Lecture - 1 : Supervised and Unsupervised Machine Learning Algorithms.

→ Classification and regression : supervised learning

→ Clustering and association : unsupervised learning.

→ between sup and unsup is called semi supervised learning..

Supervised Machine Learning :

→ Practical learning uses supervised learning. (for output $f(x)$)

→ input variable (x) and output variable (Y), use an algorithm to learn the mapping function from input to output.

$Y = f(x)$ ⟹ mapping function

new input data (x) can predict output (Y) for data.

The process of an algorithm learning from the training dataset can be thought as a teacher supervising the learning process is called Supervised Learning.

* we know the correct ans, then algo makes prediction on training data.

Classification : output variable is a category.

Ex: red or blue ; 'disease' or 'no disease'

Regression : output variable is a real value.

Ex: dollars , weight

Example; of supervised ML:

recommendation, time series prediction.

① Linear regression (regression problem)	② Random Forest (regression ; classification problem)	③ SVM (classification problem)

Unsupervised Machine Learning:

UML is where only have input data (x) and no corresponding output variables.

Goal: underlying structure in the data, so it can learn more about the data.

There is no correct answer, no teacher. Algorithms are left to their own devises to discover and present ~~to~~ the interesting structure in the data. is called unsupervised learning.

UML further grouped into clustering and association problems.

Clustering: Discover the inherent groupings in the data,

Ex: grouping customers by purchasing behaviour.

সমীর. অর্জিত সমীর, বিক্রিত সমীর ☺

Association :

Discover rules that describe Large portions of data.
(* data এর মধ্যে rules বের করে)

Ex: যারা x buy করে, তারা y ও buy করে or করতে চায়।

Example : ① K-means for clustering.

② Apriori " association.

Semi- supervised ML

A large amount of input data (x) and only some ~~to~~ of the data is labeled (Y) are called semi-supervised learning problems.

Example: photo archive, some of images are labeled and majority are ~~unabled~~ Unlabeled.

→ Unlabeled data is cheap and easy to collect and store.
→ many world fall into this area.
→ use SSML to discover and learn the structure in the input variables.

Supervised : All data is labeled and algorithms learn to predict the output from the input data.

Unsupervised : All data is ~~unlable~~ unlabeled and algorithms learn to inherent structure from the input.

Semi-supervised : some data is labeled but most of data is unlabeled, mixture of supervised and unsupervised techniques are used.

Lecture : 2

KDD = Knowledge data discovery.

Original Data } Target data } Preprocessed Data } Transformed data } Data mining algo → Pattern

Fig: KDD processes   Knowledge { Pattern

Lecture : 2

Why we need to prepare data?

Data can be incomplete, inconsistent or noisy.

① Data entry, transmission, data collection problem solve করা যায়

② duplicate data remove করা দরকার হয়

③ incomplete missing data problem solve এর জন্য

④ avoid contradictions in data.

When data can't be trusted:

* difficult to make knowledge from untrusted site.
* Better chance to discover new knowledge or pattern when data is clean.

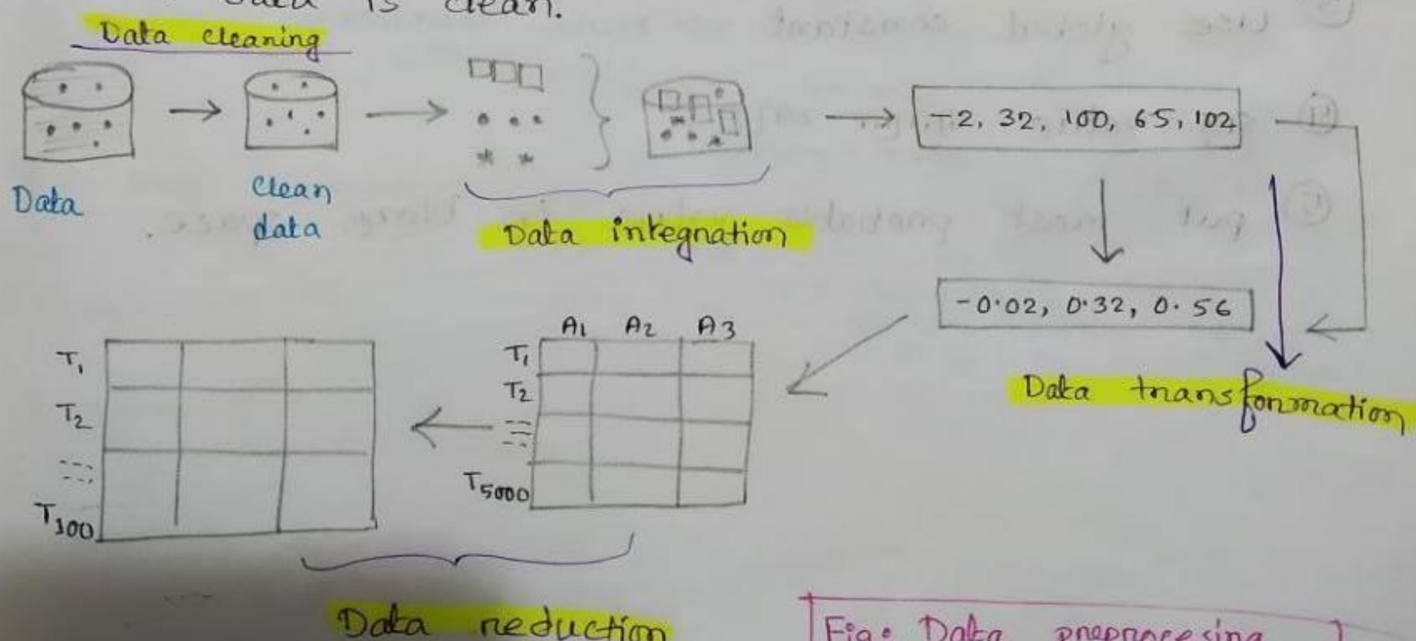


Fig: Data preprocessing

Data Cleaning: real world এ data noisy, incomplete inconsistent থাকে,

- some attribute has no value
- random error
- irrelevent records

এগুলোকে দূর data cleaning করতে হবে।

Data cleaning এর ধাপ:

① fill in missing values

② pull out noisy data

③ correct inconsistencies

④ remove irrelevent data

Solving missing data problem:

① ignore the record with missing value

② fill the missing value manually.

③ Use global constant ⇒ NULL, Unknown

④ গড় value বসানো যাবে

⑤ put most probable value in blank space.

Smoothing Noisy Data | 3 methods

Binning method.

Avg   Boundaries

Clustering   ### Regression

Binning method technique

1. Data sort করো
2. প্রতি bin এ 3টা করে value রাখো. when limit is given = 3
3. " " গড় করো
4. ঐ bin এর 3টির value গড় দিয়ে replace করো।
5. in cast of float, keep the nearest int value.
6. Then put the value according to ID in data table.

Data Transformation (Normalization)

Min max method

$$v' = \frac{v - min}{max - min}$$ when range is [0 ... 1]

Z - score normalization

$$v' = \frac{(v - mean)}{\text{Stdv (Standard Deviation)}}$$

stdv

1. $(mean - x_i)^2$

$$\frac{\sum (mean - x_i)^2}{N} = \text{Variance}$$

$$Stdv = \sqrt{Variance}$$

Data transformation: Discretization

Nominal = values from unordered list

Ordinal = " " ordered list

Numeric = real numbers.

methods
1. Binning
2. Histogram Analysis
3. Clustering "
4. Decision-tree "
5. Correlation "

Discretization use হয়: continuous attribute reduce করার জন্য

* some DM algo allow only categorical data/ Attribute, cannot handle continuous attribute value.

** Discretization can also used to generate high level concepts.

age continuous value গুলো high level concept এ convert করে discretization এর সাহায্যে,

Exmple: young, middle age, old → catagory তে ভাগ করে age কে.

(*** Continuous value কে discrete করা হয়)

Age
25
38
51
75
86

low = 25 - 40

mA = 41 - 60

Old = 60 <

→

Age
Low
low
mA
old
old

Data reduction

Often Data become too large, reduce data for improve performance.

① dimensionally reduce করা যায়
② Discretization
③ Regression / clustering reduction.

Chap = 4

① continuous value থাকলে dynamically define new discrete value with range / intervals.

② missing = assign probable value (possible)
= " most common value.

Lecture = 3

Confusion Matrix

TP = Dengue আছে, test positive } actual result.

FN = Dengue আছে. test negavite

TN = Dengue নাই. test negative } actual result

FP. = Dengue নাই, test positive.

Dengue (Original Class) \ test (Predicted Class)	1 Yes a	0 No b
1 Yes a	TP ++	FN +−
0 No b	FP −+	TN −−

	1	0
1	TP	FN
0	FP	TN

Classification

নির্দিষ্ট group, rule আছে,

(Supervise)

Clustering

নির্দিষ্ট group, rules থাকেনা,

(unsupervise)

Confussion matrix

- Accuracy = $\frac{TP + TN}{ALL}$

- Error rate = 1 − accuracy

  or = $\frac{FN + FP}{ALL}$

- Sensitivity = $\frac{TP}{P}$ (TP rate)

- Specificity = $\frac{TN}{N}$

bayes-rate should be = 60 − 40

60 ↓ pstv

40 ↓ negtv

P = Actual true

N = " Negative

Other classifier evalution metrics.

- **Precision:** original যত % positive থাকে, machine তার থেকে বেশি predict করে.

Example: original 70% লোকের heart diseases থাকে, but machine predict করছে 95% লোকের heart disease আছে।

- $\boxed{\text{machine যত গুলো positive বলছে}}$ = ~~EN~~ TP + FP.

$$\text{Precission} = \frac{TP}{TP + FP} \qquad \frac{\text{or}}{M}$$

$TP \xrightarrow{O}$ original 70

$TP + FP \xrightarrow{E}$ 95 , machine says, where HD has or not, but machine says yes.

- **Recall:** machine যত predict করছে, originally তার থেকে বেশি positive থাকবে।

~~70%~~ 50% originally samsung use করে, but machine predict করছে 30% লোকের " " " .

$$\text{recall} = \frac{TP}{TP + FN} \qquad \frac{M}{\text{or}}$$

$TP \xrightarrow{E}$ machine predict actuall user = 30

$TP + FN \xrightarrow{O}$ Original = 50

- Perfect score for both is 1.0, but there is often a a trade-off between precission and recall.

F measure : (F_1 or F-score)

harmonic mean of precission and recall,

$$F = \frac{2 \times \text{precission} \times \text{recall}}{\text{precision} + \text{recall}}$$

Overall correct rate = $\frac{TP + FN}{TP+FP+TN+FN}$

Association Analysis

Now, find, after buying Bread, how many prob to by Milk.

TID	Items
1	Bread, milk
2	Bread, Dipers, Beer, Egg
3	milk, Diaper, Beer, Cola
4	Bread, milk, Diaper, Beer
5	Bread, milk, Diaper, Cola.

■ $\underset{x \to y}{\text{Support}} = \dfrac{(x \cup y)}{N}$

■ $\underset{x \to y}{\text{Confidence}} = \dfrac{(x \cup y)}{\sigma(x)}$

TID	Bread	milk	Diaper	Beer	Egg	Cola
1	1	1	0	0	0	0
2	1	0	1	1	1	0
3	0	1	1	1	0	1
4	1	1	1	1	0	0
5	1	1	1	0	0	1

0/1 binary representation

* $x \cup y$ = x নেয়ার পর y কতগুলোর.
* N = total num of transition
* (x) = কতগুলোতে x আছে

Q: For Bread → Milk.

Support : $\frac{3}{5}$; confidence = $\frac{3}{4}$

* Total num of combination : $R = 3^d - 2^{d+1} + 1$

■ brute force approach for mining S & C for every possible rule.

d = Num of item.

B, D → M

D, M → B etc.

C, D → M

Threshold = 50%

(1)

TID	items
001	1 3 4
002	2 3 5
003	1 2 3 5
004	2 5

কিকি item purchase করে

(2)

	item	sup
	{1}	2
	{2}	3
	{3}	3
✗	{4}	1
	{5}	3

each item name/or id

ঐ item কতবার কেনা হয়

(3)

item	sup
{1}	2
{2}	3
{3}	3
{5}	3

maximum product sell = 3

threshold = 50%

so, we not accept

(3×50%) = 1.5

so, item 4 out.

(4)

itemset
{1, 2}
{1, 3}
{1, 5}
{2, 3}
{2, 5}
{3, 5}

item এর combination

↓

this combination also available in table 1

(5)

	itemset	sup
✗	{1, 2}	1
	{1, 3}	2
✗	{1, 5}	1
	{2, 3}	2
	{2, 5}	3
	{3, 5}	2

set এর product কয়টি কেনা হয়েছে.

(6)

item	sup
1, 3	2
2, 3	2
2, 5	3
3, 5	2

remove. 1.5 > item

(7)

itemset
2 3 5

Here,
1 3 5 is not available in tabe 1.
(2, 3, 5) available

(8)

item	sup
{2 3 5}	2

Here (2, 3, 5) combination occure 2 times in table : 1

Final stage.

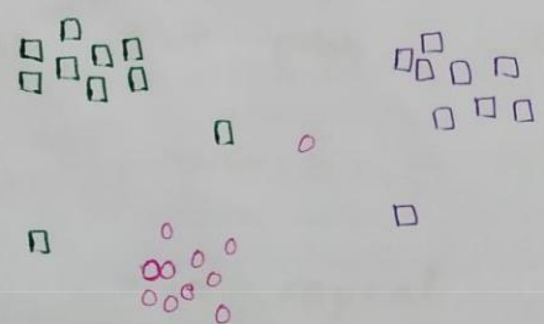
K- mean clustering

Clustering: Clustering is ~~ta~~ a task of grouping set of object in such way that objects in the same group are more similar to each other.

* same group কে $\boxed{\text{cluster}}$ বলে। প্রতি cluster এর objects এর বৈশিষ্ট্য same, একারণেই তাদের একত্র হয়।

* distance measure করে grouping করা হয়।

Example:



3 cluster

Type of clustering:

① **Hierarchical Algorithm:** Hierarchy cluster / ranking cluster build করতে চায়। clustering fall into two types

① bottom- up	② top - down
ছোট → বড়তে আকারে	বড় → ছোট

② **Partitional clustering:** determine cluster at once.

⇒ K-means and derivatives

⇒ Fuzzy c-means clustering.

⇒ QT clustering.

- **Common distance measure:** distance fn দিয়ে similarity measure করা হয়। cluster shape কে influence করে।

① The Euclidean distance ⟹ $d = \sqrt{\Sigma |x_i - y_i|^2}$

② The Manhattan Distance ⇒ $d = \Sigma |x_i - y_i|$

③ Maximum norm ⟹ $d = \max |x_i - y_i|$

④ Mahalanobis distance

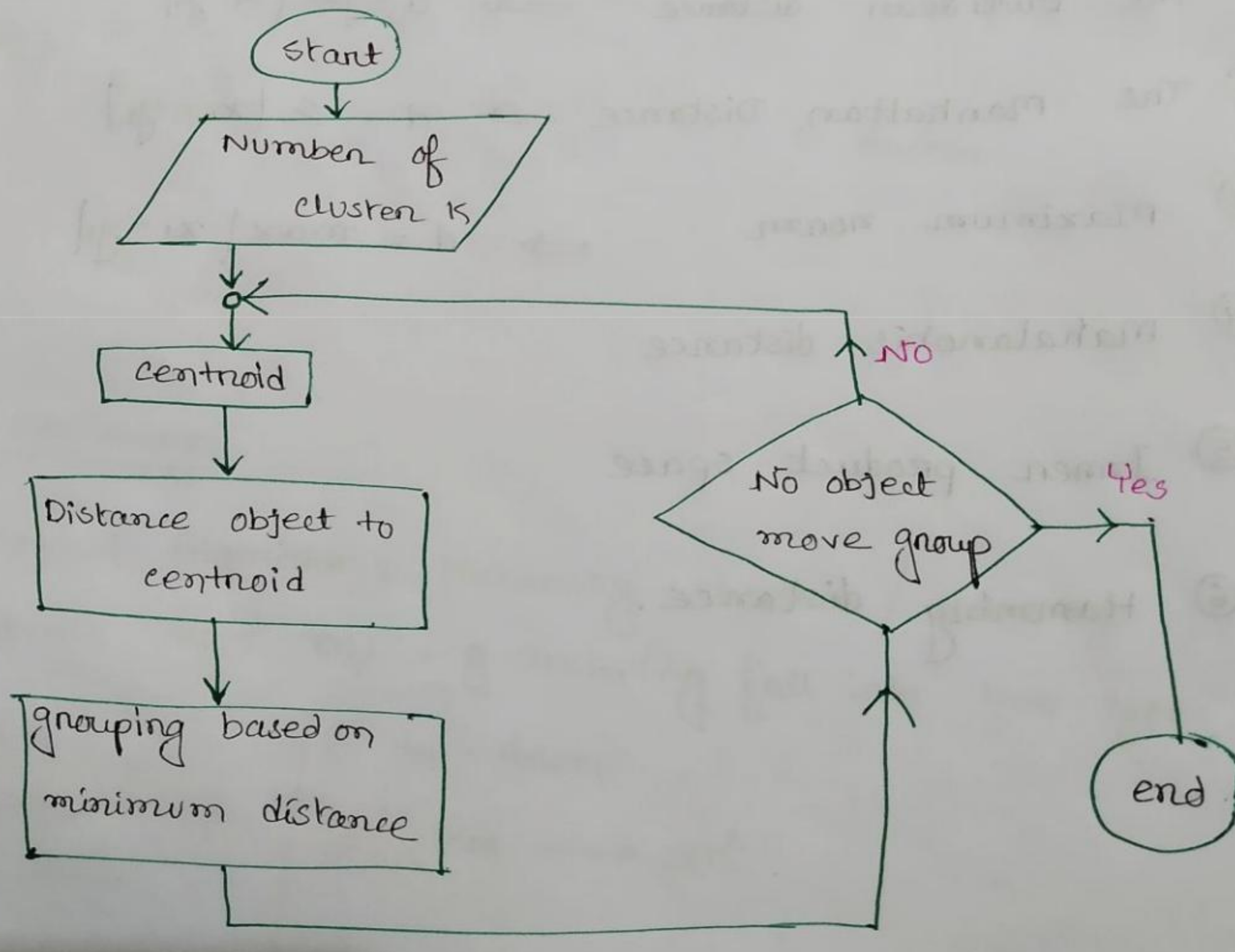
⑤ Inner product space

⑥ Hamming distance.

K-mean clustering

n = object number, K = partition, $n > K$.

যেকোনো 2টি n কে ধরে g_1, g_2 নিও, then, কাছের close object নিয়ে cluster ~~বা~~ বানাও, distance measure করো। next g_1, g_2 হবে group objectদের average. centroid প্রতিবার chng হয়। যখন কোনো g_1, g_2 পরপর 2বার same থাকবে তখন clustering done.



1 5 9 15 20 25 35 50 70

Here K = 2.

Solution : let, centroid 1 = 1, centroid 2 = 20.

distance from 1 : 0 4 8 14 19 24 34 49 69

" " 20 : 19 15 11 5 0 5 15 30 50

Now, (1) group 1 : 1 5 9 ⟶ avg = 5

(20) group 2 : 15 20 25 35 50 70 ⟶ avg = 35.

1 5 9 15 20 25 35 50 70

distance from 5 : 4 0 4 10 15 20 30 45 65

" " 35 : 34 30 26 20 15 10 0 15 35

(5) group : 1 5 9 15 ⟶ avg = 8

(35) group : 20 25 35 50 70. ⟶ avg = 40

d. f 8 : 7 3 1 7 12 17 27 42 62

d. f 40 : 39 35 31 25 20 15 5 10 30.

(8) group : 1 5 9 15 20 ⟶ avg = 10

(40) group : 25 35 50 70. ⟶ avg = 45

d.f 10 = 9 5 1 5 10 15 25 40 60

d.f 45 = 44 40 36 20 25 20 5 5 25

1 5 9 15 20 25 35 50 70

(10) group = 1 5 9 15 20 25 $\longrightarrow$ 12.5

(45) group = 35 50 70. $\longrightarrow$ 51.66

---

d.f 13 : 12 8 4 2 7 12 22 37 57

d.f 52 : 51 47 43 37 32 27 17 2 18

(13) group = 1 5 9 15 20 25

(52) group = 35 50 70.

as repeat 2 time so, group / clustering done.

Ans : [1 5 9 15 20 25]

[35 50 70]

Example of 2D data (x, y).

Euclidean Distance $= \sqrt{(x - x_1)^2 + (y - y_1)^2}$.

given that, $K = 2$.

indv	V1 (x)	V2 (y)
1	1.0	1.0
2	1.5	2.0
3	3.0	4.0
4	5.0	7.0
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5

$m1 = 1.0,\ 1.0$

$m2 = 5.0,\ 7.0$

Now, calculate distance.

For m_1:

$1 = \sqrt{(1.0 - 1.0)^2 + (1.0 - 1.0)^2} = 0$

$2 = \sqrt{(1.0 - 1.5)^2 + (1.0 - 2.0)^2} = 1.12$

$3 = \sqrt{(1.0 - 3.0)^2 + (1.0 - 4.0)^2} = 3.60$

$4 = \sqrt{(1.0 - 5.0)^2 + (1.0 - 7.0)^2} = 7.21$

$5 = \sqrt{(1.0 - 3.5)^2 + (1.0 - 5.0)^2} = 4.72$

$6 = \sqrt{(1.0 - 4.5)^2 + (1.0 - 5.0)^2} = 5.31$

$7 = \sqrt{(1.0 - 3.5)^2 + (1.0 - 4.5)^2} = 4.30$

for m_2:

$$1 = \sqrt{(5.0-1.0)^2 + (7.0-1.0)^2} = 7.21$$

$$2 = \sqrt{(5.0-1.5)^2 + (7.0-2.0)^2} = 6.10$$

$$3 = \sqrt{(5.0-3.0)^2 + (7.0-4.0)^2} = 3.61$$

$$4 = \sqrt{(5.0-5.0)^2 + (7.0-7.0)^2} = 0$$

$$5 = \sqrt{(5.0-3.5)^2 + (7.0-5.0)^2} = 2.5$$

$$6 = \sqrt{(5.0-4.5)^2 + (7.0-5.0)^2} = 2.06$$

$$7 = \sqrt{(5.0-3.5)^2 + (7.0-4.5)^2} = 2.92$$

Now.

		Cen 1	cen 2
group 1	1	0	7.21
	2	1.12	6.10
	3	3.61	3.61
group-2	4	7.21	0
	5	4.72	2.5
	6	5.31	2.06
	7	4.30	2.92

Weakness of K-mean:

① কম data থাকলে ভুল grouping এ cluster হয়

② cluster, K আগেই determine করতে হয়, cluster depends on initial assignment

③ never know the real cluster,

④ different initial condition এ ধরন different cluster হয়।

Application of K-mean:

① efficient & fast. time = $O(tkn)$

n = number of object

k = " " cluster

t = " " iteration.

② Machine learning & data mining এ use হয়

③ wave → K category এ use হয়, wave, speech data এ use হয়।

④ color plot & graphical display এ use হয়।

Lecture = 5 Classification

Classification process: 3 steps

① **model construction** (Learning) pre defined class এ record করেছে। যা determine করেছে target attribute যাকে,

value of target attribute = class labels,

Training set = learning এ use হয়।

② **Model Evaluation:** (Accuracy) estimate accuracy rate based on test set.

③ **Model use:** (Classification) model is used to classify unseen instances. Predict value.

Classification Method: ① ~~A~~ Decission tree

② Bayesian classification

③ K- Nearest neighbor

④ Neural network

⑤ SVM

⑥ Genetic Algorithm.

Naive Bayesian learning

Naive bayes play important role on prediction.

$$P(H|x) = \frac{P(x|H) \cdot P(H)}{P(x)}$$

$P(A|B)$ = prob of A, given B.

B is known.

$P(A \wedge B) = P(A|B) \cdot P(B)$. $\boxed{****}$

Bayes's rule = $P(A \wedge B) = P(A|B) \cdot P(B)$

$P(B \wedge A) = P(B|A) \cdot P(A)$.

but, $P(A \wedge B) = P(B \wedge A)$.

$\therefore P(A|B) \cdot P(B) = P(B|A) \cdot P(A)$.

$$\Rightarrow P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}.$$

So,
$$P(H|E) = \frac{P(E|H) \cdot P(H)}{P(E)}$$

$P(A|B) = P(A) \quad ; \quad P(B|A) = P(B)$

two constraints are logically equal.

$$P(A|B) = \frac{P(A \wedge B)}{P(B)} = P(A)$$

$$P(A \wedge B) = \cancel{P(A)}\ P(A|B) \cdot P(B) = P(A) \cdot P(B).$$

$$P(C_i | E) = \frac{P(C_i) \cdot P(E|C_i)}{P(E)}.$$

---

math

Problem: Too many possible combination.

Estimate = Estimated based on training data.

= compute mean & standered deviation

Smoothing: small training set, হলে estimate smooth হয়।

$P(Yes) = \frac{9}{14}$ — yes — all

$P(Sunny | Yes) = \frac{2}{9}$ — sunny → yes — total yes.

$P(\text{humidity} \mid No) = P(\text{high} \mid No) = \frac{4}{5}$

4 for high, 5 for actual no.

outlook	P	N
Sunny	2/9	3/5
Overcast	4/9	0
rain	3/9	2/5

Humd.	P	N
high	3/9	4/5
normal	6/9	1/5

Temp	P	N
hot	2/9	2/5
mild	4/9	2/5
cool	3/9	1/5

Wind	P	N
true	3/9	3/5
false	6/9	2/5

given,

$$x = \langle \text{sunny, mild, high, true} \rangle$$

$$P(x \mid NO) = \frac{3}{5} \times \frac{2}{5} \times \frac{4}{5} \times \frac{3}{5} = \frac{72}{625}$$

$$P(x \mid Yes) = \frac{2}{9} \times \frac{4}{9} \times \frac{3}{9} \times \frac{3}{9} = \frac{72}{6561}$$

Now,

$$P(x \mid NO) \cdot P(NO) = \frac{72}{625} \times \frac{5}{14} = 0.041$$

$$P(x \mid Yes) \cdot P(Yes) = \frac{72}{6561} \times \frac{9}{14} = 0.007$$

convent in probabilities,

$$P(NO|x) = \frac{0.04}{0.04+0.007} = 0.85$$

$$P(Yes|x) = \frac{0.007}{0.04+0.007} = 0.15.$$

Answ: Not play. classified as No.