CHENNAI MATHEMATICAL INSTITUTE
https://www.cmi.ac.in

1	2	3	4	5	6
3	5	5	5	5	3

Mid Semester Examination, Jan–Apr 2024

24/30 20/30

Name: Anitra Majumder	Roll Number: MS202304
Subject: DMML	Date:
Course & Year:	Total No. of Pages:

—— Begin here ——

1) number of transaction (T) $= 10^{10}$

total no of elements (not necessarily unique) are $\leq 10^{10} \cdot 10$
$= 10^{11}$

set of item has size (10^7)

an item is frequent if it appears in more than.

$10^{10} \cdot \frac{1}{1000}$ transaction
$= (10^{10-3})$ transaction

So, atmost $\frac{10^{11}}{10^{10-3}}$ item can be be frequent
$= 10^{8-4}$ ✓

So total 10^4 items make frequent

$|F_1| \leq 10^4$

Now C_2 will have form $= \{(x_1, x_2) \mid x_1 \in F_1 \wedge x_2 \in F_1\}$

So, $|F_2| \leq |C_2|$
$\leq {}^{10^4}C_2$
$= \frac{10^4(10^4-1)}{2}$
$= 5 \cdot 10^3 (10^4 - 1)$

Better estimate by directly calculating F_2, like F_1

3

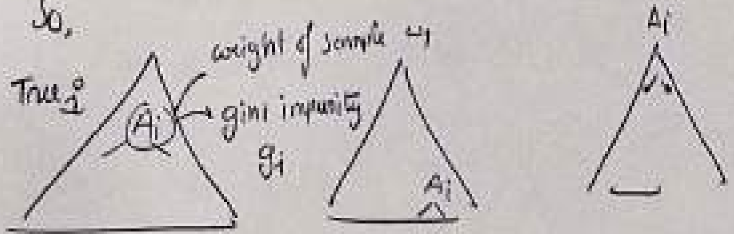
(4) In random forest classifier,   ↓ set of attributes from which to choose

There will be multiple decision-trees and further if (m) is low.

the correlation and strength will be lower. So the attribute on which the tree is split consecutively changes,

i.e. the attribute (splitter) will come at various heights of the different trees. (may not come too)

So,



$$\text{So, gini impurity}_{(A_i)} = \frac{\sum_j (w_j \cdot g_j)}{\sum_j w_j} \quad \text{if } A_i \text{ appear in Tree } j$$

Not only every A_i has a good chance to appear as a splitter but also,

it's gini impurity is measured at different levels.

Then, they (A_i)s can be ranked based on the gini impurity.

Weighted Average

But for decision-tree, some input features may not occur or they have only one chance in a path to be appeared and the one with more IG is chosen as the splitter.

A splitter may be good or bad at certain levels, decision tree doesn't handle that case. So more effective for random forest

5

~~3~~

6. locally linear embeddings are better alternative for PCA if the data points are not linear (lie along a manifold)

In the above case, lets say there are m datapoints as x_i and we take k nearest datapoints of it

we say discrepancy of x_i = $\left(x_i - \sum_{j=1}^{m} w_{ij} x_j\right)$

$$= \left(x_i - \sum_{j=1}^{m} w_{ij} x_j\right)$$

where $w_{ij} = 0$ if x_j not in neighbourhood of x_i

Now we will in general try to minimize the mean squared or mean discrepancy over the whole dataset (SSE equivalent) or MSE

So, $$J_{MSE} = \frac{1}{2m} \sum_{i=1}^{m} \left(x_i - \sum_{j=1}^{m} w_{ij} x_j \right)$$

Now $\hat{W}$ is the weight matrix st

$$\hat{W} = \arg\min_{W} J$$

This min can be achieved by gradient descent or direct calculation

Once, we get the $\hat{W}$ matrix.

for each unseen point, we will get K nearest neighbours

get the labels for them and then

predict class as based on K nearest neighbour and

weight matrix [redacted]

Weights → Embedding ?

(3)

(3) ☒ (A_1) (A_2)

I am assuming an attribute — is more significant than other.

if we only use A_1 to predict y, the result will be better than if we have used A_2 as the only attribute

while

let, A_1 has parameter θ_1, A_2 has θ_2

In that case, we know (x_i^1: 1st component of vector x; x_i^2: 2nd component of vector x)

$$\frac{1}{2}\Sigma(y_i - \theta_1 x_i^1)^2 \leq \frac{1}{2}\sum_{i=1}^{n}(y_1 - \theta_2 x_i^2)^2$$

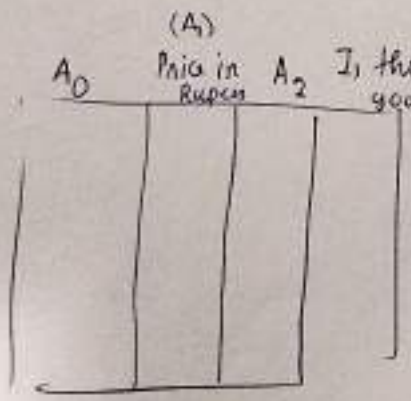
$$\Rightarrow \Sigma\left[(y_i - \theta_1 x_i^1)^2 - (y_i - \theta_2 x_i^2)^2\right] \leq 0$$

$$\Rightarrow \Sigma\left[(2y_i - \theta_1 x_i^1 - \theta_2 x_i^2)(\theta_2 x_i^2 - \theta_1 x_i^1)\right] \leq 0$$

for this condition to be true,

its not enough to make θ_1 larger than θ_2 as x_i^1 may have higher values compared to x_i^2.

To illustrate this case consider two tables



let's say price in rupees is the most important so, $\theta_1 \geq \theta_2$ and $\theta_1 \geq \theta_0$

but,

PTO

The same data can be used with inputs changed to price and obviously, the prediction will be same

So $\theta_1' = \frac{\theta_1}{100}$ which not necessarily mean $\theta_1' > \theta_2$ and $\theta_1' > \theta_0$

☒ The second interpretation could be an attribute is important if small change in it will make a good change in prediction value.

By definition. more important features will have higher values of θ than

Considering both the cases,

whether importance of θ_1 (magnitude of θ) implies most significant feature depend on the definition of significance of attribute

OK (5)

(5) Some observation:

let's take two points (x_i, x_j) among the dataset

(★) distance

if x_i and x_j are small enough. (wrt the average distance between points)

d_i, d_j

the two vectors will be close enough in N dimension

Intuition:

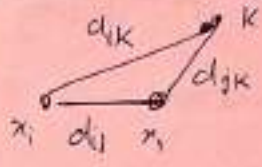
	d_i	d_j
1	d_{i1}	$\frac{1}{2}\|d_{i1} \pm \delta$
	d_{i2}	$\leq d_{i2} + \delta$
i	0	d_{ij}
j	d_{ij}	0

$\delta \leq d_{ij}$

(good approximation for)

the δ will be d_{ij} itself

as



$d_{ij} + d_{jk} > d_{ik}$

$\therefore (d_{jk} - d_{ik} < d_{ij})$

as d_{ij} is small.

two vectors will be in close vicinity, then distance will be

$\leq \sqrt{N (d_{ij})^2}$

$= d_{ij}\sqrt{N}$

between x_i and x_j

(★) if distance is large, then the how large it will be, depends on

1) d_{ij}

2) the structure of cluster

but both the cases, it will be much larger compared to the 1st case

Induction: 1) If the x_i, x_j points are far apart compared to average distance

then, assuming random distribution points are again d_{ik} and d_{jk}

d (direction)

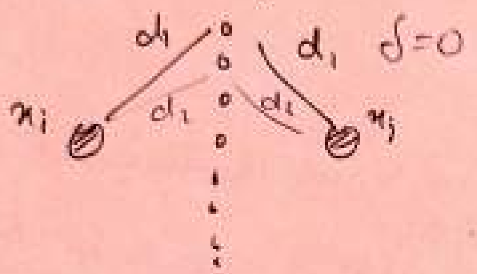
a lot
will be ~~hardly~~ different, therefore increasing their distance in N d...

In this case too $\delta \leq (d_{ij})$

but $\uparrow$ itself is large.

2>

If we consider some specially created distribution,



in those cases d_{ij} will be large enough and that will be the only one contributing to the distance of d_i and d_j in N dimension

So, if we cluster this way,

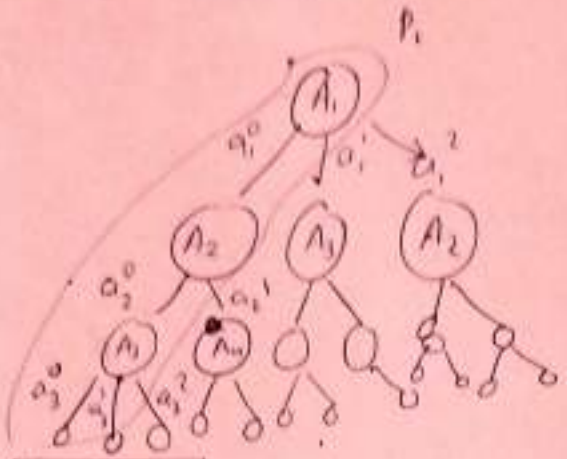
if x_i and x_j are nearby, dist (d_i, d_j) is less
so, d_i and d_j are nearby

if x_i and x_j are large, d_i and d_j are at distance too
though distance depends on the structure
(which might be actually good)

Considering those, for general clusters, clusters formed by column of D will be an overall good approximation of X

Class association rule have a attribute (class label) as their target

Now, if we take a sample decision tree,



Class label | C_1 C_2 C_2 |

Now, we can generate class association rules as follows:

take a path (P_1 from figure), then

$$(A_1 = a_1^0, A_2 = a_2^0, A_3 = a_3^0) \longrightarrow C_1$$

⋮

for all such root to leaves path, we will get such association rules

Now, to generalize a class association rule, we can do the following,

$(A_1 = a_1^0, A_2 = a_2^0, A_3 = a_3^0) \longrightarrow C_1$ for 225 samples

$(A_1 = a_1^0, A_2 = a_2^0, A_3 = a_3^1) \longrightarrow C_2$ for 100 / 101 samples

⋮

$(A_1 = a_1^0, A_2 = a_2^0, A_3 = a_3^2) \longrightarrow C_2$ " 200 / 202 "

So, $(A_1 = a_1^0, A_2 = a_2^0) \longrightarrow C_2$ for 300 / 305 samples

removing A_3 as a splitter

In this case, we have generalized the association rules by clubbing them into 1.

This is so equivalent to pruning the last level (not asking the last splitting question A_3)

This would be different as, ~~if we club~~ in general if we club or prune from bottom (as was the previous case).

what if class association generalization rule has chosen to remove A_1 instead of A_3 to generalize better

Bottom up pruning won't help, in such case

(Top down pruning/clubbing) might be helpful

remove A_1 from the class association rule $\Rightarrow$
change the 1st splitting Question to based on attribute (A_2 and the A_3)

So, class association generalization on decision tree is more exhaustive wrt usual bottom up pruning