USER-USER Collaborative filtering

Collaborative filtering:

A has not watched all the movies.
eg. in this case.
A ≈ B ⇒ similar Ratings
A + C :. ⇒ most dissimilar cases

	HP1	HP2	HP3	TW	SW1	SW2	SW3
A	4			5	1		
B	5	5	4				
C				2	4	5	
D		3					3

(users ↑, movies →)

Jaccard Similarity

$$Sim(A,B) = \frac{R_A \cap R_B}{R_A \cup R_B} = \frac{1}{5}$$

$$Sim(A,C) = \frac{2}{4}$$

∴ Sim (A, B) < Sim (A, C)

just cos A, B have watched lesser movies than A, C have in common & not if they both liked it.
problem

Cosine Similarity

> Insert unknown values : 0.

$$Sim(A,B) = \frac{\sum A \cdot B}{\sqrt{A^2}\ \sqrt{B^2}} = \frac{4*5}{\sqrt{4^2+5^2+1^2}\cdot\sqrt{5^2+5^2+4^2}} = 0.38.$$

sim (A, C) = 0.32 ← cosine ∠ btw both.

∴ Sim (A, B) > sim (A, C) but not by much.

problem: treats missing Ratings as least

Centered Cosine / Pearson Correlation / Rating Normalization

Normalization:

A : (4+5+1) / 3 = 10/3
B : (5+5+4) / 3 = 14/3
C : (2+4+5) / 3 = 11/3
D : (3+3) / 2 = 6/2

(low Rating becomes -ve, High → +ve)

Subtract this value from each cell.
then Subtract from Avg

i.e. $4 - \frac{10}{3} = \frac{2}{3}$, $5 - \frac{10}{3} =$ - - -

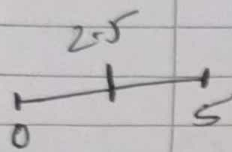
Can use Euclidean dist as well

(2)

	HP1	HP2	HP3	TW	SW1	SW2	SW3	
A	2/3			5/3	-7/3			→ Σ = 0
B	1/3	1/3	-2/3					Σ = 0
C				-5/3	1/3	4/3		
D		0					0	

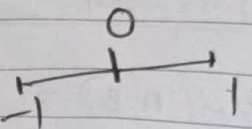
This makes Avg. Rating ∀ users = 0.

∴



Before
1) Avg = 2.5
2) Range = 0-5
3) 0-2.5 : Below Avg
4) Missing = 0

After
1) Avg = 0
2) -1 to +1
3) Below 0 : Below Avg.
4) Missing = Avg.



Now Compute Cosines.

$$Sim(A,B) = \frac{\frac{2}{3} * \frac{1}{3}}{\sqrt{\left(\frac{2}{3}\right)^2 + \left(\frac{5}{3}\right)^2 + \left(\frac{7}{3}\right)^2} \cdot \sqrt{\left(\frac{1}{3}\right)^2 + \left(\frac{1}{3}\right)^2 + \left(\frac{-2}{3}\right)^2}}$$

$$= \frac{2/9}{\sqrt{140/729}} \simeq 0.092$$

$Sim(A,C) = -0.56$.

$\therefore Sim(A,B) > Sim(A,C)$.

ITEM ITEM Collaborative filtering
(More useful, outperforms user-user CF)
∵ items are simpler than users
- items : fixed set of genres.
- users : kinda like fuzzy logic Nature.

Because you watched Crime patrol, you might also like Sardhaan India.

precision @ k , Recall @ k

$\frac{\#k \text{ Recommended}}{\#\text{Actual Relevant}}$

estimate Rating done by users for movie1

DATE

Movies \ Users	1	2	3	4	5	6	7	8	9	10	11	12
1	1		3		?	5			5		4	
2			5	4			4			2	1	3
3	2	4		1	2				4	3	5	
4		2	4		5		3	4			2	
5			4	3	4	2					2	5
6	1		3		3			2			4	

	1	2	3	4	5	6	7	8	9	10	11	12
1	-2.6	0	-0.6	0	?	1.4	0	0	1.4	0	0.4	0
2	0	0	1.84	0.84	0	0.84	0	0	0	-1.16	-2.16	-0.16
3	-1	1	0	-2	-1	0	0	0	1	0	2	0
4	0	-1.4	0.6	0	1.6	0	0	0.6	0	0	-1.4	0
5	0	0	0.7	-0.3	0.7	-1.3	0	0	0	0	-1.3	1.7
6	-1.6	0	0.4	0	0.4	0	0	-0.6	0	0	1.4	0

sim(M1, other Movies Rated by user 5) =? ∴ 2 movies with highest similarity

(1,5) =?

sim(m1, M3) = $\frac{2.6*1 + 1.4*1 + 0.4*3}{\sqrt{-}\quad\sqrt{-}}$ = 0.41 ✓

(1,3)

sim(m1, m4) = -0.10

sim(m1, m5) = -0.31

sim(m1, m6) = ——— = 0.59 ✓

(1,6)

take weighted Avgs using these

sim(m1, m1) = 1

sim(m1, m2) = -0.18

(3,5) (6,5)

∴ Ratings for [?] ⇒ $\frac{0.41*2 + 0.59*3}{0.41+0.59}$ = 2.6

∴ [?] = 2.6

classmate

PAGE

Theory.

④ DATE

User preferences → Recommender systems → Recommendations (predicting future behaviors).

Explicit feedback

implicit feedback

Collaborative:

Similar users like Similar things.
User x item Matrix

Content Based:

Considers item/user features

item x item
movie → genre, year of release, cast, director, prod: house.

user x user
age, gender, spoken language

---

1. **Content Based Filtering**

User vs Feature Matrix * Feature vs Movie
(Comedy, Action) (Comedy, Action)

↓

User vs Movie.

features come directly.

2. **Collaborative Based Filtering**.

User vs Movie with ?

↓ patterns (latent features).

User vs feature * Feature vs Movie

↓

User vs Movie | W/o ? / to fill ?

you keep guessing values for both Matrixs until you get closest Matrix Multipl° product same as from where u started.

Applications of Recommend: systems: News/Songs/...

Eg. Synthetic Control. what'll be effect of "Gun Control" policies if Implemented? you check for countries → to you that already Implemented em

experience".

- You, as a reader or student of a particular course, learn something and then become equipped or capable of carrying out various tasks based on what you learnt. You might have heard about Bloom's Taxonomy. Bloom's Taxonomy is a classification of the different objectives and skills that learners could achieve out of a particular learning or a course. Fig. 1.1.1 outlines Bloom's Taxonomy.

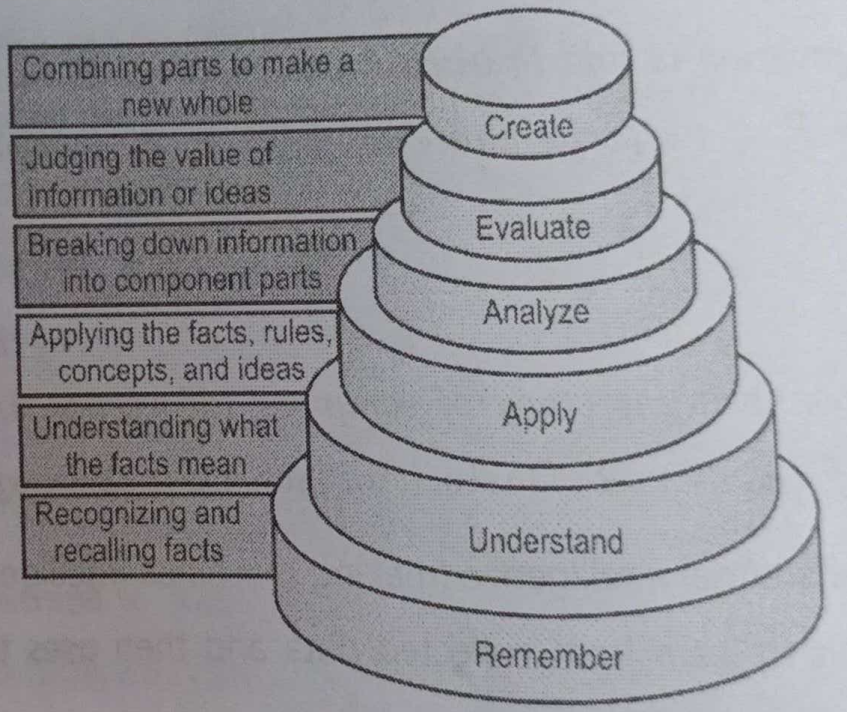


Fig. 1.1.1 : Bloom's Taxonomy

Table 1.1.2

Comparison Attributes	Structured Data	Semi-structured Data	Unstructured Data
Volume of Data	Low	Medium	High
Processing Complexity	Low	Medium	High
Data generated by	Humans and Machines	Machines	Humans
Data usually stored in	Relational Databases	Textual files	Binary files
Patterns and Schema	Fixed	Flexible	Random
Specialised Tools	Not required	Not required	Required

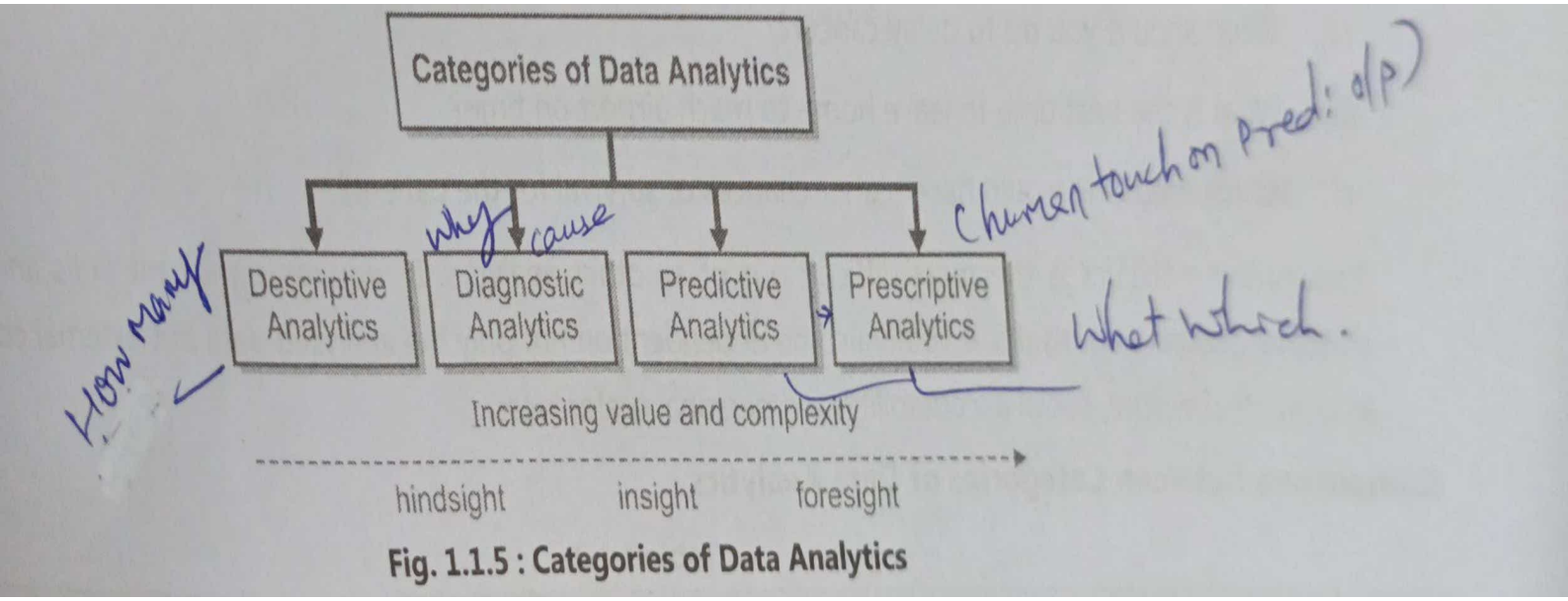


Fig. 1.1.5 : Categories of Data Analytics

Table 1.1.3

Comparison Attribute	Descriptive Analytics	Diagnostic Analytics	Predictive Analytics	Prescriptive Analytics
Complexity	Least	Medium	High	Highest
Time requirement to produce results	Low	Medium	High	Very High
Value of results	Short Term	Medium Term	Long Term	Very long term
Data enrichment level	Data	Information	Knowledge	Wisdom
Analytics Frequency	Most Common	Frequent	Not often	Rare

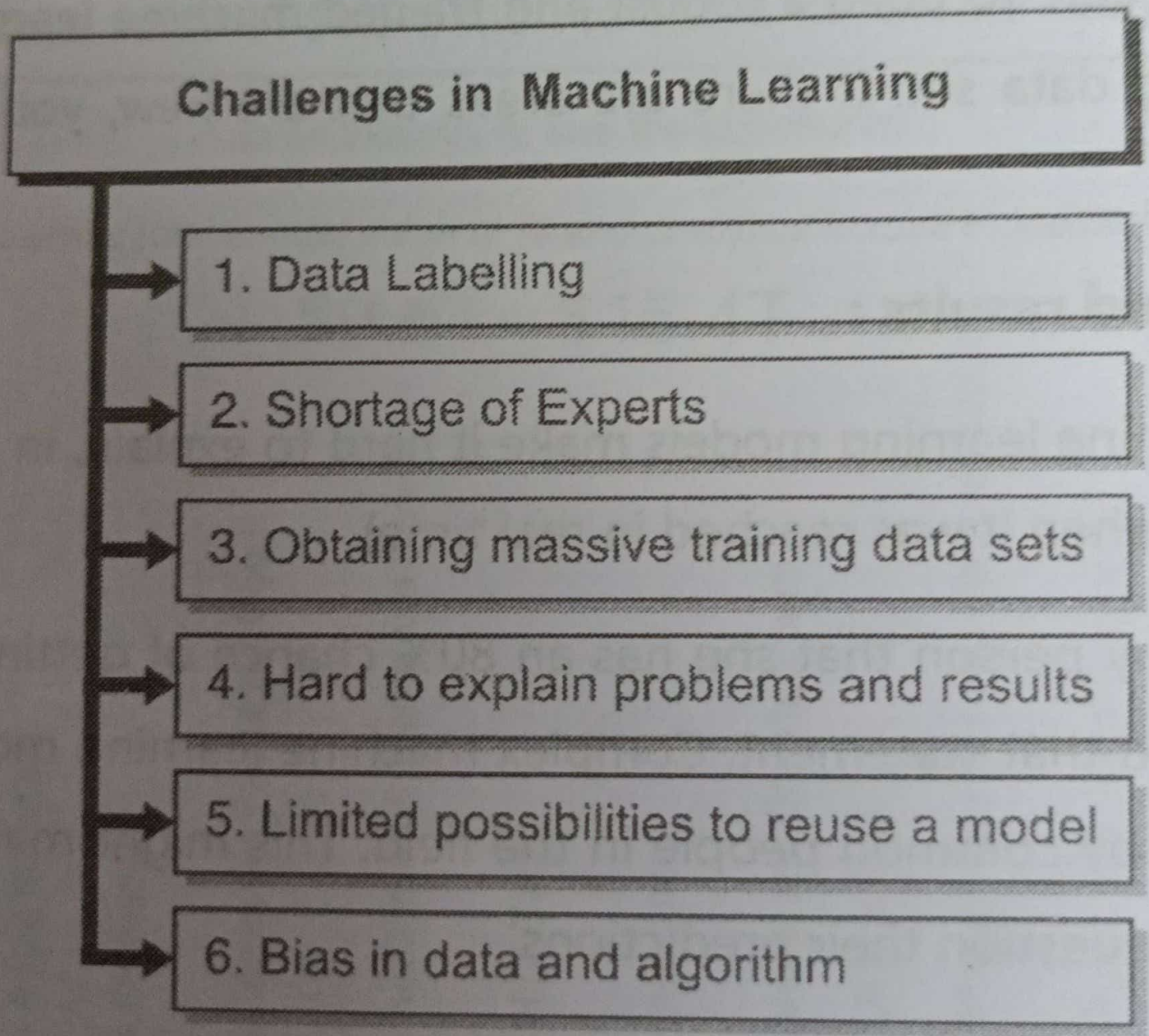
Table 1.2.1

Comparison Attribute	Supervised Learning	Unsupervised Learning
Training Dataset Contains	Both input and output	Only input
Used for	Classification and Prediction	Finding patterns and understanding data
Training Data	Is Labelled	Not labelled
Number of targets	Known beforehand	Not known
Feedback from user	Provided	Not provided
Complexity	High	Low

Table 1.2.2 subset of AI

Comparison Attribute	Machine Learning	Artificial Intelligence
Focus	Learn from data	Solve complex problems
Complexity	Low	High
Scope	Narrow	Broad
Human interaction	Minimal	High

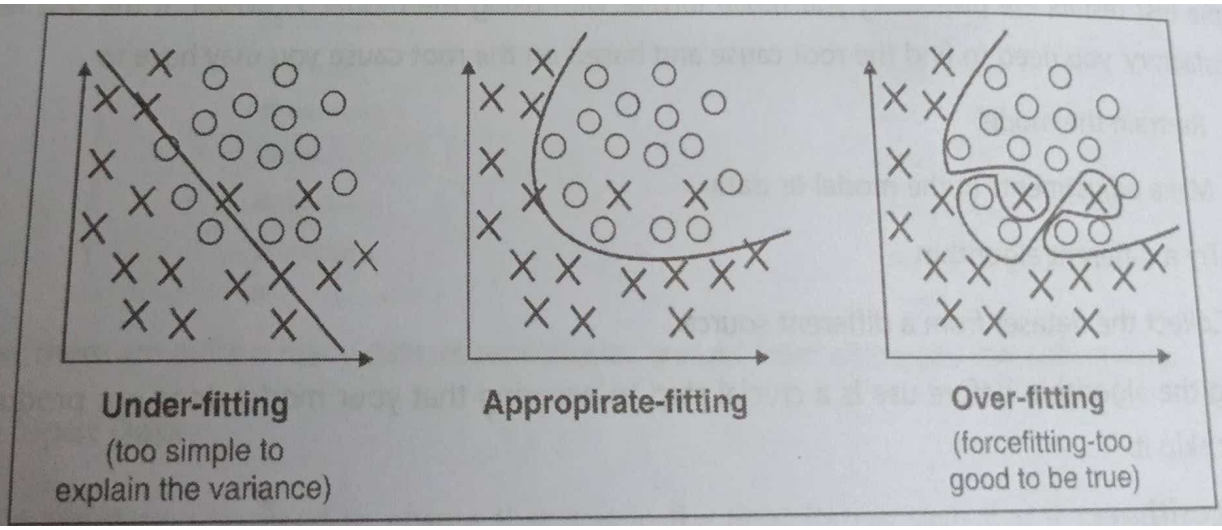
Comparison Attribute	Machine Learning	Data Mining
Building a trained model	Required	Not required
Human effort required	Only for building model	For extracting information from data
Use of specific algorithms	Frequent	Rare
Accuracy	High	Low
Tasks carried out by	Machines	Humans
Self-learning	Yes	No

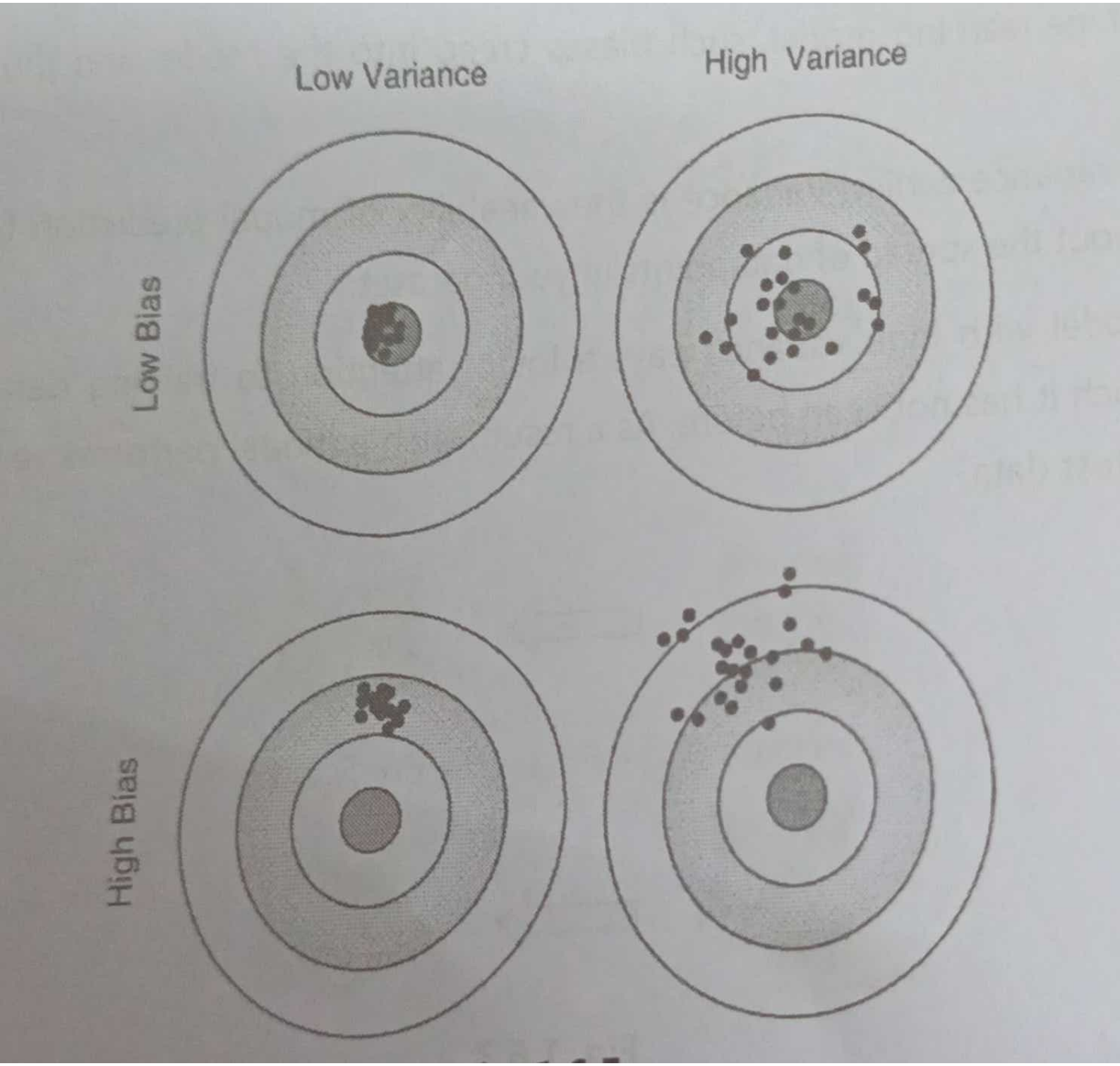


everything u read != what u learn

Fig. 1.3.1 : Challenges in Machine Learning

not everything is correct


Under-fitting
(too simple to
explain the variance)
Appropirate-fitting
Over-fitting
(forcefitting-too
good to be true)


Low Variance
High Variance
Low Bias
High Bias

Algorithm Name	Bias	Variance
Linear Regression	High	Low
Decision Tree	Low	High
Bagging	Low	High
Random Forest	Low	High

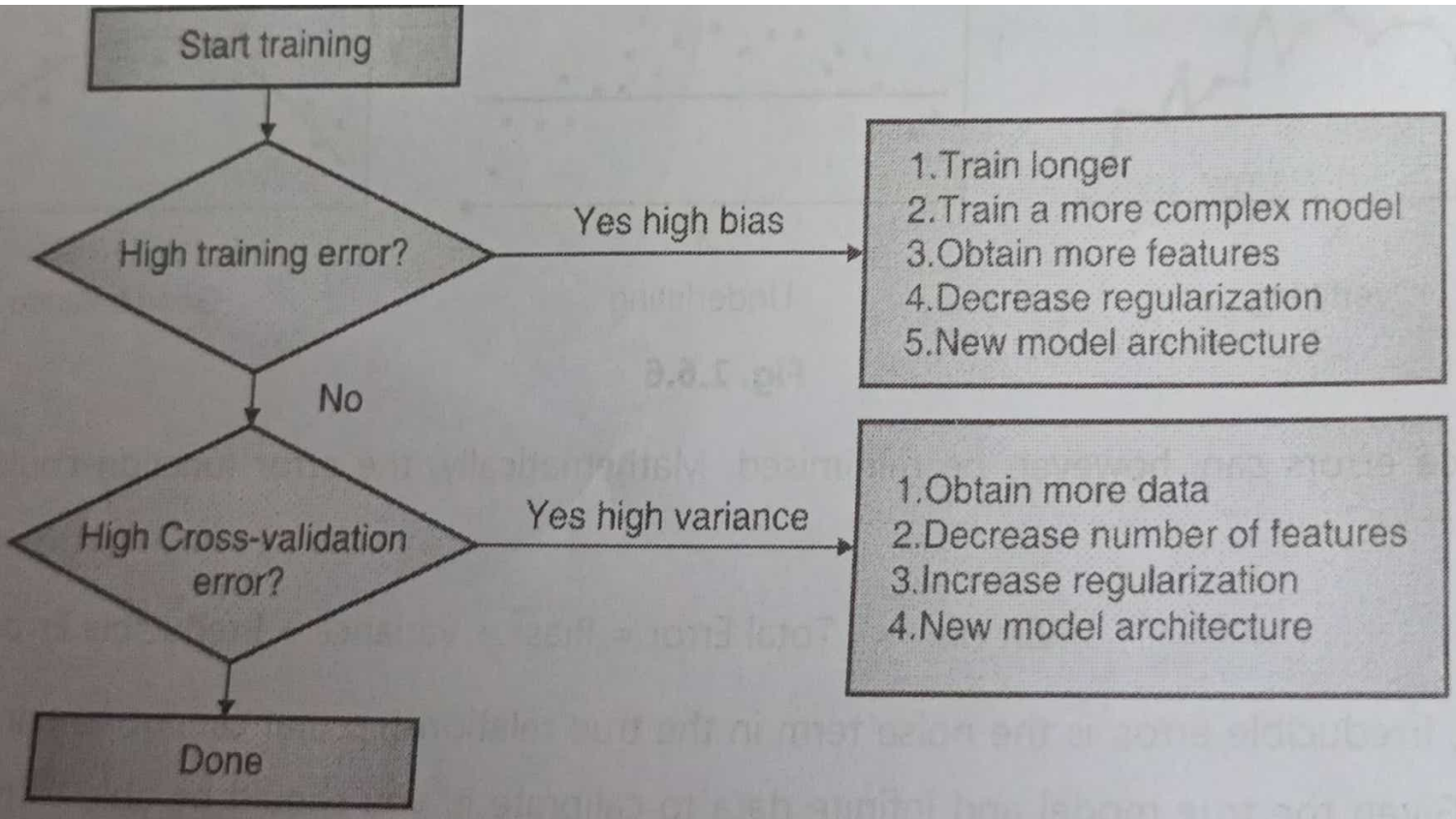

Start training
High training error?
Yes high bias
1.Train longer
2.Train a more complex model
3.Obtain more features
4.Decrease regularization
5.New model architecture
No
High Cross-validation error?
Yes high variance
1.Obtain more data
2.Decrease number of features
3.Increase regularization
4.New model architecture
Done

Fig. 1.6.8

1.6.5 Characteristics (Detection) of a High Bias Model

Characteristics of a high bias model are as following.

1. Failure to capture proper data trends
2. Low training accuracy
3. Potential towards underfitting
4. More generalised or overly simplified
5. High error rate

1.6.6 Characteristics (Detection) of a High Variance Model

Characteristics of a high variance model are as following.

1. Noise in the data set
2. Low testing accuracy
3. Potential towards overfitting
4. Complex models
5. Trying to put all data points as close as possible

Simple linear regression	***multiple linear regression***
1 independent variable (X) 1 dependent variable (Y)	n independent variable (X_1, X_2,...) 1 dependent variable (Y)
Y=mX+c	$y = b_0 + b_1*x_1 + b_2*x_2 + \dots + b_n*x_n$
only one relationship (X,Y)	more than 1 relationship. between (X1,Y) ; (X2,Y)... between dependent & independent & X1,X2 ; X1,X3.... between 2 independent.
➢ Y increases by factor of m if X increases by 1 ➢ Y = c or intercept if X=0	➢ Y increases by factor of b1 if X1 increases by 1 keeping X2 constant ➢ Y increases by factor of b2 if X2 increases by 1 keeping X1 constant ➢ Y = a or b0 if all independent variables are 0

- Adding more independent variables to a multiple regression procedure does not mean the regression will be "better" or offer better predictions; in fact it can make things worse. This is called OVERFITTING.
- The addition of more independent variables creates more relationships among them. So not only are the independent variables potentially related to the dependent variable, they are also potentially *related to each other*. When this happens, it is called MULTICOLLINEARITY.
- The ideal is for all of the independent variables to be correlated with the dependent variable but NOT with each other.

Multiple Regression Model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_p x_p + \epsilon$$

linear parameters — error

Multiple Regression Equation

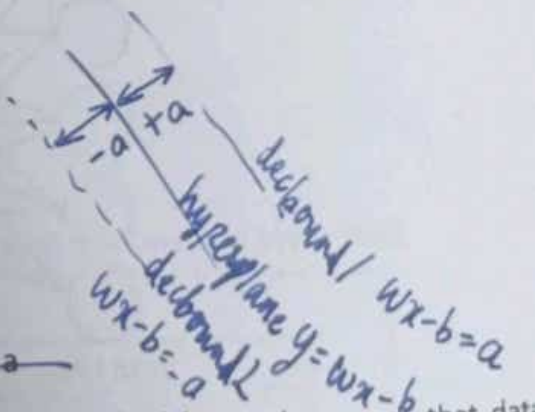
$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_p x_p$$

error term assumed to be zero

Comparison Attribute	Logistic Regression	SVM
Good for	Linear classification	Both linear and non-linear classificatio
Decision boundary	Multiple	One (best one)
Approach	Statistical	Geometrical
Errors	Comparatively higher	Comparatively Lower

- As in case of SVM, consider the two decision boundaries and a hyperplane. Your objective is to consider the po that are within the decision boundary line. The best fit line (or regression line) is the hyperplane that ha maximum number of points. Assume that the decision boundaries are at any distance, say 'a', from the hyperpl So, these are the lines that you draw at distance '+a' and '-a' from the hyperplane. Based on SVM, the equatio the hyperplane is as following.

$$Y = wx - b,$$



- The equations of decision boundaries are as following.

$$wx - b = a$$

$$wx - b = -a$$

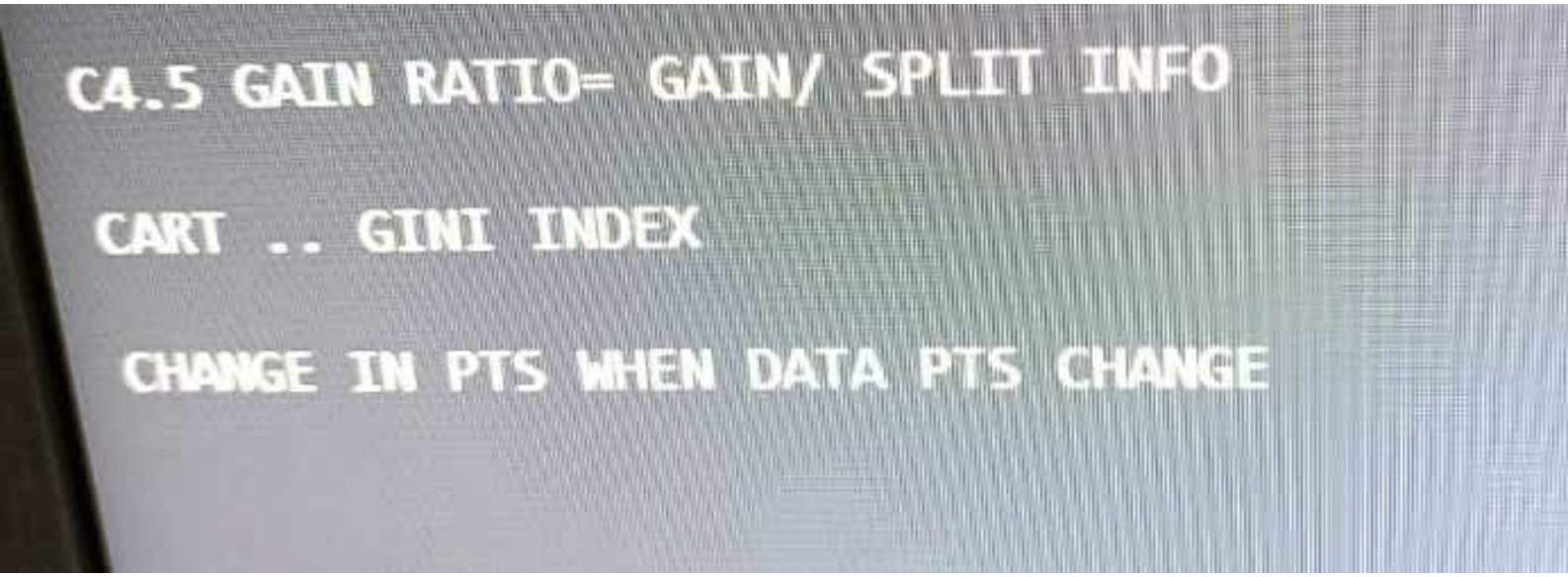
- Thus, any hyperplane that satisfies SVR should satisfy $-a \leq wx - b \leq a$
- Your goal is to decide a decision boundary at 'a' distance from the original hyperplane such that data closest to the hyperplane or the support vectors are within that boundary line.

Table 4.6.1

Comparison Attribute	One vs One (OvO)	One vs Rest (OvR)
Speed	Slower than OvR	Faster than OvO
Computation Complexity	High	Low
Suitable for	Algorithms that don't scale	Algorithms that scale
No. of binary datasets or models for C classes	$\frac{C \times (C-1)}{2}$	C
Interpretability	Low	High
Used	Less commonly	More Commonly

Table 3.1.1

Comparison Attribute	Bagging	Boosting
Weak models work	Parallelly	Sequentially
All models have	Equal weight	Adjusted weights
Reduces	Variance	Bias
Overfitting	Addressed	Not addressed
Merge predictions of	Same type	Different types


C4.5 GAIN RATIO= GAIN/ SPLIT INFO
CART .. GINI INDEX
CHANGE IN PTS WHEN DATA PTS CHANGE

LINK

- The distance matrix is, AVG[dist(P3,P6),P1]

dist((P3,P6),P1) = ½ (dist(P3,P1)+dist(P6,P

= ½ (0.22+0.23)

= ½ (0.45)

= 0.23

www.anuradhabhatia.com

chical Agglomerative Clusterir
Average Link]

6 yr ago Data Warehouse and Mining ...mo

nuradha Bhatia 10.5K

Share Remix

ents 51

Add a comment...

Single Linkage	This is the distance betwee members of the two clusters.
Complete Linkage	This is the distance between that are farthest apart.
Average Linkage	This method involves loo distances between all pairs all of these distances. This i UPGMA - Unweighted Pair Averaging.

chical Agglomerative Clusterin
Average Link]

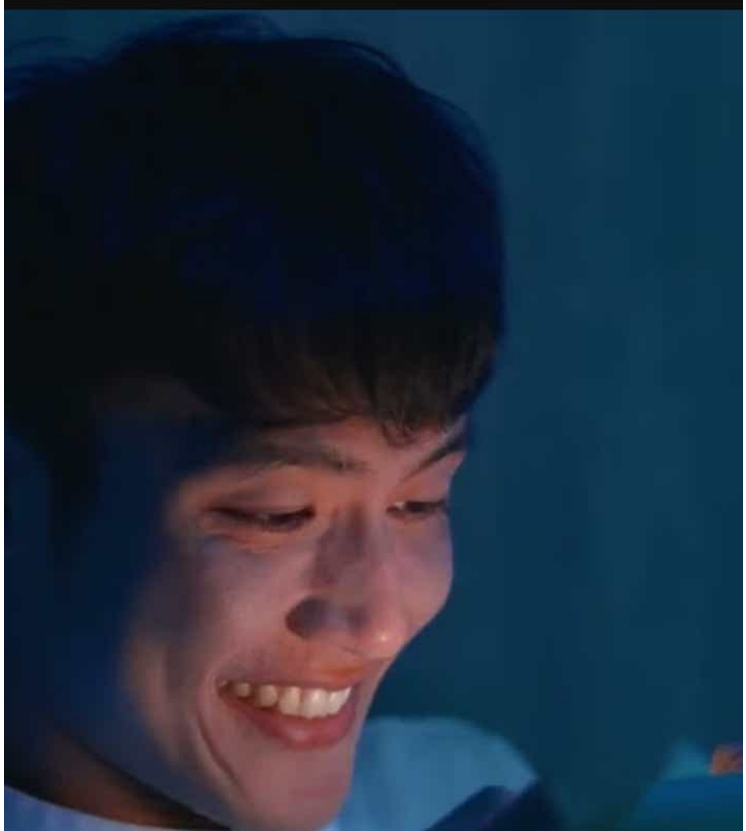
6 yr ago Data Warehouse and Mining ...mo

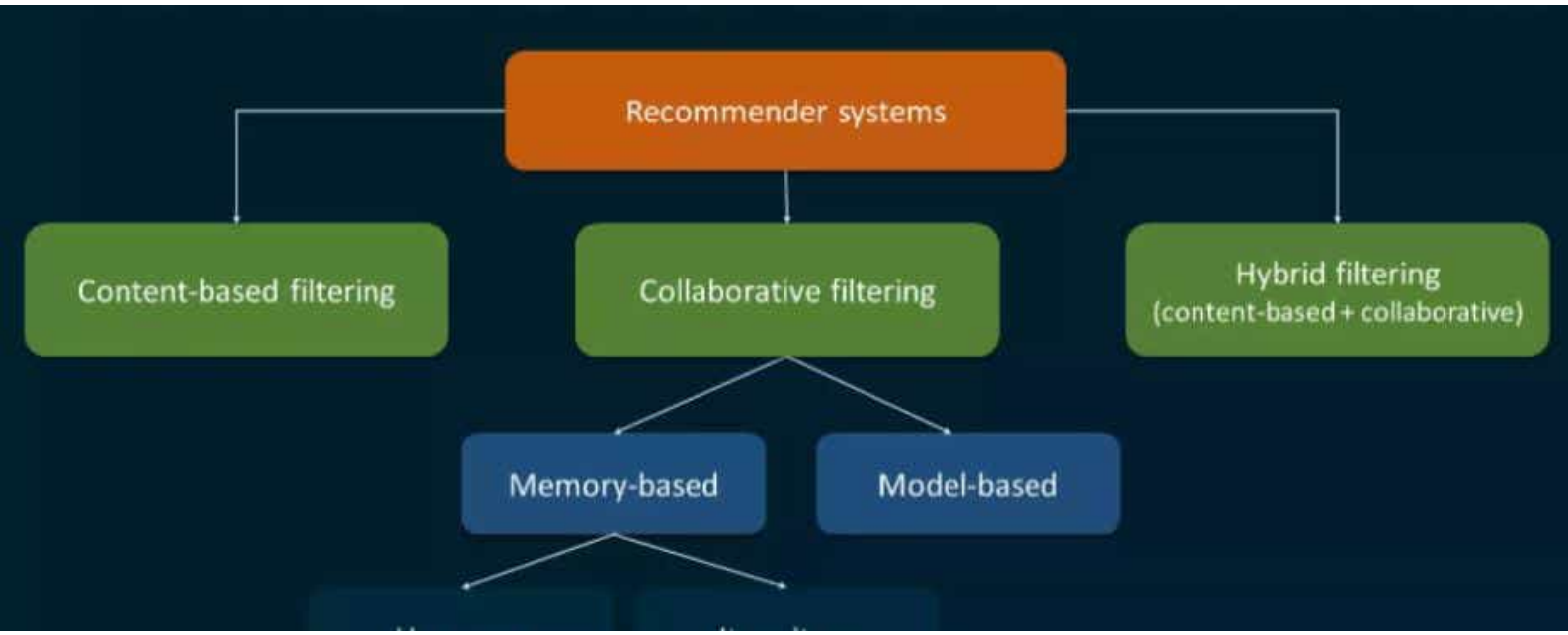
nuradha Bhatia 10.5K Su

K Share Remix

ents 51

Add a comment...




Recommender systems
Content-based filtering
Collaborative filtering
Hybrid filtering
(content-based + collaborative)
Memory-based
Model-based