Big Data: Large amounts of data
→ Data that is so large and complex that it is difficult to process using traditional methods
→ Includes structured, unstructured, semi structured data
→ act of storing and analysing
↓ cloud ↓ for insights

Types of data
Structured: Table, pdfs, word [CSV, Relational data]
↳ Stored, accessed with fixed format
Unstructured: Google search results
↳ might have missing / wrong values, unknown form
semistructured: XML, HTML, <tag>
↳ combination

why is big data important?
i) cost reduction due to storing data in hadoop etc.
ii) Time reduction, speed in memory analysis
iii) new product development & optimisation
iv) smart decision making
v) multiple key technologies like ML, AI, Mining etc.

Characteristics
3V: Volume, velocity (speed & data generation), Variety (heterogenous data)
extra 2V: Veracity, Value
↓
Validating Big Data: Find out feasibility, reasonability, value, integrity and sustainability to validate data and accept a technique and tool.

class label is in supervised learning and it is the supervisor

Uses: Healthcare, Govmt, Education, Telecommunication, Manufacture, Financial Sector.

Different sources of data
→ sensor, social media, E-commerce, transaction, log, web-related data.

Veracity: verify truthfulness of data.
Value: enhances business profit by providing decision making systems.

Traditional Data	Big Data
→ manual data generation	→ machine generates data
→ standard format	→ messy and ugly format
→ able to store and process with the help of traditional database management tools	→ unable to capture & store & process using traditional DB management tools within elapsed time

Then Vs Now
Healthcare → Paper
→ EHR (Electronic Health Record)
↙ MRI ↘ Lab ...

Telecommunication: Bandwidth, call rate, N/W capacity, number of users, customer Feedback, demographic location
Government: aadhar, job, income, tax
Education: Students count, interest, age,
Financial sector: Bank account, credit score, loan required, income, profession, age, address, phone
→ targeted marketing

To process large data, one can replace a single high end computer with many many computers with resources.

erce, transaction, log.

data.
providing decision

Big Data
→ machine generates data
→ messy and ugly format
→ unable to capture & store & process using traditional DB managem -ent tools within elapsed time

ealth Record)

h, call rate, N/W capacity
Feedback, demographic

me, tax
rest, age,
, credit score, loan
ge, address, phone

can replace a
many many computers

Distributed computation
writing distributed pgms require high level of expertise } Issues and Challenges

Points to be considered when designing a sys
i) Scalable ii) Cost effective iii) Flexibility
iv) Ease of use

HDFS : Hadoop Distributed File System
YARN : yet another Resource Navigator
Map Reduce : Programming Model to handle big data

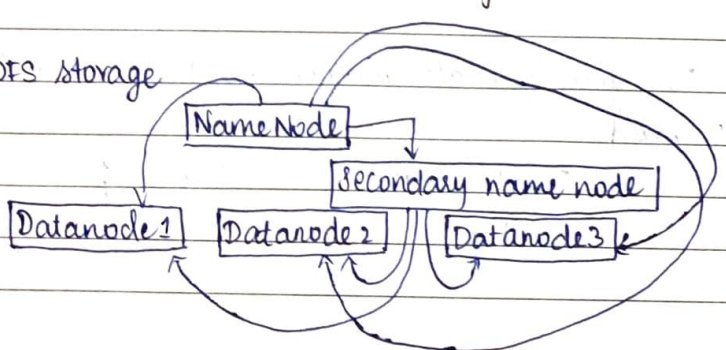
Hbase	Flink	Giraph	zookeeper
Cassandra	spark	Hive	
MongoDB	storm	Pig	

Hadoop : A scalable, fault tolerant, cost effective open source software that handles large data in a distributed system

Hadoop
HDFS Map Reduce (MR)

Map and Reduce
→ 2 phases : mapper phase and reducer phase
→ in between these phases there is shuffle and sort phase
→ programmer needs to identify those phases.
→ output of phases is key-value pair

HDFS storage



Namenode:

One of the three main components in HDFS storage

It is the master

① Master Daemon

② Metadata information

→ Size of file

→ Permission of file

→ Location of a file

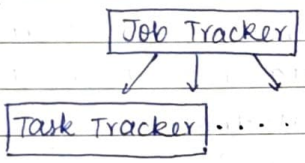
③ Receives heartbeat from the datanode

Datanode

① Serving read requests and write requests from the client

② Actual data is stored

Map Reduce



(ie master allocates job to tasks

hdfs-site.xml

	Hadoop V1	Hadoop V2
Large I/P file is divided into	64 mb	128 mb

QN 1: How many ~~tasks~~ blocks will be created if file of size 514 mb is copied to hdfs?

If you allocate hadoop V2 then we will need 5 blocks

Block 1: 128 ..... Block 4: 128,

Block 5: 2 mb

Replication factor (Defau
in 3 different places.

Rack Awareness Algorithm

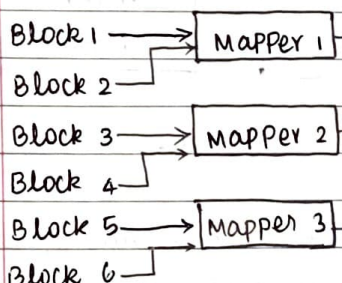
→ Blocks should
are replication data.

RACK 1	RACK 2
Datanode 1 Block 1	Datanode Block 2
Datanode 3	Datanode Block 3 (r
Datanode 5	Datanod Block 3 (r

→ original and
racks.

Map Reduce Programm

→ Divide &



Divide and
conq

Replication factor (Default) is 3. Blocks can be replicated in 3 different places.

Rack Awareness Algorithm

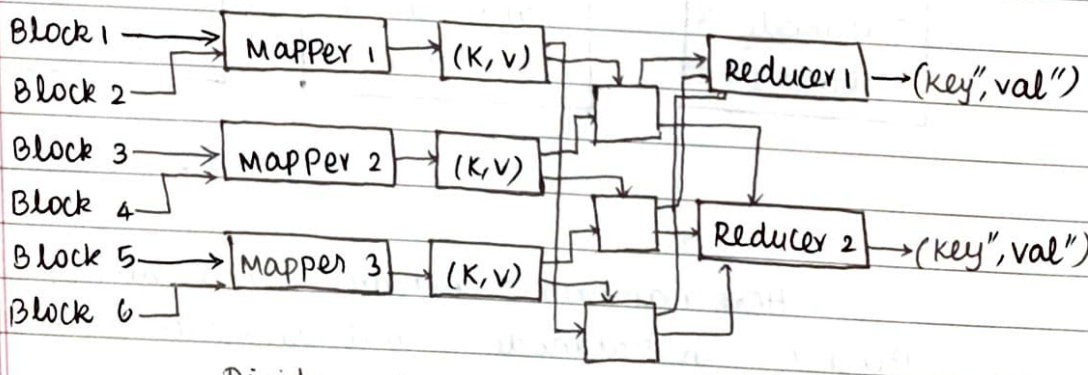
→ Blocks should come under same rack if they are replication data. For example,

RACK 1	RACK 2	RACK 3	RACK 4
Datanode 1 Block 1	Datanode 2 Block 2	Datanode 7 Block 1 (rep)	Datanode 8 Block 3
Datanode 3	Datanode 4 Block 3 (rep)	Datanode 9 Block 1 (rep)	Datanode 10 Block 2 (rep)
Datanode 5	Datanode 6 Block 3 (rep)	Datanode 11	Datanode 12 Block 2 (rep)

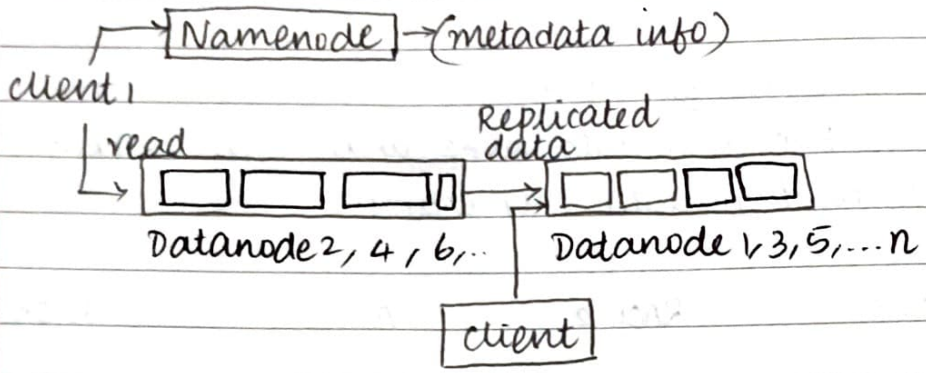
→ Original and replication data must be in diff racks.

Map Reduce Programming Paradigm

→ Divide & Conquer Approach



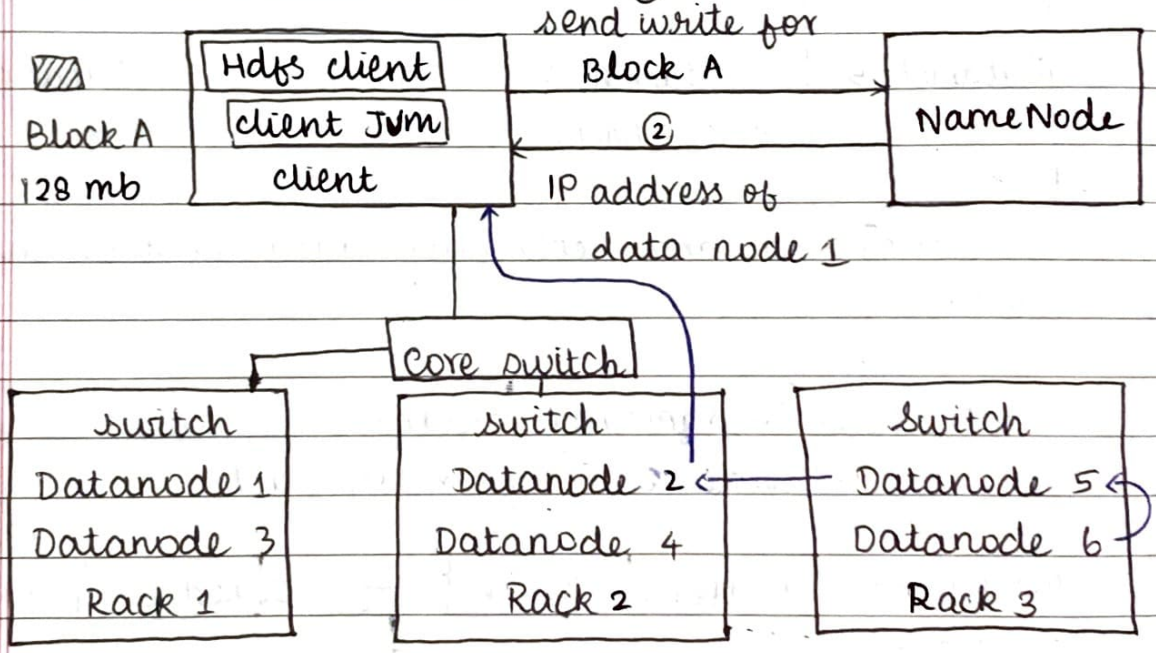
hdfs architecture



Hdfs read/write mechanism

→ hdfs write mechanism
→ hdfs write mechanism with ACK
→ hdfs multi write pipeline : has many blocks
→ hdfs read mechanism : sends read request @ ①

Hdfs write mechanism with ACK



→ Acknowledgement

Here name node contains size, location, permisno Block A is in Datanode 2 and replications are in datanode 5 and 6.

Classification (Type of Supervised learning)

uses classification algorithm/model
- ↳ decision tree algorithm
- ↳ naive based Bayes algorithm
- ↳ support vector machine
- ↳ neural network

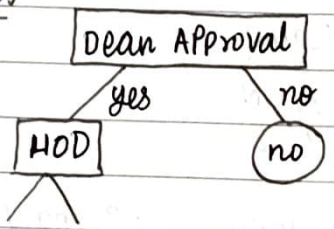
Decision Trees (Greedy approach)

consists of Internal nodes, leaves, branches

attribute — class label

Root node: Attribute that provides maximum information

Example (IV or not)



Attribute Selection

Gini Index — Info Gain : Entropy formula

$$Gain(D) - Gain_A(D) \qquad \sum_{i=1}^{n} -P_i \log_2 P_i$$

Example

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
1	Sunny	Hot	High	weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	weak	yes
4	Rainy	Mild	High	weak	Yes
5	Rainy	cool	Normal	weak	yes
6	Rainy	cool	Normal	strong	No
7	Overcast	Cool	Normal	strong	Yes
8	Sunny	~~High~~ Mild	High	weak	No
9	Sunny	Cool	Normal	weak	Yes
10	Rainy	Mild	Normal	weak	yes

11	Sunny	~~High~~ Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rainy	~~High~~ Mild	High	Strong	No

Solution

Step 1: Find Entropy

$$\text{Entropy} = \frac{-9}{14}\log_2\left(\frac{9}{14}\right) - \frac{5}{14}\log_2\left(\frac{5}{14}\right) = 0.94$$

Step 2: Find 1st highest information gain

① Outlook

Sunny Overcast Rainy

Y ∧ N Y ∧ N Y ∧ N

2 3 4 0 3 2

$$\text{Entropy}(\text{Outlook} = \text{Sunny}) = \frac{-2}{5}\log_2\left(\frac{2}{5}\right) - \frac{3}{5}\log_2\left(\frac{3}{5}\right)$$

$$= 0.971$$

$$\text{Entropy}(\text{Outlook} = \text{Overcast}) = \frac{-4}{4}\log_2\left(\frac{4}{4}\right) = 0$$

$$\text{Entropy}(\text{Outlook} = \text{Rainy}) = \frac{-3}{5}\log_2\left(\frac{3}{5}\right) - \frac{2}{5}\log_2\left(\frac{2}{5}\right)$$

$$= 0.971$$

$$\text{Gain}_{\text{Outlook}} = \frac{5}{14} \times 0.971 + \frac{4}{14} \times 0 + \frac{5}{14} \times 0.971$$

$$= \frac{5}{14} \times 0.971 \times 2 = \frac{5}{7} \times 0.971 = 0.693$$

$$\text{Info Gain}(\text{Outlook}) = 0.94 - 0.693 = 0.247$$

② Temperature

Hot ~~High~~ Mild Cold

2 ∧ 2 4 ∧ 2 3 ∧ 1

Entropy(Temp = Hot) = 1

Entropy(Temp = Mild) = $\frac{-4}{6}\log_2\left(\frac{4}{6}\right) - \frac{2}{6}\log_2\left(\frac{2}{6}\right) = 0.910$

Entropy(Temp = Cold) = $\frac{-3}{4}\log_2\left(\frac{3}{4}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right) = 0.811$

Gain$_{Temp}$ = $\frac{4}{14} \times 1 + \frac{6}{14} \times 0.910 + \frac{4}{14} \times 0.811$

= 0.911

Info Gain(Temp) = 0.94 − 0.911 = 0.0292

③ Humidity

Normal | High

6 1 | 3 4

Entropy(Humidity = Normal) = $\frac{-6}{7}\log_2\left(\frac{6}{7}\right) - \frac{1}{7}\log_2\left(\frac{1}{7}\right)$

= 0.591

Entropy(Humidity = High) = $\frac{-3}{7}\log_2\left(\frac{3}{7}\right) - \frac{4}{7}\log_2\left(\frac{4}{7}\right)$

= 0.983

Gain$_{Humidity}$ = $\frac{7}{14} \times 0.591 + \frac{7}{14} \times 0.983$

= 0.787

Info Gain(Humidity) = 0.94 − 0.787 = 0.153

④ wind

Strong | weak

3 3 | 6 2

Entropy(wind = weak) = $\frac{-6}{8}\log_2\left(\frac{6}{8}\right) - \frac{2}{8}\log\left(\frac{2}{8}\right)$

= 0.811

Entropy(wind = Strong) = 1

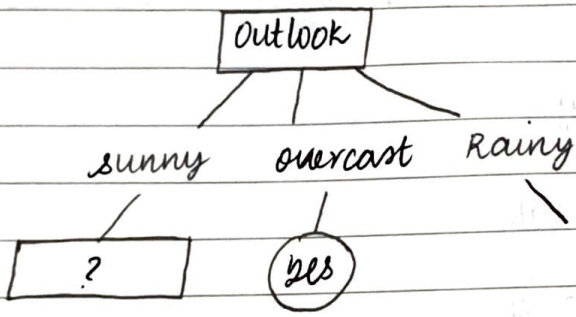
$Gain_{wind} = \frac{6}{14} \times 1 + \frac{8}{14} \times 0.811$

$= 0.892$

Info gain (wind) $= 0.94 - 0.892$

$= 0.048$

Step 3 : Find root



Entropy

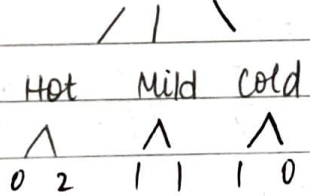
$= \frac{-2}{5} \log_2\left(\frac{2}{5}\right)$

$- \frac{3}{5} \log_2\left(\frac{3}{5}\right)$

$= 0.971$

Repeat Step 2 :

① Temperature



Entropy (Temp = Hot) = 0

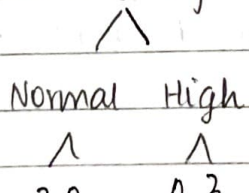
Entropy (Temp = Mild) = 1

Entropy (Temp = Cold) = 0

Gain (Temp) $= 1 \times \frac{2}{5} = \frac{2}{5} = 0.4$

Info Gain (Temp) $= 0.971 - 0.4 = 0.571$

② Humidity

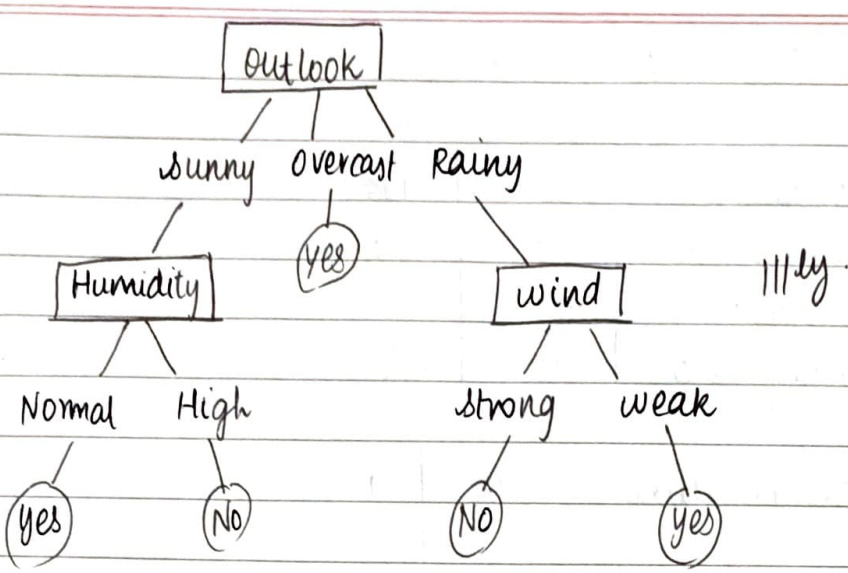


Entropy (Humid = Normal) = 0

Entropy (Humid = High) = 0

Gain (Humidity) = 0

Info Gain (Humidity) = 0.971



IIIly.

② Bayesian Classifier (Supervised learning)

i) Based on class independence : the effect of values of attribute is independent of values of another attribute

(e) Age, Income, student/not are independent.

ii) Each feature equally contributes / has equal importance in predicting the class label

Baye's Theorem :

$$P(\text{Hypothesis})(X_1, X_2, \ldots X_n) = \frac{P(X/H) \cdot P(H)}{P(X)}$$

$(X_1, X_2, \ldots X_n)$ ↓ given sample tuple

$P(X)$ ↓ evidence

Posterior probability

likelihood

prior probability

$$P(C/X_i) = P(C/X_1) * P(C/X_2) * P(C/X_3) \ldots * P(\text{class})$$

$$P(C/X_i) = \sum_{i=1}^{n} P(C/X_i)$$

$$P(C/X_i) \propto \sum_{i=1}^{n} P(C/X_i)$$

$$P(C/X_i) = \text{argmax}\ P(C/X_i)$$

Let us assume Play Tennis dataset

Test data : Outlook Temperature Humidity wind

sunny Hot Normal weak

$P(Yes) = \frac{9}{14}$ $P(No) = \frac{5}{14}$

Outlook	Yes	No
Sunny = 5	2/9	3/5
Overcast = 4	4/9	0/5
Rainy = 5	3/9	2/5

Temperature	Yes	No
Hot = 4	2/9	2/5
Mild = 6	4/9	2/5
Cold = 4	3/9	1/5

Humidity	Yes	No
Normal = 7	6/9	1/5
High = 7	3/9	4/5

wind	Yes	No
strong = 6	3/9	3/5
weak = 8	6/9	2/5

P(Yes / Outlook = sunny, Temp = hot, humidity = normal, wind = weak)

= Likelihood × Prior Probability / Evidence

$= \frac{P(Outlook=sunny)}{P(Yes)} \times \frac{P(Temp=hot)}{P(yes)} \times \frac{P(Humid=N)}{P(yes)} \times \frac{P(wind=weak)}{P(yes)} \times \frac{9}{14}$

$= \frac{2}{9} \times \frac{2}{9} \times \frac{6}{9} \times \frac{6}{9} \times \frac{9}{14}$

$= 8/567 \simeq 0.014$

P(No / Outlook = sunny, Temp = hot, humidity = normal, wind = weak)

$= \frac{P(O=sunny)}{P(No)} \times \frac{P(T=hot)}{P(no)} \times \frac{P(Humid=Normal)}{P(no)} \times \frac{P(wind=w)}{P(no)} \times P(no)$

$= \frac{3}{5} \times \frac{2}{5} \times \frac{1}{5} \times \frac{2}{5} \times \frac{5}{14} = \frac{6}{875}$

$\simeq 0.006$

∴ For the given data, PlayTennis = YES

For BuysComputer Dataset

1/4/22

K-Means clustering (Unsupervised learning)

Within a cluster, similarities are high

Application of clustering

i) Outlier detection Eg: fraud detection / credit card
ii) grouping for google search

Clustering can also be called cluster analysis / data segmentation

Challenges

i) Scalability
ii) Need to consider dealing with high dimensional data
iii) Ability to deal with data of arbitrary shape (types)
iv) Sensitive to noise
v) Constraint based
vi) Different set of attributes
vii) Choosing for input parameters
viii) Incremental updates

iii) Discovery of clusters with arbitary shape

iv) Requirement for domain knowledge to determine input parameters
v) Ability to deal with noisy data
vi) Incremental clustering & insensitive to input
vii) Capability of clustering high dimensional data
viii) Interpretability and usability
ix) Constraint based clustering

Clustering algo Types → Partitioning clustering Eg: K-means
→ Hierarchial clustering
→ Density based clustering
→ Grid based clustering

These algos should cover the following
i) Seperation of clusters ii) Similarity measure
iii) Clustering method [Euclidean, Manhattan, Jacard distance etc.]
iv) partitioning criteria

Intracluster distance: within a cluster (must be low)
Intercluster distance: datapoints b/w two clusters (must be high)
→ similarity high
similarity is low

Euclidean distance
$= \sqrt{(x_2-x_1)^2+(y_2-y_1)^2}$

K-means clustering
i) Assume a centroid value for some cluster
ii) Find distance b/w centroid and datapoint.
iii) Put datapoint under cluster where distance ii) is small

Qn: Cluster the following 8 points [with (x,y) representing location] into three clusters.

$A_1(2,10)$ $B_1(5,8)$ $C_1(1,2)$
$A_2(2,5)$ $B_2(7,5)$ $C_2(4,9)$
$A_3(8,4)$ $B_3(6,4)$

Distance function is Euclidean distance. Suppose initially we assign A_1, B_1, and C_1 as the center of each cluster respectively, use K-means algorithm to show
a) 3 cluster centers after first round of execution
b) Final 3 clusters

FIRST ROUND
K=3; Centroids are A_1, B_1, C_1

cluster	1	2	3
	A_1	A_3	A_2
		B_1	C_1
		B_2	
		B_3	
		C_2	

$A_2(2,5)$
i) $A_1(2,10)$
Distance $= \sqrt{(2-2)^2+(10-5)^2} = \sqrt{25} = 5$
ii) $B_1(5,8)$
Distance $= \sqrt{(5-2)^2+(8-5)^2} = \sqrt{9+9} = 4.2$
iii) $C_1(1,2)$
Distance $= \sqrt{(2-1)^2+(5-2)^2} = \sqrt{1+9} = \sqrt{10} = 3.1$
$\Rightarrow A_2$ belongs to cluster 3

$A_3(8,4)$
i) $A_1(2,10)$
Distance $= \sqrt{(8-2)^2+(4-10)^2} = \sqrt{36+36} = 10.09$
ii) $B_1(5,8)$
Distance $= \sqrt{(8-5)^2+(4-8)^2} = \sqrt{9+16} = 5$
iii) $C_1(1,2)$
Distance $= \sqrt{(8-1)^2+(4-2)^2} = \sqrt{49+4} = 7.2$
$\Rightarrow$ Belongs to cluster 2

$B_2(7,5)$
i) $A_1(2,10)$
Distance $= \sqrt{(7-2)^2+(5-10)^2} = \sqrt{25+25} = 7.07$

ii) $B_1(5,8)$

$D = \sqrt{(7-5)^2+(5-8)^2} = \sqrt{4+9} = 3.6$

iii) $C_1(1,2)$

$D = \sqrt{(7-1)^2+(5-2)^2} = \sqrt{36+9} = 6.7$

$\Rightarrow$ Cluster 2

$B_3(6,4)$

i) $A_1(2,10)$

$D = \sqrt{(6-2)^2+(4-10)^2} = \sqrt{16+36} = 7.2$

ii) $B_1(5,8)$

$D = \sqrt{(6-5)^2+(4-8)^2} = \sqrt{1+16} = 4.12$

iii) $C_1(1,2)$

$D = \sqrt{(6-1)^2+(4-2)^2} = \sqrt{25+4} = 5.3$

$\Rightarrow$ cluster 2

$C_2(4,9)$

i) $A_1(2,10)$

$D = \sqrt{(4-2)^2+(9-10)^2} = \sqrt{4+1} = \sqrt{5} = 2.2$

ii) $B_1(5,8)$

$D = \sqrt{(5-4)^2+(8-9)^2} = \sqrt{2} = 1.4$

iii) $C_1(1,2)$

$D = \sqrt{(4-1)^2+(9-2)^2} = \sqrt{9+49} = 7.6$

$\Rightarrow$ cluster 2

Centroid of cluster 1 : $A_1(2,10)$

2 : ~~$(5.4, 6)$~~ $(6,6)$

3 : $(1.5, 3.5)$

$\left(\frac{5+8+7+6+4}{5}, \frac{8+4+5+4+9}{5}\right)$

~~$(5.4, 6)$~~ $(6,6)$

$\left(\frac{1+2}{2}, \frac{2+5}{2}\right)$

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

$=$

SECOND ROUND

centroid	(2,10)	(6,6)	(1.5,3.5)
clusters	1	2	3
	A1	A3	A2
	C_2	B1	C_1
		B2	
		B3	

$A_2(2,5)$

i) $A_1(2,10)$

$D = \sqrt{0+5^2} = 5$

ii) (6,6) $D = \sqrt{(6-2)^2+(6-5)^2} = \sqrt{16+1} = 4.12$

iii) (1.5, 3.5)

$D = \sqrt{(2-1.5)^2+(5-3.5)^2} = 1.58$ //

$A_3(8,4)$

i) $D = \sqrt{(8-2)^2+(4-10)^2} = \sqrt{36+36} = 10.09$

ii) $D = \sqrt{(8-6)^2+(4-6)^2} = \sqrt{4+4} = 2.8$ //

iii) $D = \sqrt{(8-1.5)^2+(4-3.5)^2} = 6.51$

$B_1(5,8)$

i) $D = \sqrt{(5-2)^2+(10-8)^2} = \sqrt{9+4} = 3.6$

ii) $D = \sqrt{(5-6)^2+(8-6)^2} = \sqrt{1+4} = 2.23$ //

iii) $D = \sqrt{(5-1.5)^2+(8-3.5)^2} = 5.7$

$B_2(7,5)$

i) $D = \sqrt{(7-2)^2+(5-10)^2} = \sqrt{25+25} = 7.07$

ii) $D = \sqrt{(7-6)^2+(5-6)^2} = \sqrt{2} = 1.4$ //

iii) $D = \sqrt{(7-1.5)^2+(5-3.5)^2} = 5.70$

$C_1(1,2)$

i) $D = \sqrt{(2-1)^2+(10-2)^2} = \sqrt{1+64} = 8.06$

ii) $D = \sqrt{(6-1)^2+(6-2)^2} = \sqrt{25+16} = 6.4$

iii) $D = \sqrt{(1.5-1)^2+(3.5-2)^2} = 1.5$ //

$C_2(4,9)$

i) $D = \sqrt{(4-2)^2+(9-10)^2} = 2.2$ //
ii) $D = \sqrt{(6-4)^2+(6-9)^2} = \sqrt{4+9} = 3.6$
iii) $D = \sqrt{(4-1.5)^2+(9-3.5)^2} = 6.04$

clusters
i) $\left(\frac{2+4}{2}, \frac{10+9}{2}\right) = (3, 9.5)$

ii) $\left(\frac{5+8+7+6}{4}, \frac{8+4+5+4}{4}\right) = (6.5, 5.25)$

iii) $(1.5, 3.5)$

THIRD ROUND

Centroid: (3, 9.5) (6.5, 5.25) (1.5, 3.5)
Clusters: 1: A1, B1, C2 — 2: A3, B2, B3 — 3: A2, C1

A1 (2, 10)
i) $\sqrt{(3-2)^2+(9.5-10)^2} = \sqrt{1.25} = 1.1$ //
ii) $\sqrt{(6.5-2)^2+(5.25-10)^2} = \sqrt{42.8} =$
iii) $\sqrt{(2-1.5)^2+(10-3.5)^2} = 6.51$

A2 (2, 5)
i) $\sqrt{(3-2)^2+(9.5-5)^2} = 4.60$
ii) $\sqrt{(6.5-2)^2+(5.25-5)^2} = 4.50$
iii) $\sqrt{(2-1.5)^2+(5-3.5)^2} = 1.58$ //

A3 (8, 4)
i) $\sqrt{(8-3)^2+(4-9.5)^2} = 7.43$
ii) $\sqrt{(8-6.5)^2+(4-5.25)^2} = 1.95$ //

iii) $\sqrt{(8-1.5)^2+(4-3.5)^2} = 6.51$

B1 (5, 8)
i) $\sqrt{(5-3)^2+(8-9.5)^2} = 2.06$ //
ii) $\sqrt{(6.5-5)^2+(5.25-8)^2} = 3.13$
iii) $\sqrt{(5-1.5)^2+(8-3.5)^2} = 5.70$

B2 (7, 5)
i) $\sqrt{(7-3)^2+(5-9.5)^2} = 6.02$
ii) $\sqrt{(7-6.5)^2+(5-5.25)^2} = 0.55$ //
iii) $\sqrt{(7-1.5)^2+(5-3.5)^2} = 5.70$

B3 (6, 4)
i) $\sqrt{(6-3)^2+(4-9.5)^2} = 6.26$
ii) $\sqrt{(6.5-6)^2+(5.25-4)^2} = 1.34$ //
iii) $\sqrt{(6-1.5)^2+(4-3.5)^2} = 4.53$

C1 (1, 2)
i) $\sqrt{(3-1)^2+(9.5-2)^2} = 7.76$
ii) $\sqrt{(6.5-1)^2+(5.25-2)^2} = 6.38$
iii) $\sqrt{(1.5-1)^2+(3.5-2)^2} = 1.51$ //

C2 (4, 9)
i) $\sqrt{(4-3)^2+(9-9.5)^2} = 1.11$ //
ii) $\sqrt{(6.5-4)^2+(5.25-9)^2} = 4.5$
iii) $\sqrt{(4-1.5)^2+(9-3.5)^2} = 6.04$

R-Programming

- ↳ Statistical Programming language
- ↳ graphical visualization
- ↳ used to develop Machine learning algo
- ↳ open source software
- ↳ platform independent
- ↳ packages

Assign variables

a ← 5

subject = "bda"

Datatypes

* Number → int eg: x ← 10L
  → numeric eg: m = 7.0
* complex 7i+9
* character " "
  ↓ single quotes automatically becomes double
* Boolean/Logical True or false

Data structures

i) vector
ii) List
iii) Matrix
iv) Array
v) Dataframe
vi) Factors

i) c() → c function stands for concatenate or combine
ii) supports different datatype

iii) matrix (c(elements, nrow=2, ncol=3))
iv) array (c(elements, dim = (row, column, dimensions)))
v) data.frame ( ) → create tabular data
vi) Shows levels which are distinct values in the factor

Name1 Association Rule mining algorithm

→ helps for product suggestion or recommendation
→ finding frequent item set which leads to the discovery of association rules and correlations

- → support value — no. of times item is present in db
- → confidence value — % of interest in buying things together

Name2 Eg: Market-Basket Analysis

- low: large number of items can be retrived
- high: nothing will be there

Two Step Process

↳ find the frequent itemset
↳ association rules from the frequent itemset

Name3 Apriori algorithm : Iterative approach

Scan the entire database and generate C1: candidate frequent itemset and L1: first frequent itemset.
The second frequent itemset L2 can be created by joining operation

C_K can be generated by joining L_{K-1}

* Join property and prune property

Example:

Transaction Id	List of items
100	I_1 I_2 I_5
101	I_2 I_4
103	I_2 I_3
104	I_1 I_2 I_4
105	I_1 I_3
106	I_2 I_3
107	I_1 I_3
108	I_1 I_2 I_3 I_5
109	I_1 I_2 I_3

suppose min support count = 2
min confidence value = 70%

Steps

i) Generate first frequent itemset

$I_1 \rightarrow 6$			$I_1 \rightarrow 6$
$I_2 \rightarrow 7$			$I_2 \rightarrow 7$
$I_3 \rightarrow 6$	$\Rightarrow$		$I_3 \rightarrow 6$
$I_4 \rightarrow 2$	Support = 2		$I_4 \rightarrow 2$
$I_5 \rightarrow 2$			$I_5 \rightarrow 2$
C1			L1

first frequent itemset

ii) Generate second frequent itemset (Join operation)

$I_1, I_2 \rightarrow 4$		$I_1, I_2 \rightarrow 4$
$I_1, I_3 \rightarrow 4$		$I_1, I_3 \rightarrow 4$
$I_1, I_4 \rightarrow 1$		$I_1, I_5 \rightarrow 2$
$I_1, I_5 \rightarrow 2$		$I_2, I_3 \rightarrow 4$
$I_2, I_3 \rightarrow 4$	$\longrightarrow$	$I_2, I_4 \rightarrow 2$
$I_2, I_4 \rightarrow 2$	support = 2	$I_2, I_5 \rightarrow 2$
$I_2, I_5 \rightarrow 2$		
$I_3, I_4 \rightarrow 0$		L2
$I_3, I_5 \rightarrow 1$		
$I_4, I_5 \rightarrow 0$		
C2		

candidate pair

iii) Generate third frequent itemset

$I_1, I_2, I_3 \rightarrow 2$		$I_1\ I_2\ I_3 \rightarrow 2$
$I_1, I_2, I_5 \rightarrow 2$	support = 2	$I_1\ I_2\ I_5 \rightarrow 2$
~~I_1, I_3, I_5~~	$\longrightarrow$	
$I_1, I_2, I_4 \rightarrow 1$		
~~$I_1, I_3, I_5 \rightarrow 1$~~		F3
$I_1, I_3, I_2, I_4 \rightarrow 0$		
$I_1, I_3, I_2, I_5 \rightarrow 1$		
C3		

not all subset are in L2, so delete

iv) Generate fourth frequent itemset

$I_1\ I_2\ I_3\ I_5 \Rightarrow 1$ X $\longrightarrow$ support = 2 $\quad C_4 = \phi \quad F_4 = \phi$

(I_2, I_3, I_5) is not there (ie) prune property

v) Generate association rules from frequent item set

Frequent item set $\{I_1\}$ $\{I_2\}$ $\{I_3\}$ $\{I_4\}$ $\{I_5\}$ $\{I_1, I_2\}$ $\{I_1, I_3\}$ $\{I_1, I_5\}$ $\{I_2, I_3\}$ $\{I_2, I_4\}$ $\{I_2, I_5\}$ $\{I_1, I_2, I_3\}$ $\{I_1, I_2, I_5\}$

Let us consider I_1, I_2, I_5

subsets are $\{I_1, I_2\}$ (I_2, I_5) (I_1, I_5) (I_1) (I_2) (I_5)

Min Conf = 70

Rule 1: $I_1, I_2 \rightarrow I_5$

$$\frac{\text{support}(I_1, I_2 I_5)}{\text{support}(I_1, I_2)} = \frac{2}{4} = \frac{1}{2} = 50\% \quad \text{X}$$

Rule 2: $I_1, I_5 \rightarrow I_2$

$$\frac{\text{Support}(I_1, I_2, I_5)}{\text{support}(I_1, I_5)} = \frac{2}{2} = 1 = 100\% \quad \checkmark$$

Rule 3: $I_2, I_5 \rightarrow I_1$

$$\frac{\text{Support}(I_1, I_2, I_5)}{\text{Support}(I_2, I_5)} = \frac{2}{2} = 1 = 100\% \quad \checkmark$$

Rule 4: $I_1 \rightarrow I_2, I_5$

$$\frac{\text{support}(I_1, I_2, I_5)}{\text{Support}(I_1)} = \frac{2}{6} = \frac{1}{3} = 33.3\% \quad \text{X}$$

Rule 5: $I_2 \rightarrow I_1, I_5$

$$\frac{\text{Support}(I_1, I_2, I_5)}{\text{support}(I_2)} = \frac{2}{7} = 28.5\% \quad \text{X}$$

Rule 6: $I_5 \rightarrow I_1, I_2$

$$\frac{\text{Support}(I_1, I_2, I_5)}{\text{support}(I_5)} = \frac{2}{2} = \frac{1}{1} = 33.3\% \ 100\% \quad \checkmark$$

⇒ Association Rules

I1, I5 → I2

I2, I5 → I1

I5 → I1, I2

If support value = 5% then do $\frac{5}{100}$ and work.

Qn: Suppose the association rule hot dogs → hamburgers is mined. give a min support threshold of 25% and a minimum confidence of 50% does the association mining hold

	hotdog	$\overline{\text{hotdog}}$	Total
hamburger	2000	500	2500
$\overline{\text{hamburger}}$	1000	1500	2500
Total	3000	2000	5000

hotdogs → hamburgers

$$\frac{\text{support(hotdog, hamb)}}{\text{support(hotdog)}} = \frac{2000}{3000} = \frac{2}{3} \Rightarrow 66.6\%$$

Min confidence = 50%

⇒ Association holds

21/4

in RStudio

Packages (Decision Tree)

caret : classification & regression

rpart, e1071

26/4

Assignment 1

Basics of R programming: Conditional stmt, Looping, Function, array, matrix, vector, factors, file read write

Descriptive analysis and Predictive analysis

Decision Tree algo: Manual dataset (handwritten code + SS) & benchmark (just SS)

Naive Bayes

Qn): A database has five transactions. Let minimum support = 60% and min confidence = 80%.

$\frac{60}{100} \times 5 = 3$

Transaction Id	items bought
T100	{M, O, N, K, E, Y}
T200	{D, O, N, K, E, Y}
T300	{M, A, K, E}
T400	{M, U, C, K, Y}
T500	{C, O, O, K, I, E}

⇒ 27

i) Find all frequent item sets using apriori algorithm.

ii) List all the strong association rules with support s and confidence c matching the following meta-rule where x is a variable representing customers, $item_i$ denotes variables representing items

$\forall x \in$ transaction, $buys(x, item_1) \wedge buys(x, item_2) \Rightarrow buys(x, item_3)$

Ans: First frequent itemset

M → 3		M → 3
O → 4		O → 4
N → 2		K → 5
K → 5		E → 4
E → 4	→ Support = 3	Y → 3
Y → 3		
D → 1		
A → 1		
U → 1		
C → 2		
I → 1		

Second iteration

M, O → 1	O, K → 3	K, Y → 3		M, K → 3
M, K → 3	O, E → 3	E, Y → 2	→ S = 3	O, K → 3
M, E → 2	O, Y → 2			O, E → 3
M, Y → 2	K, E → 4			K, E → 4
				K, Y → 3

Third iteration

Eliminating by prune

~~M K O~~ →
~~M K O E~~ →
~~M K E~~ →
~~M K Y~~ →
O K E → 3
~~O K Y~~ →
~~K E Y~~ →
~~O E K Y~~ →

→ $s = 3$

OKE → 3
Third frequent itemset

OKE by prune property ϕ generate association rules

Frequent itemset {O} {K} {E} {OK} {KE} {OE} for {OKE}

Let us consider

i) O, K → E

$$\frac{\text{Support}(OKE)}{\text{support}(OK)} = \frac{3}{3} = 100\%$$

ii) O, E → K

$$\frac{\text{Support}(OKE)}{\text{Support}(OE)} = \frac{3}{3} = 100\%$$

iii) K, E → O

$$\frac{\text{Support}(OKE)}{\text{Support}(KE)} = \frac{3}{4} = 75\%$$

iv) Confidence value 80%

Association rules

O, K → E
O, E → K

Q2: Database has four transactions. Let min support = 60% min confidence = 80%

CustId	TId	Items_bought
01	T100	~~{A, B, C, D}~~ {KC, SM, DC, BB}
02	T200	{BC, DM, GA, TP, WB}
01	T300	{WA, DM, WB, TP}
03	T400	{WB, SM, DC}

KC - King's crab
SM - Sunset milk
DC - Dairyland cheese
BB - Best Bread
BC - Best cheese
DM - Dairyland Milk
GA - Goldenfarm Apple
TP - Tasty Pie
WB - Wonder Bread
WA - Westcoast Apple

i) At the granularity of item category (item could be milk) for the following rule template

$\forall x \in$ transaction, buys (X, item$_1$) ∧ buys (X, item$_2$) ⇒ buys (X, item$_3$)

Find and list k frequent itemset and all the strong association rules with a support s and confidence c.

ii) At the granularity of brand item category (item could be sunset milk) following rule template

$\forall_x \in$ customer, buys(x, item$_1$) ∧ buys (X, item$_2$) ⇒ buys (X, item$_3$)

① Item category granularity

			Support val
01	T100	{G, M, C, B}	$\frac{60}{100} \times 4$
02	T200	{C, M, A, P, B}	⇒ $= 2.4$
01	T300	{A, M, B, P}	
03	T400	{B, M, C}	

First freq.

CY → 1
M → 4
C → 3
B → 4
A → 2
P → 2

Support = 2.4

M → 4
C → 3
B → 4

Second freq

M,C → 3
M,B → 4
C,B → 3

Support = 2.4

M,B → 4
C,B → 3
M,C → 3

Third freq

M,B,C → 3

Support = 2.4

ϕ

Freq item set

{M} {B} {C}
{M,B,C}
{M,B} {M,C} {B,C}

M,B ⇒ C

$$\frac{\text{Support}(M,B,C)}{\text{Support}(M,B)} = \frac{3}{4} = 75\%$$

M,C ⇒ B

$$\frac{\text{Support}(M,B,C)}{\text{Support}(M,C)} = \frac{3}{3} = 100\%$$

B,C ⇒ M

$$\frac{\text{Support}(M,B,C)}{\text{Support}(B,C)} = \frac{3}{3} = 100\%$$

Association Rule

M,C ⇒ B
B,C ⇒ M

② Brand item granularity

First freq.

KC → 1
SM → 2
DC → 2
BB → 1
BC → 1
DM → 2
GA → 1
TP → 2
WB → 3
WA → 1

Support = 2.4 →

WB → 3

Antecedent → Consequent

Eg: Milk → Bread

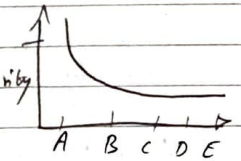
$$\text{Lift} = \frac{\text{Support}(\text{Milk, Bread})}{\text{Support}(\text{Milk}) * \text{Support}(\text{Bread})}$$

$$\text{Confidence} = \frac{\text{Support}(\text{Milk, Bread})}{\text{Support}(\text{Milk})}$$

5/5/22

Recommendation System / Algo

Search vs recommendation system
→ long tail phenomenon



Types of recommendation systems
→ collaborative based rec.
→ content based rec. → based on past historical info
→ Knowledge based

find similarity b/w 2

Eg: based on movie rating given by user features of that movie are extracted for recommendation

Collaborative → memory based
→ model based

construct utility function / utility matrix
$R : X \times S$
$X \rightarrow$ set of customers
$S \rightarrow$ set of items

	Item 1	Item 2	Item 3	Item 4
User 1	5		3	
User 2		6	4	2
User 3		7		
User 4				5
User 5	3			

→ Data gathering
↳ collected in 2 ways
↳ explicit data collection
↳ implicit data collection
→ Extrapolating the unknown values
→ Evaluate the values

2 major challenges in recommendation system
1. Cold start problem
2. data sparsity problem

↳ new user will not have historical data so finding similarity in collaborative technique is not possible.

Aggregate recommendation system
Recent videos in youtube, top 10, most popular.
Personalized recommendation system
Similarity
$sim(u_i, u_j) = \cos(u_i, u_j) = \frac{u_i \cdot u_j}{\|u_i\| \|u_j\|}$

Pearson Correlation Coefficient

$$sim(u_i, u_j) = \frac{\sum_k (r_{i,k} - \bar{r}_i)(r_{j,k} - \bar{r}_j)}{\sqrt{\sum_k (r_{i,k} - \bar{r}_i)^2} \sqrt{\sum (r_{j,k} - \bar{r}_j)^2}}$$

$r_{i,k}$ - rating of user i for the product k

$r_{j,k}$ - rating of user j for the product k

$$r_{u,i} = \bar{r}_u + \frac{\sum_{v \in W(u)} sim(u,v)(r_{u,i} - \bar{r}_v)}{\sum_{v \in N(u)} sim(u,v)}$$

So finding unknown rating in matrix
$r_{v,i}$ → observed rating of user v for the product i

$\bar{r}_v$ → user v mean rating

Memory based collaborative → user based
→ Item based

User based
1) Calculate average rating
2) Similarity
3) Neighbourhood size
4) Predict the unknown rating value

	Item1	Item2	Item3	Item4	Item5
Anu	5	3	4	4	?
Priya	3	1	2	3	3
Sam	4	3	4	3	5
Ravi	3	3	1	5	4
Raj	1	5	5	2	1

Step 1: Calculate ~~User-wise~~ Avg rating

$$\bar{r}_{anu} = \frac{5+3+4+4}{4} = 4$$

$$\bar{r}_{priya} = \frac{3+1+2+3+3}{5} = 2.4$$

$$\bar{r}_{sam} = \frac{4+3+4+3+5}{5} = 3.8$$

$$\bar{r}_{ravi} = \frac{3+3+1+5+4}{5} = 3.2$$

$$\bar{r}_{Raj} = \frac{1+5+5+2+1}{5} = 2.8$$

Step 2: Similarity

$$Sim(anu, priya) = \frac{(5\times3)+(3\times1)+(4\times2)+(4\times3)}{\sqrt{(5)^2+(3)^2+(4)^2+(4^2)}\cdot\sqrt{(3)^2+(1)^2+(2^2)+(3)^2}}$$

$$= \frac{15+3+8+12}{\sqrt{66}\cdot\sqrt{23}}$$

$$= \frac{38}{\sqrt{66}\cdot\sqrt{23}} = 0.975$$

$$Sim(anu, sam) = \frac{(5\times4)+(3\times3)+(4\times4)+(4\times3)}{\sqrt{5^2+3^2+4^2+4^2}\cdot\sqrt{4^2+3^2+4^2+3^2}}$$

$$= \frac{57}{\sqrt{66}\cdot\sqrt{50}} = \frac{57}{57.44} = 0.992$$

$$Sim(anu, ravi) = \frac{(5\times3)+(3\times3)+(4\times1)+(4\times5)}{\sqrt{66}\cdot\sqrt{3^2+3^2+1^2+5^2}}$$

$$= \frac{5.908}{\sqrt{44}} = 0.890$$

$$sim(anu, raj) = \frac{(5\times1)+(3\times5)+(4\times5)+(4\times2)}{\sqrt{66}\cdot\sqrt{1^2+5^2+5^2+2^2}}$$

$$= \frac{48}{\sqrt{66}\cdot\sqrt{55}} = 0.796$$

Step 3: Neighbourhood size

Assume neighbourhood size = 2

⇒ Top 2 are

i) Anu + Sam

ii) Anu + Priya

Step 4: Predict unknown rating value

$$\bar{r}_{anu,items} = 4 + \frac{0.992(5-3.8)+0.975(3-2.4)}{0.992+0.975}$$

(ii) $$= \bar{r}_{anu} + \frac{sim(anu,sam)(r_{sam,items} - \bar{r}_{sam}) + sim(anu,priya)(r_{priya,items} - \bar{r}_{priya})}{sim(anu,sam)+sim(anu,priya)}$$

$$= 4 + \frac{1.188+0.582}{1.96}$$

$$\bar{r}_{anu,items} = 4.903$$

Item based (Consider same data)

① Avg rating

$\bar{r}_{i1} = \frac{5+3+4+3+1}{5} = 3.2$

$\bar{r}_{i2} = \frac{3+1+3+3+5}{5} = 3$

$\bar{r}_{i3} = \frac{4+2+4+1+5}{5} = 3.2$

$\bar{r}_{i4} = \frac{4+3+3+5+2}{5} = 3.4$

$\bar{r}_{i5} = \frac{3+5+4+1}{4} = 3.25$

② Similarity

$sim(i1, i5) = \frac{(3\times3)+(4\times5)+(3\times4)+(1\times1)}{\sqrt{3^2+5^2+4^2+1^2}\cdot\sqrt{3^2+4^2+3^2+1^2}}$

$= \frac{42}{\sqrt{51}\cdot\sqrt{35}} = 0.994$

$sim(i2, i5) = \frac{(3\times1)+(5\times3)+(4\times3)+(1\times5)}{\sqrt{51}\ \sqrt{1^2+3^2+3^2+5^2}}$

$= \frac{35}{\sqrt{51}\sqrt{44}} = 0.738$

$sim(i3, i5) = \frac{(3\times2)+(5\times4)+(4\times1)+(1\times5)}{\sqrt{51}\cdot\sqrt{2^2+4^2+1^2+5^2}}$

$= \frac{35}{\sqrt{51}\cdot\sqrt{46}} = 0.722$

$sim(i4, i5) = \frac{(3\times3)+(5\times3)+(4\times5)+(1\times2)}{\sqrt{51}\cdot\sqrt{3^2+3^2+5^2+2^2}}$

$= \frac{46}{\sqrt{51}\cdot\sqrt{47}} = 0.939$

③ Neighbourhood zone = 2

i) Item1, Item5

ii) Item4, Item5

④ Find unknown rating

$\bar{r}_{items,\ anu} = 3.25 + \frac{0.994(5-3.2)+0.939(4-3.4)}{0.994+0.939}$

$= 3.25 + \frac{1.7892+0.558}{1.92}$

$= 4.46$

Content based filtering

I/P : User i's profile information, item description for each items $j \in \{1,2,3,\ldots n\}$ keywords, & number of recommendation

Return n number of recommended items

$u = \{u_1, u_2, u_3, \ldots u_n\}$ user profile vector

$(I_j)_{j=1}^{n} = \{I_{j,1}, I_{j2}, I_{j3}, \ldots I_{jk}\}$

↳ Item description vector

$S_{i,j} = sim(U_i, I_j)$

Return top r with maximum similarity

Qn: Data: Movie Item Feature Matrix

	Fantasy	Action	Cartoon	Drama	Comedy
M1	0	0	1	0	1
M2	0	0	0	1	0
M3	1	0	0	0	1
M4	1	1	0	0	1
M5	1	1	0	1	0
M6	0	1	1	0	1
M7	0	0	0	1	0

Qn: User Item Matrix

User1: M1 = 7 M3 = 10 M5 = M7 =
M2 = 4 M4 = M6 =

Similarity b/w User and Item

$$= \begin{bmatrix} 7 \\ 4 \\ 10 \end{bmatrix} \times \begin{bmatrix} 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 0 & 0 & 7 & 0 & 7 \\ 0 & 0 & 0 & 4 & 0 \\ 10 & 0 & 0 & 0 & 10 \end{bmatrix} = [10 \;\; 0 \;\; 7 \;\; 4 \;\; 17]$$ Sum = 38

$$= [0.26 \;\; 0 \;\; 0.18 \;\; 0.10 \;\; 0.44]$$

$$\downarrow \quad \frac{10}{38} \quad \frac{0}{38} \quad \frac{7}{38} \quad \frac{4}{38} \quad \frac{17}{38}$$

Preference
i) Comedy
ii) Fantasy
iii) Cartoon
iv) Drama

Future recommendation

$$\begin{bmatrix} 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix} \times [0.26 \;\; 0 \;\; 0.18 \;\; 0.10 \;\; 0.44]$$

	Fantasy	Action	Cartoon	Drama	Comedy	
=	0.26	0	0	0	0.44	M4
	0.26	0	0	0.10	0	M5
	0	0	0.18	0	0.44	M6
	0	0	0	0.10	0	M7

$$= \begin{bmatrix} 0.7 \\ 0.36 \\ 0.62 \\ 0.10 \end{bmatrix} \begin{matrix} M4 \\ M5 \\ M6 \\ M7 \end{matrix}$$

Top two recommended: M4, M6 by content based filtering

{Implement content & collab based filtering in R/Python using inbuilt movie dataset (IMDB)}.