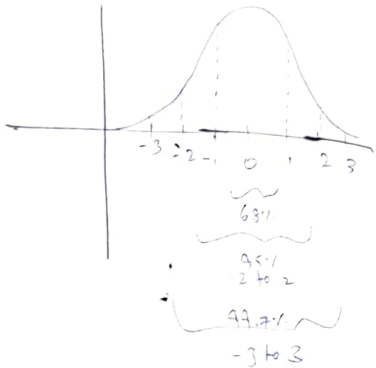
Median income	House age	Average no. of rooms per household	Avg bedrooms	Population	Avg occupants	latitude	longi
1.0793	1.5606	0.087				37.88	-122
1.0760	0.7843	0.076					
0.9320	2.00	0.105					
0.7093	2.0	0.070					
0.4695	2.0	0.077					

After ~~min~~ min. max scaler ↓

1.13	0.423
1.1307	0.424
1.1235	0.426
1.1285	0.418
1.1285	0.418

Methods to detect outliers & removal

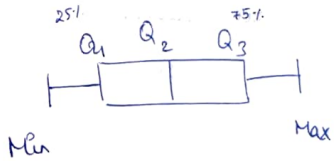
1) Standard deviation



→ Beyond -3 or +3 is outlier data.
→ Within -1 to +1 : Normal data

~~2) Interquartile range~~

2) Box plot

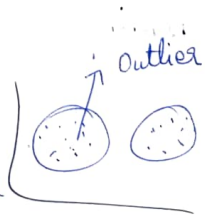


$$IQR = Q_3 - Q_1$$

Outliers $\begin{cases} < Q_1 - 1.5 * IQR \\ > Q_3 + 1.5 * IQR \end{cases}$

3) DBScan clustering

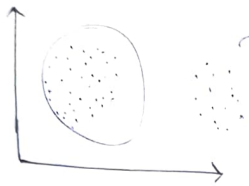
→ based on distance btw centroids
→ very large dist: outlier



4) Isolation Forest
5) Random Cut Forest → used by Amazon ?

} specialised techniques

Isolation Forest



→ deviated from original forest

Random cut forest

→ based on low & high values.

low → Normal ₹2000 - ₹6K
high → outliers ₹20K

Eg: Credit card purchases

→ Threshold like ₹ > 50,000

Sklearn, numpy can be used. (Anomaly)

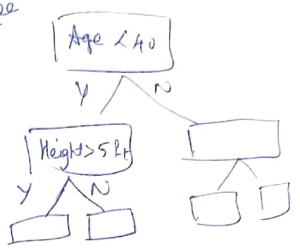
Applications 1) Quality assurance
2) Salary (not in expected range for given category)

11/4

Machine learning

1) Supervised (labelled data) - Decision tree, KNN, SVM, Random forest
2) Unsupervised (Unlabelled data)
3) Reinforcement (Agent - Reward)

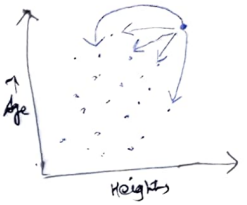
Decision tree



→ leaf nodes are class labells.

K - nearest neighbour (KNN)

k = 1, ... n



Find dist. of test data pt with all other data points. Find the closest k points & avg. of their weight values will be weight of test data pt.

1) Euclidean distance $\left(\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}\right)$
2) Manhattan distance $\sum_{i=1}^{k}|x_i - y_i|$ for continuous variables
3) Hanning distance (D_H) $\sum_{i=1}^{k}|x_i - y_i|$ for categorical variables

$x = y \Rightarrow D = 0$

$x \neq y \Rightarrow D = 1$

Q.

ID	Height	Age	Weight
1	5	45	77
2	5.11	26	47
3	5.6	30	55
4	5.9	34	59
5	4.8	40	72
6	5.8	36	60
7	5.3	19	40
8	5.8	28	60
9	5.85	23	45
10	5.6	32	58

Testing: 11 5.5 38

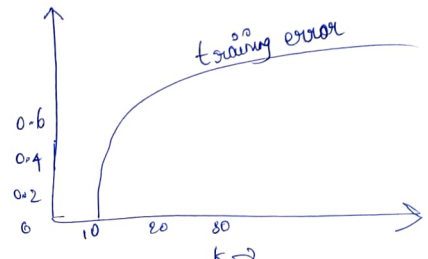
Distances wrt = test data points

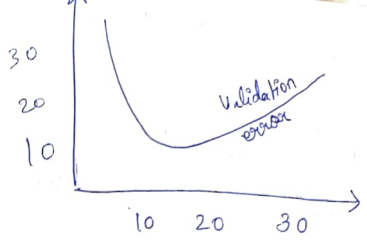
1) 7.017
2) 12.007
3) 8.001
4) 4.02
5) 2.119
6) 2.023
7) 19.02
8) 10.005
9) 15
10) 6.001

Min dist points are 6, 5, 4 (k = 3)

Avg of the weights of these pts. = $\frac{191}{3} = 63.67$

ID	Ht.	Age	Weight
4	5.9	34	~~59~~ 59
5	4.8	40	~~72~~ 72
6	5.8	36	60





Name	Age	Gender	Class of sports	Dist from test
Ajay	32	0	Football	17.03
Mark	40	0	Neither	25.02
Sarah	16	1	Cricket	1
zaira	34	1	Cricket	19
Sachin	55	0	Neither	40.013
Rahul	40	0	Cricket	25.02 ~~15.034~~
Pooja	20	1	Neither	5
Smith	15	0	Cricket	1
Lakhan	55	1	Football	40
Michal	15	0	Football	1
Test → Angelina	15	1	?	

Min dist pts are Sarah, Smith, Michal

Cricket, Cricket, Football

⇒ Cricket.

Steps used for supervised learning algo.

1) Determine type of training dataset
2) Collect / gather labelled training data
3) Split training dataset into training, test & validation dataset.
   ↳ subset of training dataset
4) Determine input features of training dataset which should have enough knowledge which the model can accurately predict output
5) Determine suitable algorithm for the model such as SVM, decision tree, Random forest etc.
6) Execute algorithm on training dataset. Sometimes we need validation set as a control parameter which are subset of training dataset.
7) Evaluate accuracy of your model by providing test dataset. If the model predict correct output, it means model is accurate.

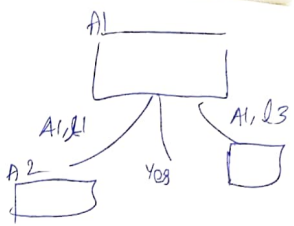
13/4

Decision Tree

Entropy, $E(S) = -\sum_{i=1}^{n} P(x_i) \log_2 P(x_i)$

Information gain, $IG(S, A) = E(S) - E(S, A)$

Total entropy − Average entropy



=0 → No split

≠0 → Further split.

Outlook	Temperature	Humidity	Wind	Play-Football
Sunny	Hot	High	Weak	N
Sunny	Hot	High	Strong	N
Overcast	Hot	High	Weak	Y
Rain	Mild	High	Weak	Y
Rain	Cool	Normal	Weak	Y
Rain	Cool	Normal	Strong	N
Overcast	Cool	Normal	Strong	Y
Sunny	Mild	High	Weak	N
Sunny	Cool	Normal	Weak	Y
Rain	Mild	Normal	Weak	Y
Sunny	Mild	Normal	Strong	Y
Overcast	Mild	High	Strong	Y
Overcast	Hot	Normal	Weak	Y
Rain	Mild	High	Strong	N

$$E(S) = -\left[\frac{9}{14}\log_2\left(\frac{9}{14}\right) + \frac{5}{14}\log_2\left(\frac{5}{14}\right)\right]$$

$$= 0.94$$

Outlook

Outlook

Outlook	Play Football: Yes	No	Total
Sunny	3	2	5
Overcast	4	0	4
Rainy	2	3	5
			14

$$E(S, Outlook) = \frac{5}{14} E(3,2) + \frac{4}{14} E(4,0) + \frac{5}{14} \cdot E(2,3)$$

$$= \frac{5}{14} \cdot -\left(\frac{3}{5} \cdot \log \frac{3}{5} + \frac{2}{5} \log \frac{2}{5}\right) + \frac{4}{14} \cdot 0 + \frac{5}{14} \left(-\frac{2}{5} \log \frac{2}{5} - \frac{3}{5} \log \frac{3}{5}\right)$$

$$= 0.693$$

$$IG(S, outlook) = 0.94 - 0.693 = 0.247$$

Temperature

Temperature	Play Football: Yes	No	Total
Hot	2	2	4
Mild	4	2	6
Cool	3	1	4
			14

$$E(S, temp) = \frac{4}{14} \cdot E(2,2) + \frac{6}{14} E(4,2) + \frac{4}{14} E(3,1)$$

$$= \frac{4}{14} -\left(\frac{2}{4} \log \frac{2}{4} + \frac{2}{4} \log \frac{2}{4}\right) + \frac{6}{14} -\left(\frac{4}{6} \log \frac{4}{6} + \frac{2}{6} \log \frac{2}{6}\right) + \frac{4}{14} -\left(\frac{3}{4} \log \frac{3}{4} + \frac{1}{4} \log \frac{1}{4}\right)$$

$$= 0.911$$

$$IG(S, temp) = 0.94 - 0.911 = 0.029$$

~~E(S, Humidity)~~ =

Humidity	Yes	No	Total
High	3	4	7
Normal	6	1	7

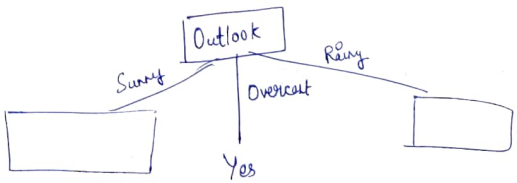
$$E(S, humidity) = \frac{7}{14} E(3,4) + \frac{7}{14} E(6,1) = 0.788$$

$$IG(S, hum) = 0.94 - 0.788 = 0.152$$

$$E(S, Wind) = 0.8932$$

$$IG(S, hum) = 0.048$$

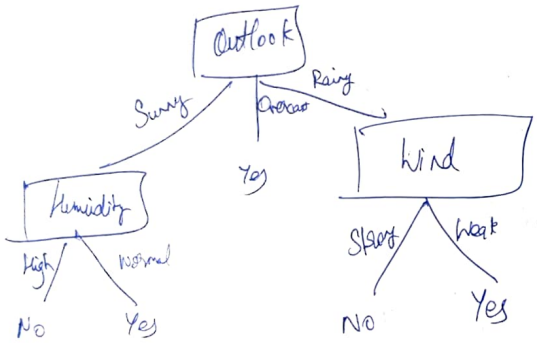
Highest IG is for ~~humidity~~ outlook



Outlook	Temp	Humidity	Wind	Play Football
			W	N
Sunny	High	H	S	N
Sunny	H	H		N
Sunny	Mild	H	W	Y
Sunny	Cool	N	S	Y
Sunny	Mild	N		

$$E(S) = -\left[\frac{3}{5}\log\frac{3}{5} + \frac{2}{5}\log\frac{2}{5}\right]$$

∴ Humidity has highest IG



Gini (E)

$$= 1 - \sum_{j=1}^{k} P_j^2$$

→ used for CART

Classification & Regression Tree

18/4

For outlook = Sunny

$$E(S) = -\left\{\frac{3}{5}\log\frac{3}{5} + \frac{2}{5}\log\frac{2}{5}\right\} = 0.97$$

$$IG(\text{Outlook} = \text{Sunny}) = 0.94 - 1.732$$

$$=$$

Temp

	Yes	No	Total
Hot	0	2	2
Cool	1	0	1
Mild	1	1	2
			5

$$E(S, \text{Temp}) = -\frac{2}{5}\left(\frac{1}{2}\log\frac{1}{2} + \frac{1}{2}\log\frac{1}{2}\right)$$

$$= -\frac{2}{5}\log\frac{1}{2} = \frac{2}{5}$$

$$IG(S, \text{Temp}) = 0.97 - 2/5 = 0.57$$

Humidity

	Yes	No	Total
High	0	3	3
Normal	2	0	2
			5

$$E(S, \text{Humidity}) = 0$$

$$IG(S, \text{Humidity}) = 1.732$$

$$\log_{a_2} a = \frac{\log_{10} a}{\log_{10} a_2}$$

CART

$$\text{Gini}(S) = 1 - \sum_{i=1} P_i^2$$

$$= 1 - \left[\left(\frac{9}{14}\right)^2 + \left(\frac{5}{14}\right)^2\right]$$

$$= 0.4591$$

$$\text{Gini}(S, \text{Outlook}) = \frac{5}{14}\,\text{Gini}(S,2) + \frac{4}{14}\,\text{Gini}(4,0) + \frac{5}{14}\,\text{Gini}(2,3)$$

$$= \frac{5}{4}\left(1 - \left[\left(\frac{3}{5}\right)^2 + \left(\frac{2}{5}\right)^2\right]\right) + \frac{4}{14} \times 0 + \frac{5}{14} \times \left(1 - \left[\left(\frac{2}{5}\right)^2 + \left(\frac{3}{5}\right)^2\right]\right)$$

$$= 0.171 + 0 + 0.171 = 0.342$$

$$\text{Gini gain}(S, \text{Outlook}) = 0.459 - 0.342 = 0.117$$

$$\text{Gini gain}(S, \text{temperature}) = 0.459 - 0.4405 = 0.0185$$

$$\text{Gini gain}(S, \text{Humidity}) = 0.4459 - 0.3674 = 0.0916$$

$$\text{Gini gain}(S, \text{Windy}) = 0.459 - 0.4286 = 0.0304$$

Apply decision tree for (CGPA, Job offer dataset ← back pages) using Information gain (Not gini)

$$E(S) = -\left[\frac{8}{10}\log\frac{8}{10} + \frac{2}{10}\log\frac{2}{10}\right]$$

$$= 0.882$$

CGPA

	Yes	No	Total
>=9	3	1	4
>=8	4	0	4
<8	0	2	2
			10

$$E(S, \text{CGPA}) = \frac{4}{10}E(S,1) + \frac{4}{10} \times 0 + \frac{2}{10} \times 0$$

$$= \frac{4}{10}\left(-\left(\frac{3}{4}\log\frac{3}{4} + \frac{1}{4}\log\frac{1}{4}\right)\right)$$

$$= 0.3246$$

$$IG(S, \text{CGPA}) = 0.5574$$

Interactive

	Yes	No	Total
Yes	5	1	5
No	2	2	4
			10

$$E(S, \text{Interactive}) = -\frac{6}{10}\left(\frac{5}{6}\log\frac{5}{6} + \frac{1}{6}\log\frac{1}{6}\right) - \frac{4}{10}\left(2\cdot\frac{2}{4}\log\frac{2}{4}\right)$$

$$= \frac{6}{5}(0.65) + \frac{4}{10}$$

$$= 0.39$$

$$IG(S, \text{Interactive}) = 0.492$$

(CGPA: Highest)

Practical knowledge

	Yes	No	Total
Very good	2	0	2
good	4	1	5
avg	1	2	3
			10

CGPA

>=8 / <8 / >=9

Yes No

$$E(S, \text{Practical knowledge}) = \frac{5}{10}E(1,1) + \frac{3}{10}E(1,2)$$

$$= -0.5\left(\frac{4}{5}\log\frac{4}{5} + \frac{1}{5}\log\frac{1}{5}\right) - 0.3\left(\frac{1}{3}\log\frac{1}{3} + \frac{2}{3}\log\frac{2}{3}\right)$$

$$= 0.361 + 0.276 = 0.637$$

$$IG = 0.2456$$

Comm. skills

	Yes	No	Total
Good	4	1	5
Moderate	3	0	3
Poor	0	2	2
			10

$$E(S, \text{Com}) = -\frac{5}{10}\left(\frac{4}{5}\log\frac{4}{5} + \frac{1}{5}\log\frac{1}{5}\right)$$

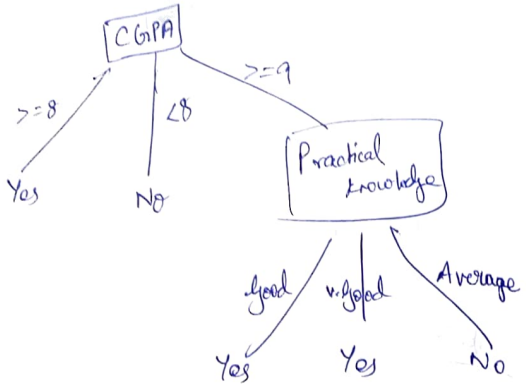
$$= 0.361$$

$$IG = 0.521$$

CGPA>=9

CGPA	Interactive	Practical knowledge	Comm. skills	Job offer
>=9	Y	VG	G	Y
>=9	N	A	P	N
>=9	Y	G	M	Y
>=9	N	VG	G	Y

$$E(S) = -\left(\frac{3}{4}\log\frac{3}{4} + \frac{1}{4}\log\frac{1}{4}\right) = 0.8113$$



Interactive

	Yes	No	
Y	2	0	2
N	1	1	2
			4

$= \frac{2}{4}\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log\frac{1}{2}\right) = \frac{1}{2}\log\frac{1}{2}$

$= -\frac{1}{2}\log 2^{-1} = 1/2 = 0.5$

$IG_1 = 0.3113.$

Prac

	Yes	No	Total
VG	2	0	2
G	1	0	1
A	0	1	1
			4

$E = 0, \quad IG_1 = 0.8113.$

Comm.

	Yes	No	Total
G	2	0	2
M	1	0	1
P	0	1	1

$E = 0, \quad IG_1 = 0.8113$

Prac know = VG.

Prk	Job
VG	Y
VG	Y

Prack = G

Prk	Job
G	Y

Prac k = Avg

Prk	Job
Avg	No

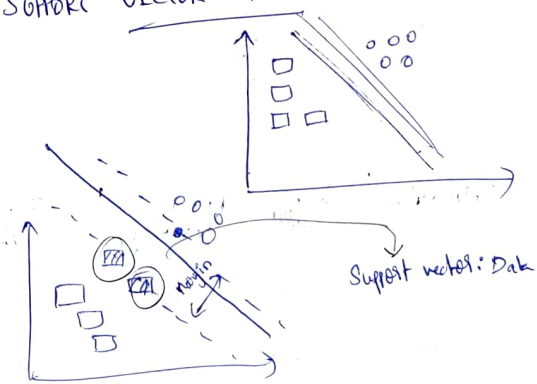
CGPA <8

CGPA	...	Job
<8		No
<8		No

CGPA >=8

CGPA	...	Job
>=8		Y
>=8		Y
>=8		Y
>=8		Y

SUPPORT VECTOR MACHINE



Support vectors: Data pts. close to the margin

Cost Function

$$C(x, y, f(x)) = \begin{cases} 0 & \text{if } y \cdot f(x) \geq 1 \\ 1 - y\, f(x) & \text{else} \end{cases}$$

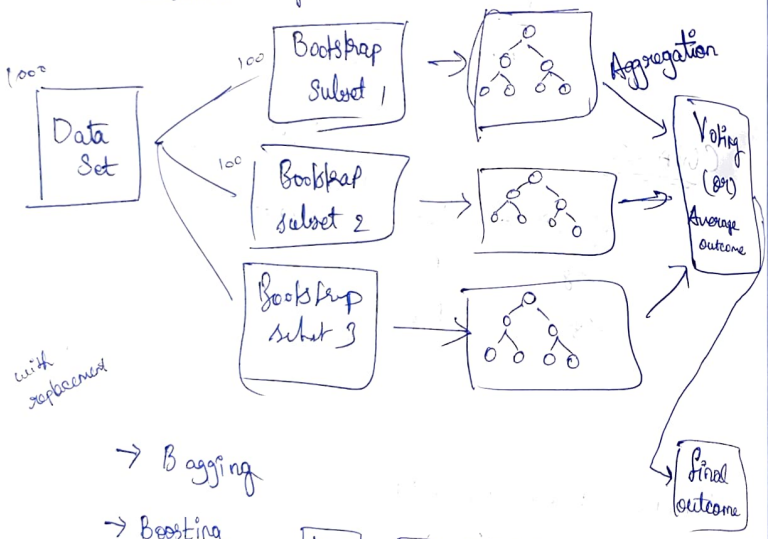
Gradient updates

$w = w - \alpha(2\lambda w)$ → No misclassification

$w = w + \alpha(y_i * x_i - 2\lambda w)$ → final gradient update for proper classification

20/4 Random Forest

→ Ensemble algorithm



→ Bagging

→ Boosting — Weak — [ ] — Strong

Weak to Strong classifier

Q.

	CGPA	Interactives	Comm. skills	Prog. knowledge	Job offer
1.	>=9	Y	Good	G	Y
2.	<9	N	Moderate	G	Y
3.	>=9	No	M	Avg	N
4.	<9	N	M	A	N
5.	>=9	Y	M	G	Y

Bootstrap Subset S1

1.	>=9	Y	G	G	Y
2.	<9	N	M	G	Y
3.	>=9	N	M	A	N
4.	>=9	N	M	A	N
5.	>=9	Y	M	G	Y

S2

1.	<9	N	M	G	Y
2.	>=9	N	M	A	N
3.	>=9	N	M	A	N
4.	>=9	Y	M	G	Y
5.	>=9	Y	M	G	Y

S3

1.	>=9	Y	G	G	Y
2.	>=9	Y	G	G	Y
3.	<9	N	M	G	Y
4.	>=9	N	M	A	N
5.	>=9	N	M	A	N

Choose 2) Construct decision tree for each subset.

3) Choose no. of features to be used in each tree. (2)

4) Choose depth of each decision tree to be 2

5) Repeat the following steps for 3 decision tree that are constructed.

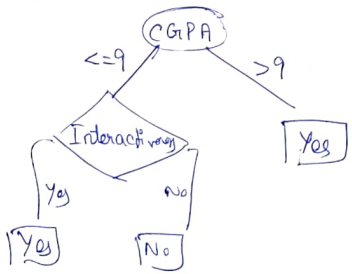
a) Choose bootstrap sample S1 of 5 data items to construct decision tree 1.

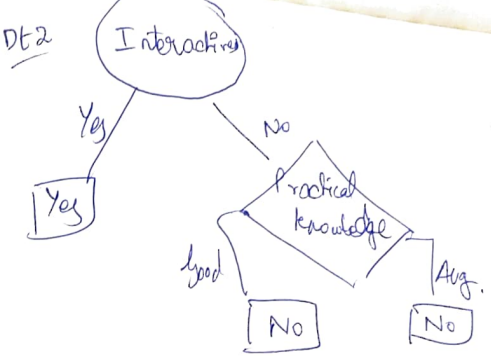
b) Compute the best split feature to be placed at each node.

c) Construct decision tree using CART algorithm

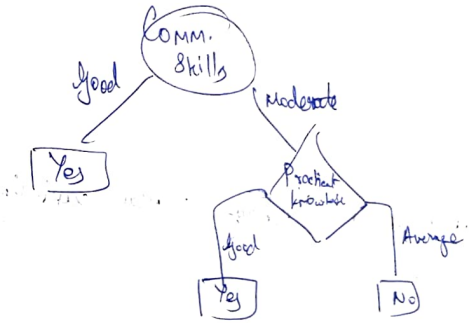
6. Combine predictions from all the classifiers taking an average for regression problem or majority voting for classification.

1st dataset using CGPA & interactiveness.





DT3 Comm & Pk



4th instance as test data

DT1 : Yes , DT2: No , DT3: No

After Voting : No

Advantages:

→ Can solve both classification & regression

→ Reduces overfitting & variance compared to single decision tree

→ Provides more accuracy

→ Performs well even with small data set

Dis:

→ Increased complexity in construction & prediction

→ Uses more computational resources

27/4

Types

1. Partition based clustering / Centroid based clustering
2. Hierarchical based / Connectivity based clustering
3. Density based / Model based
4. Distribution based clustering
5. Fuzzy based clustering

Partition based Clustering (K means clustering)

Object	x-coordinate	y-coordinate	
1	2	4	
2	4	6	C_1
3	6	8	
4	10	4	
5	12	4	C_2

C1	C2
(4,6)	(12,4)

Iteration 1

Dist (1, centroid 1) = $2\sqrt{2}$

Dist (1, centroid 2) = 10

Dist (3, Centroid 1) = $2\sqrt{2}$

Dist (3, centroid 2) = $\sqrt{52}$

Dist (4, C1) = $\sqrt{40}$

" (4, C2) = $\sqrt{4}$

After iteration 1

C1	C2
(4,6)	(10,4)
(2,4)	(12,4)
(6,8)	
	(11,4)

New centroid : (4,6)

Iteration -2

Dist (2, C1) = 0

Dist (2, C2) = $\sqrt{53}$

Dist (1, C1) = $\sqrt{8}$

Dist (1, C2) = $\sqrt{81}$

Dist (3, C1) = $\sqrt{8}$

Dist (3, C2) = $\sqrt{41}$

Dist (4, C1) = $\sqrt{40}$

Dist (4, C2) = $\sqrt{1}$

Dist (5, C1) = $\sqrt{68}$

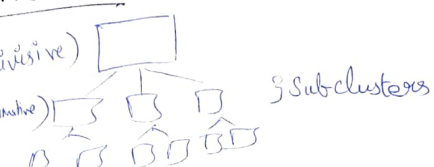
Dist (5, C2) = $\sqrt{1}$

After Iteration 2

C1	C2
(2,4)	(10,4)
(4,6)	(12,4)
(6,8)	

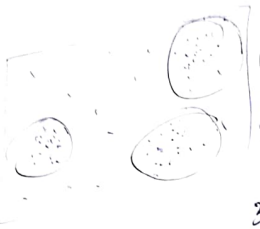
Hierarchical based

1. Top down (divisive)
2. Bottom up (Agglomerative)



→ Clusters & sub clusters.

Density based

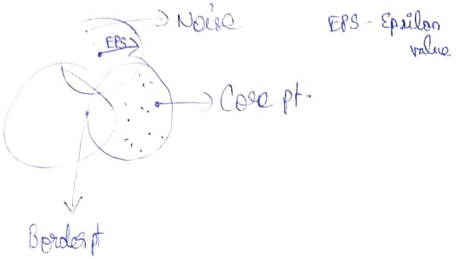


1. Core pt.
2. Border point
3. Noise

~~DB Scan~~

DBSCAN

Density Based Spatial Clustering Application with Noise



EPS - Epsilon value

BIRCH

Balanced Iterative Reducing and Clustering using Hierarchies

1. CF Tree (Clustering Feature tree)
2. Threshold

2/5 Use k-means algo (k=2) and show the result for the following

S.no	X	Y
1	3	5
2	7	8
3	12	5
4	16	9

Let $C1 = (3,5)$, $C2 = (12,5)$

C1	C2
(3,5)	(12,5)
(7,8)	(16,9)

$Dist(2, C1) = \sqrt{25}$

$Dist(2, C2) = \sqrt{34}$

New C1 = (5, 6.5) New C2 = (14, 7)

C1	C2
(3,5)	(12,5)
(7,8)	(16,9)

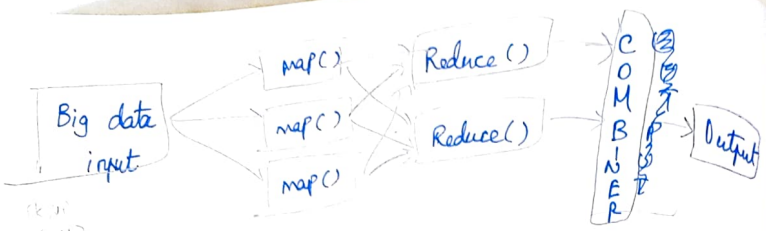
$Dist(2, C1) = 2.5$

$Dist(2, C2) = 7.07$

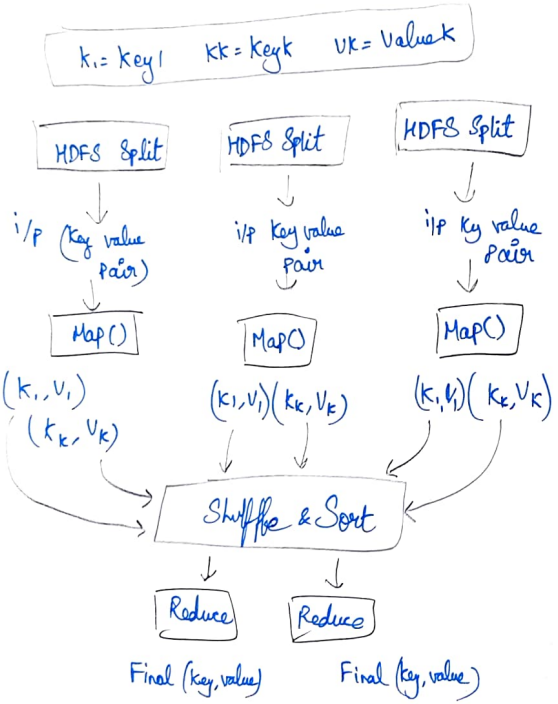
Hadoop

→ Distributed File System (HDFS)
→ Distributed processing (MapReduce)
→ Size of HDFS block 128 - 256 MB
→ Default replication factor - 3

Important components:
1) HDFS
2) Map Reduce
3) YARN framework
4) Hadoop common utilities (Pig, Hbase, Hive)



Reduce operations: 1) Sort()
2) Shuffle()



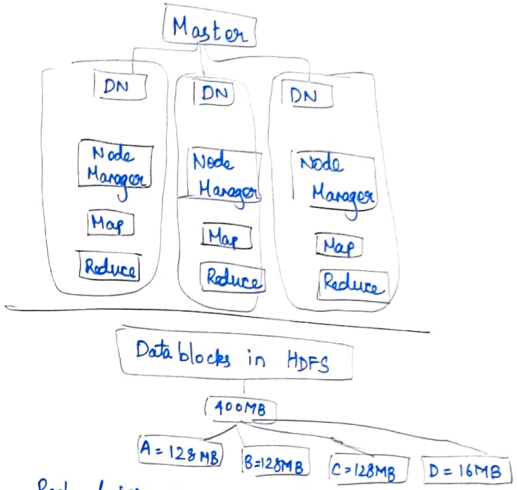
HDFS

→ Based on Google FS (GFS)

→ Size of each block 128 - 256 MB

→ Name node & data node
(Master) (Stores file blocks)
(Store metadata)

→ Rack aware Eg. Rack 1 has data nodes DN1 - 10
Rack 2 " 11 - 20



Repl. factor = 3 ⇒ Total file blocks = 3 * 4 = 12 blks

→ Usually 500 DN are used.

Advantages
1) → Size of unix file blk is 4KB while HDFS has 128 MB
2) Replication
3) High throughput
4) Commodity hardware (Works with inexpensive hardware)
5) High Fault tolerant
6) Software merging failure

Installation
1. Pseudo distributed
2. Fully distributed (Min 2 DN present)

XML files
- Core-Site.xml
- Yarn-Site.xml
- HDFS-site.xml
- Map Red-Site.xml

} Config files

pig
hive
Hbase
Sqoop

Data Science & Analytics

Hive - External table

> Create EXTERNAL TABLE emp.employee_external (id int, name string, age int, gender string) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' LOCATION '/user/hive/data/employee_external'

Temporary Table

>CREATE temporary TABLE emp.employee_temp(id int, ...) ROW ... BY ',';

Partitioning Table

> Create TABLE zipcodes (RecordNumber int, Country string, city string, zipcode int) PARTITIONED BY (State String) ROW FORMAT... ;

> Describe zipcodes;

> DROP TABLE <tablename>

> DELETE FROM <tablename> WHERE ...

MongoDB

→ Document based, dynamic schema

C++

Cross platform

High performance / Scalable / Availability

RDBMS	MongoDB	Column, Field
DB	DB	Join, Embedded doc
Table	Collection	Primary key, _id
Tuple	Document	

→ index on any attribute ; replication, auto sharding, rich queries

Data Types

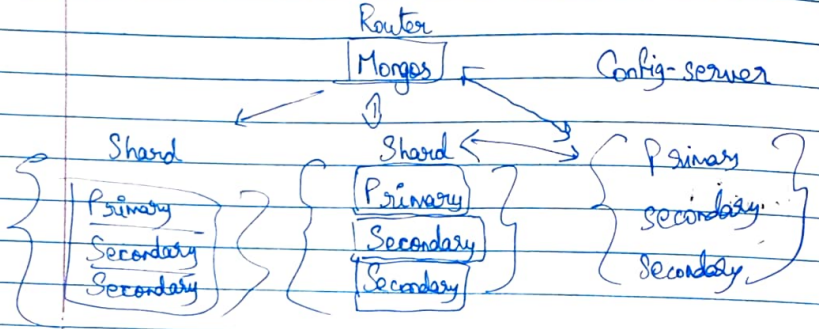
String, Integer, boolean, double, min/max keys, arrays, timestamp, object, null, symbol, date, object Id, binary Data, code, reg ex

db.find()

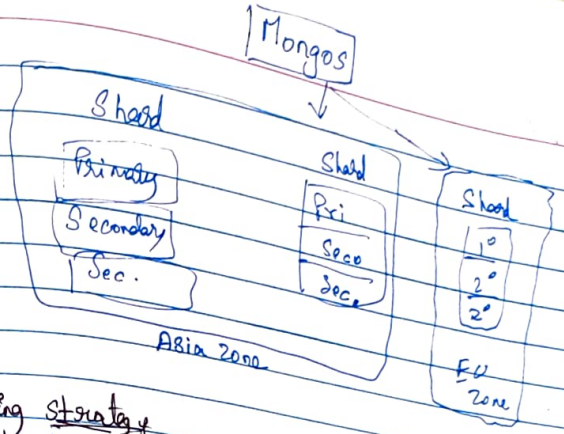
.find().pretty()?

11/5 Partitioning in MongoDB

Sharding



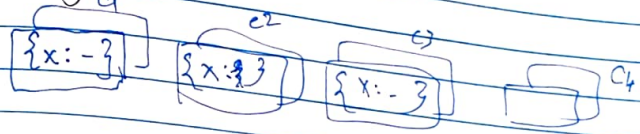
Zone Sharding

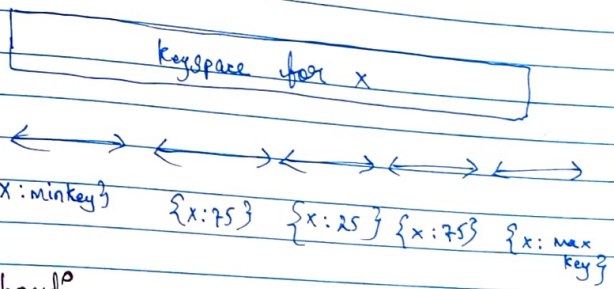


Sharding Strategy

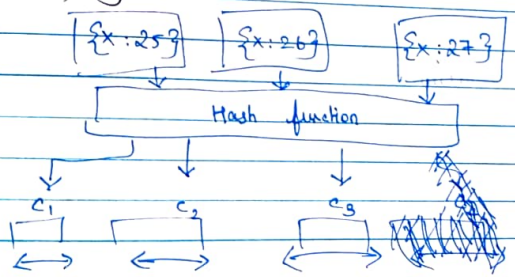
- Ranged Sharding
- Hashed Sharding

Ranged Sharding



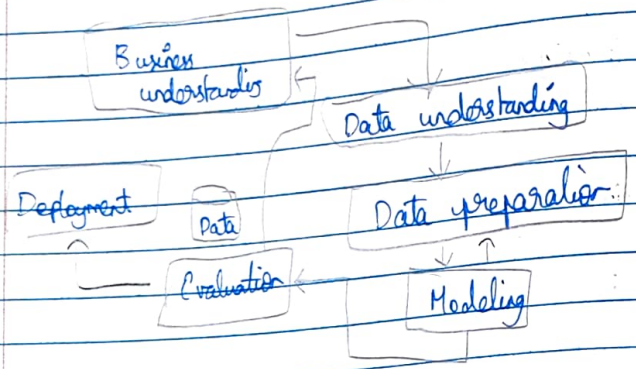


Hashed Sharding



Benefits
→ Increased Read/Write throughput
→ Increased storage capacity
→ Data locality

13/5 Data Munging (Wrangling)



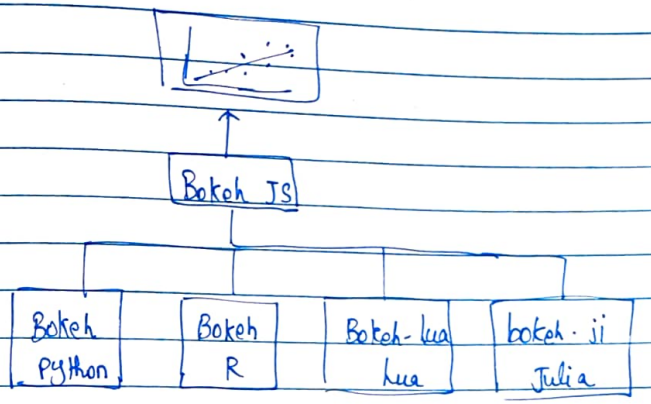
1. Pre processing
2. Standardize dataset into understandable format
3. Cleaning data from noise & erroneous elements
4. Consolidate data from various sources.
5. Matching data with existing dataset.
6. Filtering data through determined settings.

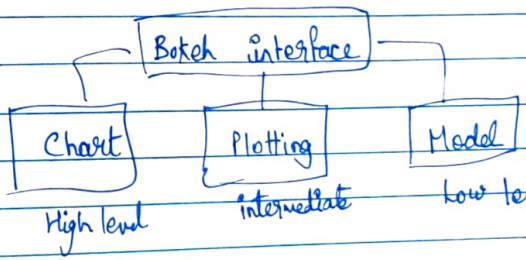
Data cleaning
- Data audit
- Workflow specification & execution
- Post processing control

Tools for data munging
→ Excel Power query / spreadsheets
→ OpenRefine
→ Google Data Prep
→ Tabula
→ Data wrangler
→ CSV kit

16/5 Data Pipeline

Interactive data visualization using Bokeh





Methodology to create chart

1. Import the library and functional methods
2. Prepare a data
3. Set output mode (notebook / web browser / server)
4. Create chart with styling options
5. Visualise the chart

Barchart on web browser

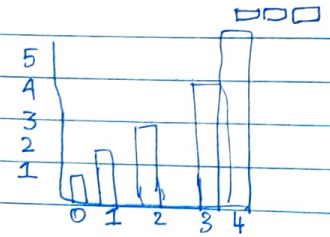
```
from bokeh.charts import Bar, output_file, show

data = {'y': [1, 2, 3, 4, 5]}

output_file('bar.html', title='Bar Chart Example')

p = Bar(data, title='Bar chart example',
        xlabel='x', ylabel='values', width=400,
        height=400)
show(p)
```

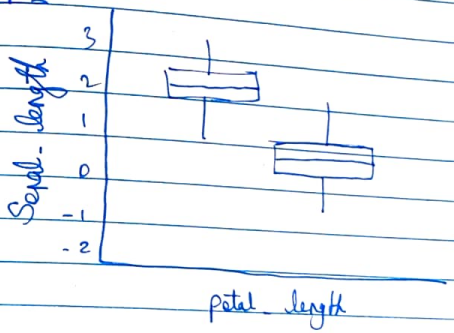
Output



Compare distribution of sepal length & petal length of iris dataset using box plot on notebook

```
from sklearn.datasets import load_iris
import pandas as pd
iris = load_iris()
df = pd.DataFrame(iris)
df.columns = ['petal width', 'petal length', 'sepal width',
              'sepal length']
from bokeh.charts import Boxplot, output_notebook, show
data = df[['petal length', 'sepal length']]
output_notebook()
p = Boxplot(data, width=400, height=400)
show(p)
```

Output

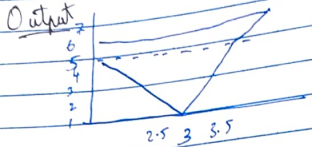


Create line plot on bokeh server

```
>bokeh-server
from bokeh.plotting import figure,
out_server('line') output_server, show
```

```
p = figure (plot_width = 400, plot_height = 400)
p.line ([5,2,3,4,5], [5,7,2,4,5], line_width=2)
show(p)
```

Output



Create a square mark on xy frame of notebook

Combining line plots

```
from bokeh.plotting import figure, output_notebook, show
output_notebook()
p = figure (plot_width = 400, plot_height = 400)
p.square ([2,5,6,4], [2,3,2,1,2], size=20, color='navy')
p.line ([1,2,3,4,5], [2,2,2,2,2], line_width=2)
show(p)
```

Output