1	2	3	4	5	6
5	4	5	3.5	5	3

$\frac{25.5}{30}$

(1)

Name: Siddhant Shah	Roll Number: BMC202171
Subject: DMML	Date: 2/3/24
Course & Year: BSc Year 3	Total No. of Pages: 4

--- Begin here

* Question 1 :-

No of transaction = 10^{10}

Size of each transaction ≤ 10

~~Total~~ Max no of tokens: 10^{11}

Each frequent token occurs ~~at~~ in ~~most~~ at least $10^{-4} \times 10^{10}$ ~~times~~ transaction $= 10^6$

Maximum such tokens $= 10 \times \frac{10^7}{10^6} = 100$

Each transaction has at most 10 tokens

$\therefore n(F_1) \leq 100$

~~Now say all combin~~ F_2 is made of union of two element of F_1, such that the union is also frequent.

~~Now $n(F_2) \leq n(F_1)$~~

$\Rightarrow \{x, y\} \in F_2 \iff freq(\{x,y\}) \geq 10^6$ and $x, y \in F_1$

Maximum such pairs $= \frac{1}{2} \times 10 \times \left(\frac{10^7}{10^6}\right) = 50$

length of sequence

$n(F_2) \leq 50$

* Question 3 :-

The claim is not justified as it doesn't take into account the average value of the attribute.

For example, say ~~x_1 = years of education~~ x_1 measures distance in cm and $\theta_1 = 3$

and x_2 measures length in inches and $\theta_2 = 2$

Taking these values directly will not ~~be~~ give an accurate answer as they have to be normalized.

Another example where this would fail is when we consider x_1 that varies between 1 and 10 ~~with~~ with an average value of 5, $\theta_1 = 0.1$

and x_2 that varies between 1 and 100 with an average value of 60, $\theta_2 = 0.01$

If we simply compare θ_1 and θ_2, then we get that the first attribute has ~~at~~ a larger contribution.

However the average contribution is $\theta_1 \text{Avg}(x_1) = 0.5$ and $\theta_2 \text{Avg}(x_2) = 0.6$ respectively, which contradicts our earlier claim.

5

* Question 5 :-

The i^{th} column of D, D^i ~~has~~ contains the distances of all the points from x_i

$$D^i = \begin{bmatrix} d(x_1, x_i) \\ d(x_2, x_i) \\ \vdots \\ d(x_n, x_i) \end{bmatrix}$$

Running clustering on the columns of D will give us points that are at a similar distance from each of the points in the ~~plan~~ plane.

For example, if D^i, D^j are in the same cluster, ~~st~~ with the ~~max~~ ~~di~~ radius ϵ, we get that.

$$\sqrt{(d(x_1,x_j) - d(x_1,x_i))^2 + \dots + (d(x_n,x_j) - d(x_n,x_i))^2} \leq \epsilon$$

$$\Rightarrow |d(x_k, x_j) - d(x_k, x_i)| \leq \epsilon \quad \forall k \; \cancel{k} = 1, \dots n$$

let $k = i$ or $k = j$ to get

$$|d(x_i, x_j)| = d(x_i, x_j) \leq \epsilon$$

This can be done for ~~point~~ points x_i, x_j such that D^i, D^j are in the same cluster.

5

* Question 6 —

Suppose we have n points. We express each point x_i as a linear sum of its k nearest neighbours (k is predecided)

$$\therefore \quad x_i = \sum_{j=1}^{n} w_{ij} x_j$$

$w_{ij} = 0$ if x_j is not one of the k-nearest neighbours of x_i ~~$(w_{ij} = 0$~~ $(w_{ii} = 0)$

~~So~~ This gives us a matrix W.

We have to find W so that the squared distance is minimized

$$\therefore \quad \hat{W} = \underset{W}{\operatorname{argmin}} \sum_{i=1}^{n} \left(x_i - \sum_{j=1}^{n} w_{ij} x_j\right)^2$$

This gives us the locally linear embedding.

→ w_{ij} that are such that x_j is a k-nearest neighbour are learnable parameters, for all $i = 1, \ldots n$

Weights → Embedding

3

* Question 2 :-

Given a decision tree.

→ For categorical variables, add the values that are permitted for traversal along that path ~~to~~ to a list

Each categorical variable is queried only once

→ For numerical variable, do the same thing but create a range

If all values < 10 are permitted, add $[-\infty, 10]$ to the list

If the ~~value~~ variable is queried again, take the intersection of the previous range for that variable, with the current proposed range

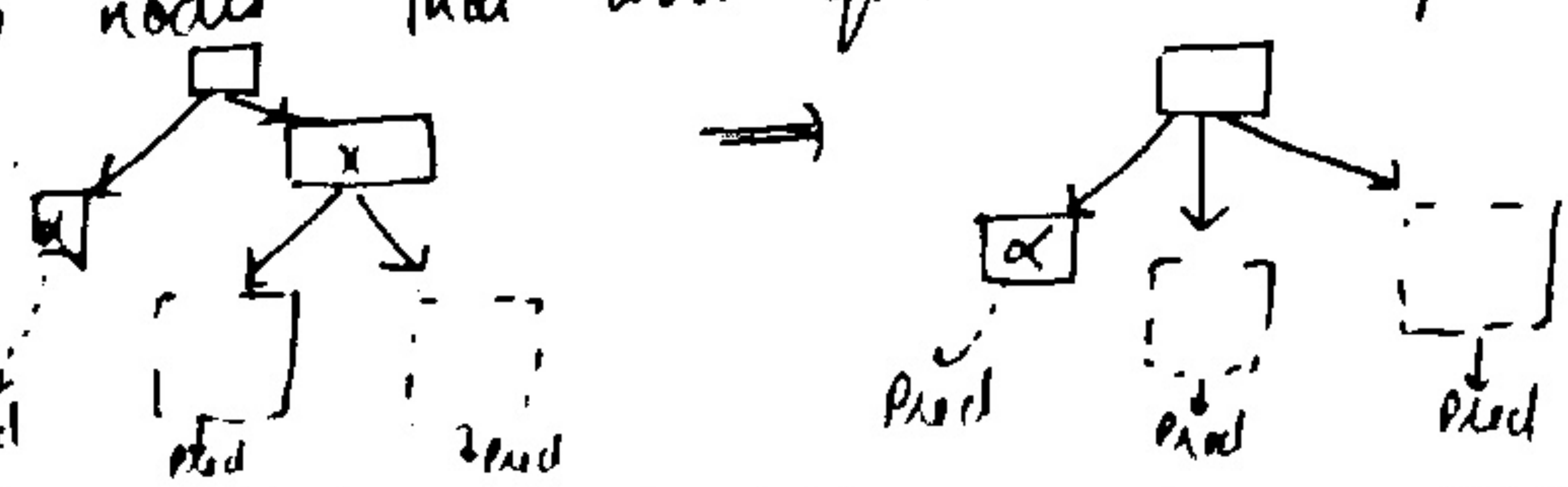
→ At the ~~xx~~ end of the path, we get the prediction of the tree. Add that to the left side of the association rule

Thus, we will be left with rules that look like

$\{$ categ var i_1 = value i_1, categ var i_1 = value i_2, ... ,
num var $a_1 \in (l_1, m_1)$, ... $\} \longrightarrow$ Prediction

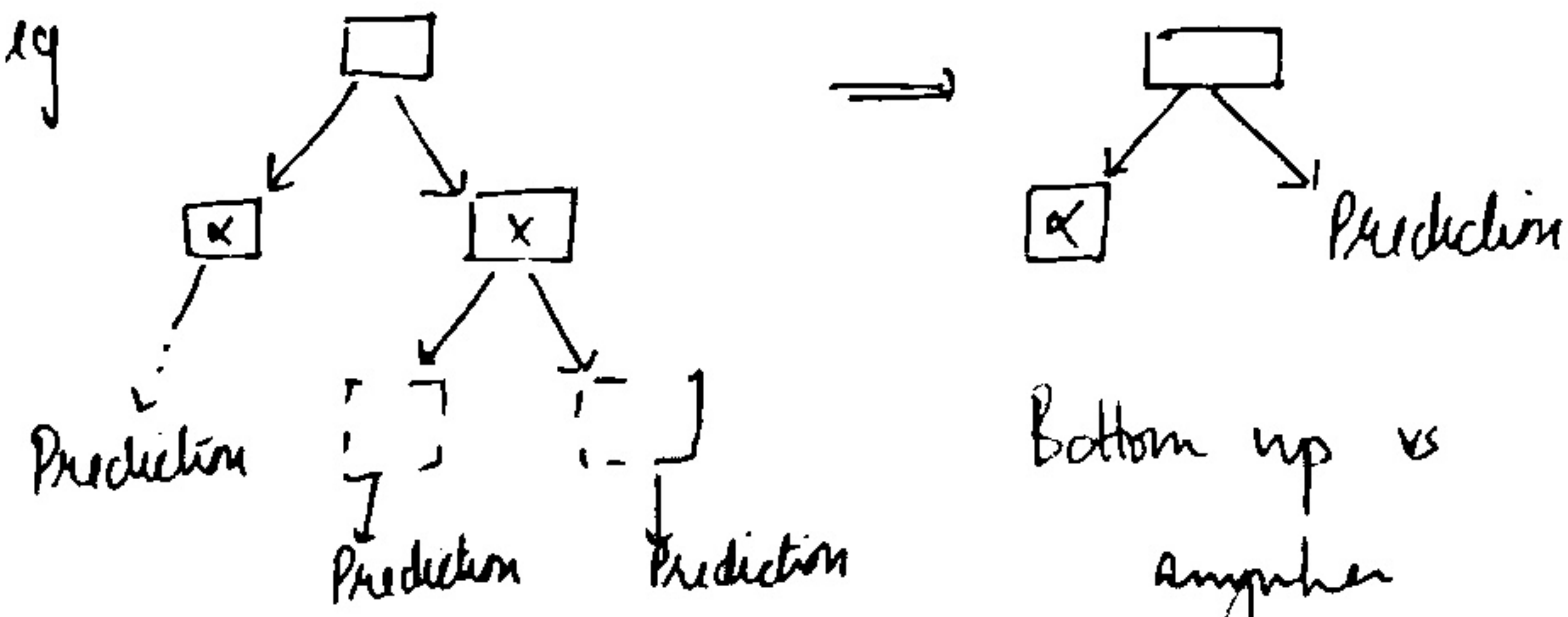
Removing an attribute from the left side would be to not query that attribute at all in the tree but to keep the part of the tree that was under the nodes that ~~xxx~~ queried the respective attribute.

eg.



The usual methods of pruning either limit the max depth or place a restriction on splitting a node. When we remove a node, we merge all the subnodes into the parent nodes.

eg



Bottom up vs anywhere

4

---

* <u>Question</u> 4 :—

In a ~~non~~ random forest built on the same attributes as those we wish to rank,

For each attribute, we take some metrics to quantify the central tendency of the impurity of any node that queries on that attribute (mean, max, min)

weighted impurity gain

We rank the attributes by the <u>mean impurity</u>. in decreasing order, as an attribute with lower impurity classifies the data better.

The random forest is better than a single decision tree as the decision tree will only be formed by maximizing the information gain at

each step but won't be able to get a globally min config for the data due to its high variance.
Averaging over the random classifier increases the accuracy ~~#~~ and decreases variance.

eg It is possible that a single decision tree won't create a node that has 50% impurity as there may be another attribute with lesser impurity.
But it may so happen that after creating ~~that~~ the earlier node, ~~create~~ adding one more node reduces the impurity drastically.
This ~~is~~ ~~that~~ tree is likely to be formed in the random forest and it should be taken into account in ranking the attributes.

3.5

---

* Question 1 :—

Each freq item has to occur in at least $10^{10} \times 10^{-3}$
$= 10^{7}$ transactions

Consider that each frequent item occurs exactly 10^{7} times
No. of frequent items $= \left(\frac{10^{10}}{10^{7}}\right) \times 10 = 10^{4}$

$\therefore n(F_1) \leq 10^{4}$ ✓

Now for F_2, ~~if each item in a transaction~~
if the same transaction is repeated 10^{7} times, every pair of items in it will be in F_2

There can be such 10^3 ~~pairs~~ transactions.
Each transaction gives us such $\binom{10}{2} = 45$ pairs

$\therefore$ Total pairs $= 45 \times 10^3 \geq n(F_2)$

No. of elements used $= 10 \times 10^3 = 10^4 \leq 10^7$ ~~but this is not possible because $10^5 \geq 10^4 \geq n(F_1)$~~

$\Rightarrow n(F_2) \leq 4.5 \times 10^4$

$\therefore$ $n(F_1) \leq 10^4$
$n(F_2) \leq 4.5 \times 10^4$

5