Last class (September 1)

1) Blinkit Problem Statement
   - issue with logistic regression
2) Geometric Intuition of KNN
3) KNN Algorithm
4) KNN scratch code
5) Assumptions of KNN
6) sklearn's KNN implementation
7) Bias-Variance trade-off
8) Train & Test time complexity of KNN

Today's class

0) Recap - Quizzes
1) KNN with categorical features
2) Different distance metrics
3) LSH for KNN
4) Missing data with KNN — imputation
5) Employee Attrition Problem Statement
6) Decision Tree Intuition
7) How to split the nodes
8) Purity and Impurity of the nodes

Target Encoding

x_2	x_1	x_2	y
0.75	0.8	RED	1
0.5	0.1	GREEN	0
0.75	2.2	RED	1
0.75	-1.5	RED	1
0.5	-0.9	GREEN	1
0.75	3.2	RED	0

Target / Output

$$P(y=1 \mid x_2 = \text{GREEN}) = \frac{P(y=1 \cap x_2 = \text{Green})}{P(x_2 = \text{Green})} = \frac{1}{2} = 0.5$$

$$P(y=1 \mid x_2 = \text{RED}) = \frac{P(y=1 \cap x_2 = \text{RED})}{P(x_2 = \text{RED})} = \frac{3}{4} = 0.75$$

OHE

3 values → RED, GREEN, YELLOW

$3 - 1 = 2$ → sufficient

→ Yellow

Not Red	is_Green	is_Red	is_Yellow
Not Green	0	0	1
Red	0	1	0
Green	1	0	0

N unique cat values ⇒ OHE needs N-1 dimension

num feat → n-f

≤ 6 → OHE

> 6 → Target Encoding

Categorical values: A, B, C

3 classes: 0, 1, 2

$P(y=0 \mid x = A/B/C) = \underset{A}{0.5}, \underset{B}{0.25}, \underset{C}{0.25}$

$P(y=1 \mid x = A/B/C) = \underset{A}{0.3}, \underset{B}{0.6}, \underset{C}{0.1}$

$P(y=2 \mid x = A/B/C) = \underset{A}{0.4}, \underset{B}{0.45}, \underset{C}{0.15}$

$V_1 = [-5, 5]$

$V_2 = [-3, 3]$

Similarity $\propto \frac{1}{\text{Distance}}$

$\cos(V_1, V_2) =$

L2-norm of $V_1 = \sqrt{(-5)^2 + 5^2} = \sqrt{50}$

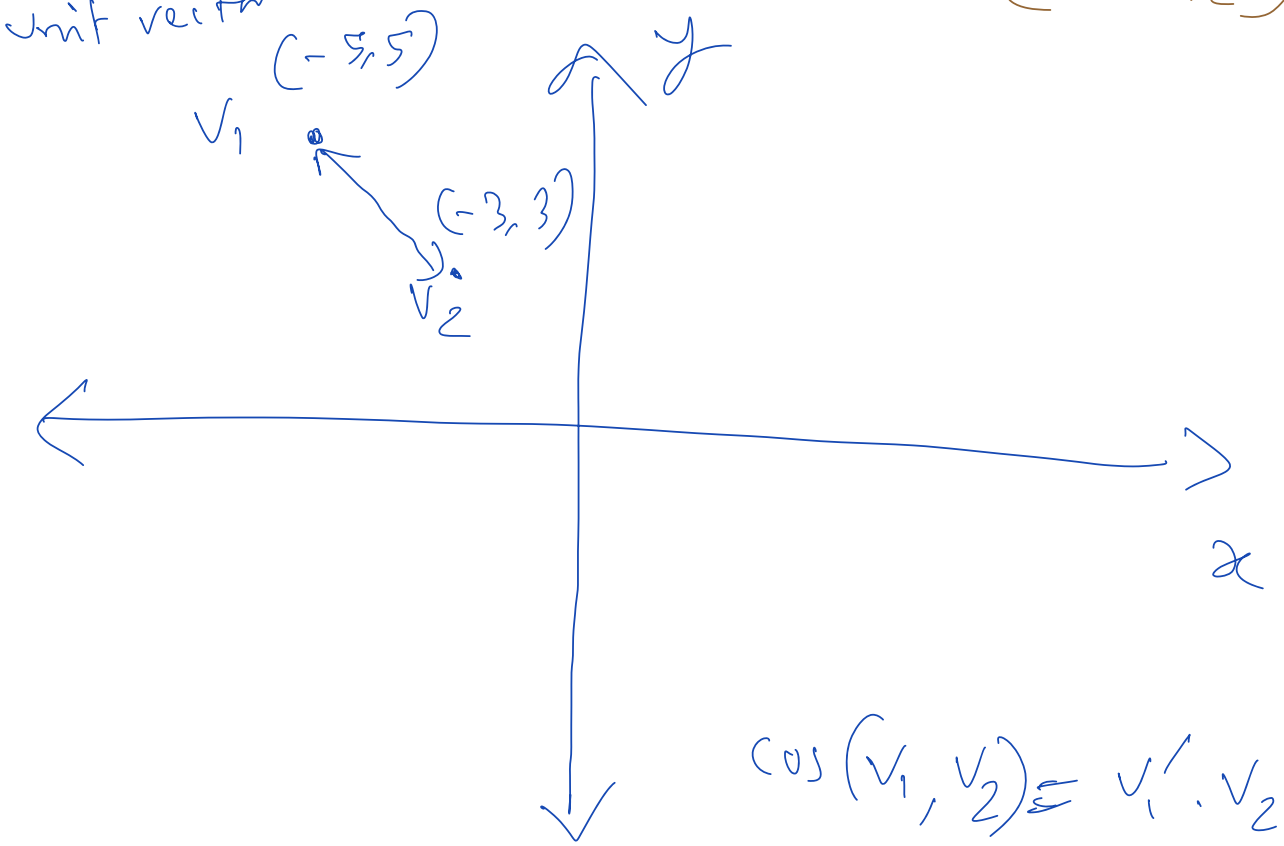
$V_1' = [-\frac{5}{\sqrt{50}}, \frac{5}{\sqrt{50}}]$

unit vector $= [-\frac{5}{5\sqrt{2}}, \frac{5}{5\sqrt{2}}] = [-\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}]$

L2-norm of $V_2 = \sqrt{(-3)^2 + 3^2} = \sqrt{18} = 3\sqrt{2}$

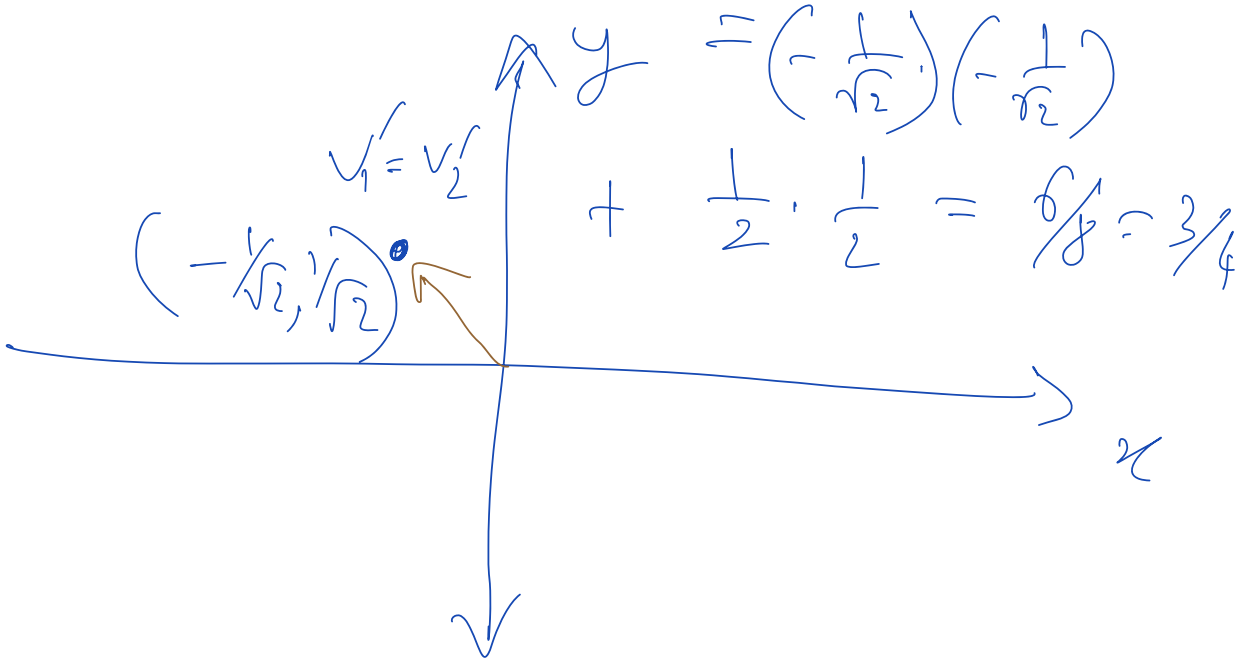
$$v_2' = \left[-\frac{3}{3\sqrt{2}}, \frac{3}{3\sqrt{2}} \right] = \left[-\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}} \right]$$

unit vector



$$\cos(v_1, v_2) = v_1' \cdot v_2'$$

$$= \left(-\frac{1}{\sqrt{2}}\right)\left(-\frac{1}{\sqrt{2}}\right) + \frac{1}{2} \cdot \frac{1}{2} = 6/8 = 3/4$$



Angle between v_1 & v_2 → small,

⇒ cosine similarity → 1

very high

$\cos(\theta) = -1$ to $+1$

Cosine similarity $\propto \frac{1}{\text{Euclidean Distance}}$

cosine distance = 1 − cosine similarity

Test time for kNN: $O(nd + n \log n)$

assuming k is very small

$k << n$

1 B $\Rightarrow 10^9$

100 $\rightarrow 10^2$ $\Big\}$ $\downarrow 10^7$ times

	f_1	f_2	f_3	f_4	f_5	y
x_1	0.5	1.1	-1.1	10	7	
x_2	10.0	18	NAN	15	14	
x_3	13.0	7	8	9	13	
			⋮			
x_{100}						

→ 1.5

Drop f_3

	f_1	f_2	f_4	f_5	y
x_1	0.5	1.1	10	7	
x_2	10.0	18	15	14	
x_3	13.0	7	19	13	
	⋮	⋮	⋮	⋮	⋮
x_{100}					

KNN

↓

Get 4 NN of x_2

x_1

x_5

x_{10}

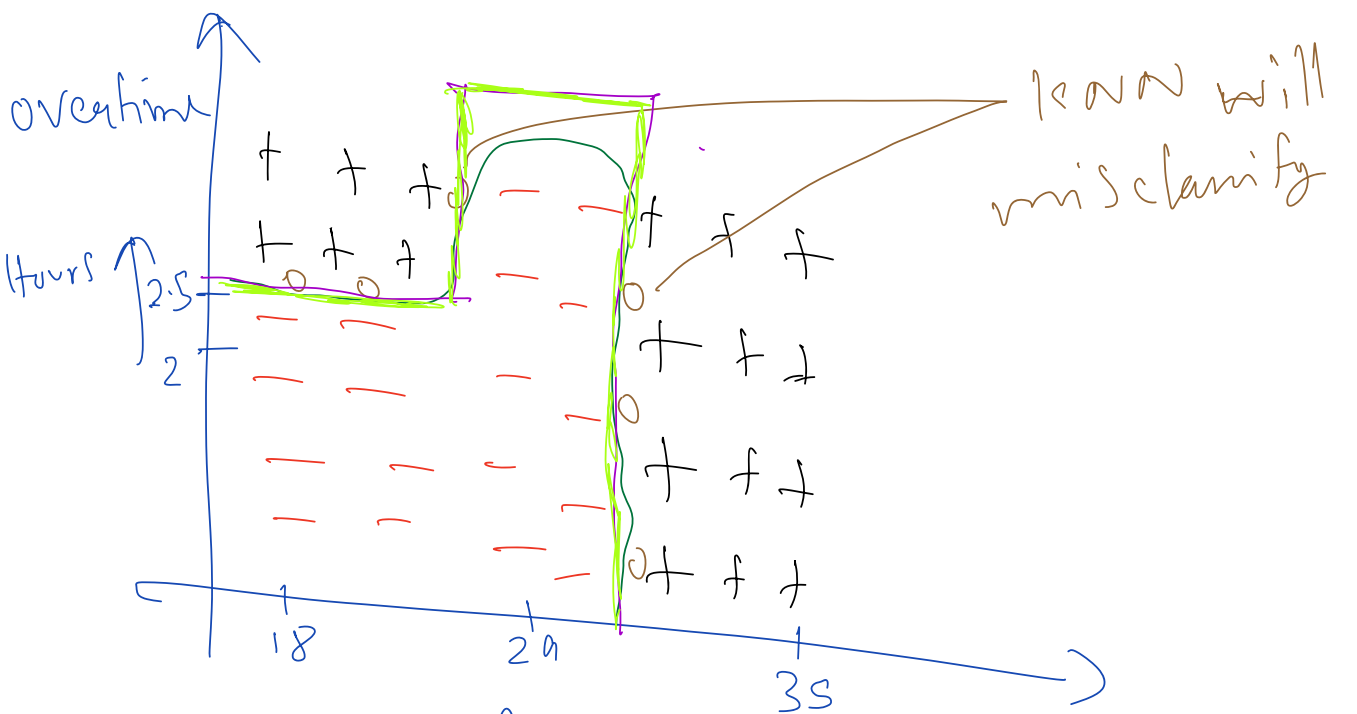
x_{100}

4 Nearest neighbors

$x_{23} = 2$

$\text{avg}(x_{1,3}, x_{53}, x_{10,3}, x_{100,3})$

$x_{23} = 1.5$



+ve class
-ve class

Node A

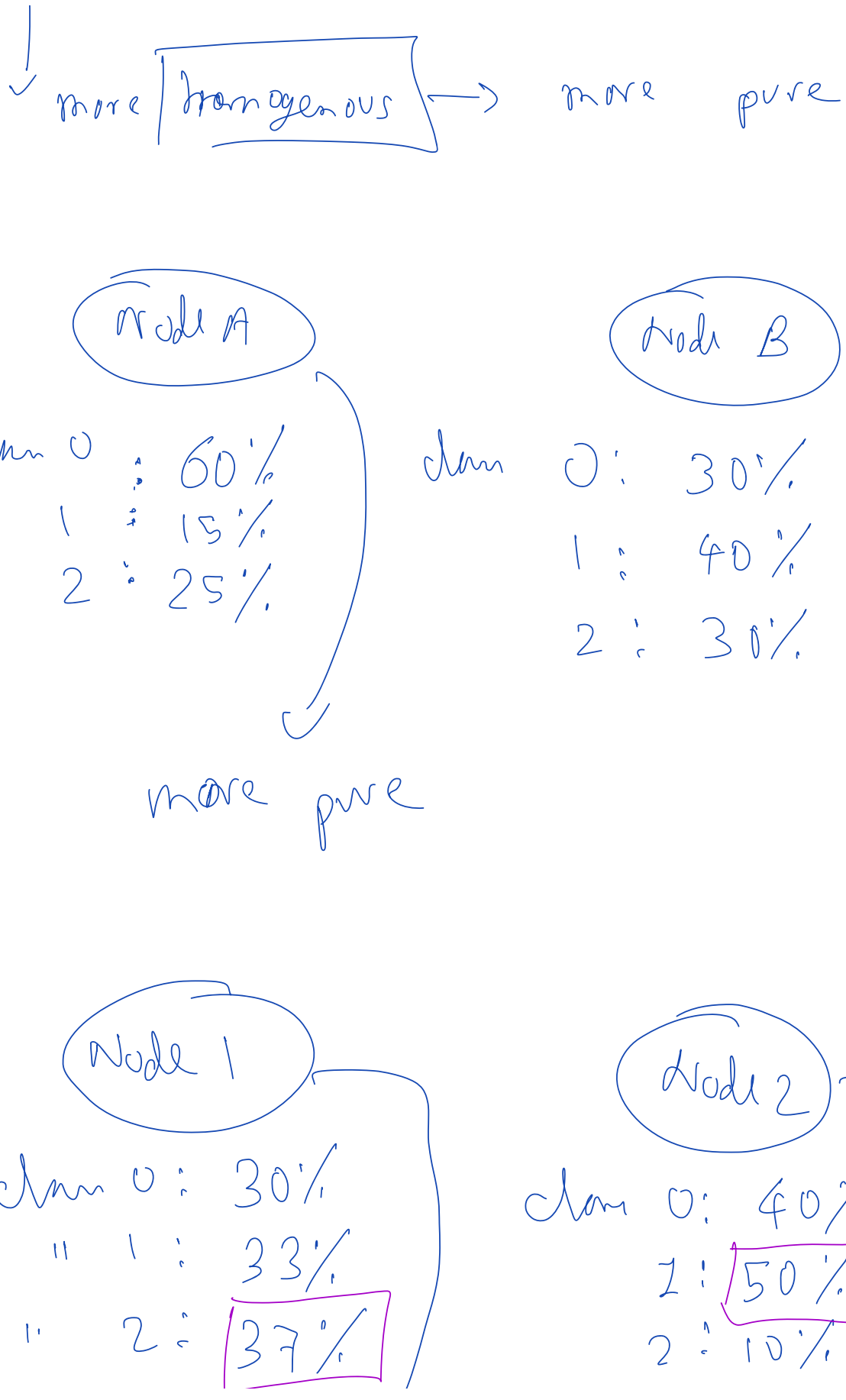
+ve : 90 → 90%

-ve : 10

Node B

+ve : 60 60%

-ve : 40


more homogenous
more pure
Node A
class 0 : 60%
1 : 15%
2 : 25%
more pure
Node B
class 0: 30%
1: 40%
2: 30%
Node 1
class 0: 30%
" 1: 33%
" 2: 37%
Node 2
class 0: 40%
1: 50%
2: 10%

f_1	f_2	$f_3 \cdots$	f_{10}
$\tau_{1,1}$	$\tau_{2,1}$		
$\tau_{1,2}$	$\tau_{2,2}$		
$\vdots$	$\vdots$		
$\tau_{1,10}$	$\tau_{2,10}$		

100 +ve
100 -ve

200 → x_i's

$f_1 > \tau_{1,1}$ <

$f_1 > \tau_{1,10}$ <

1000

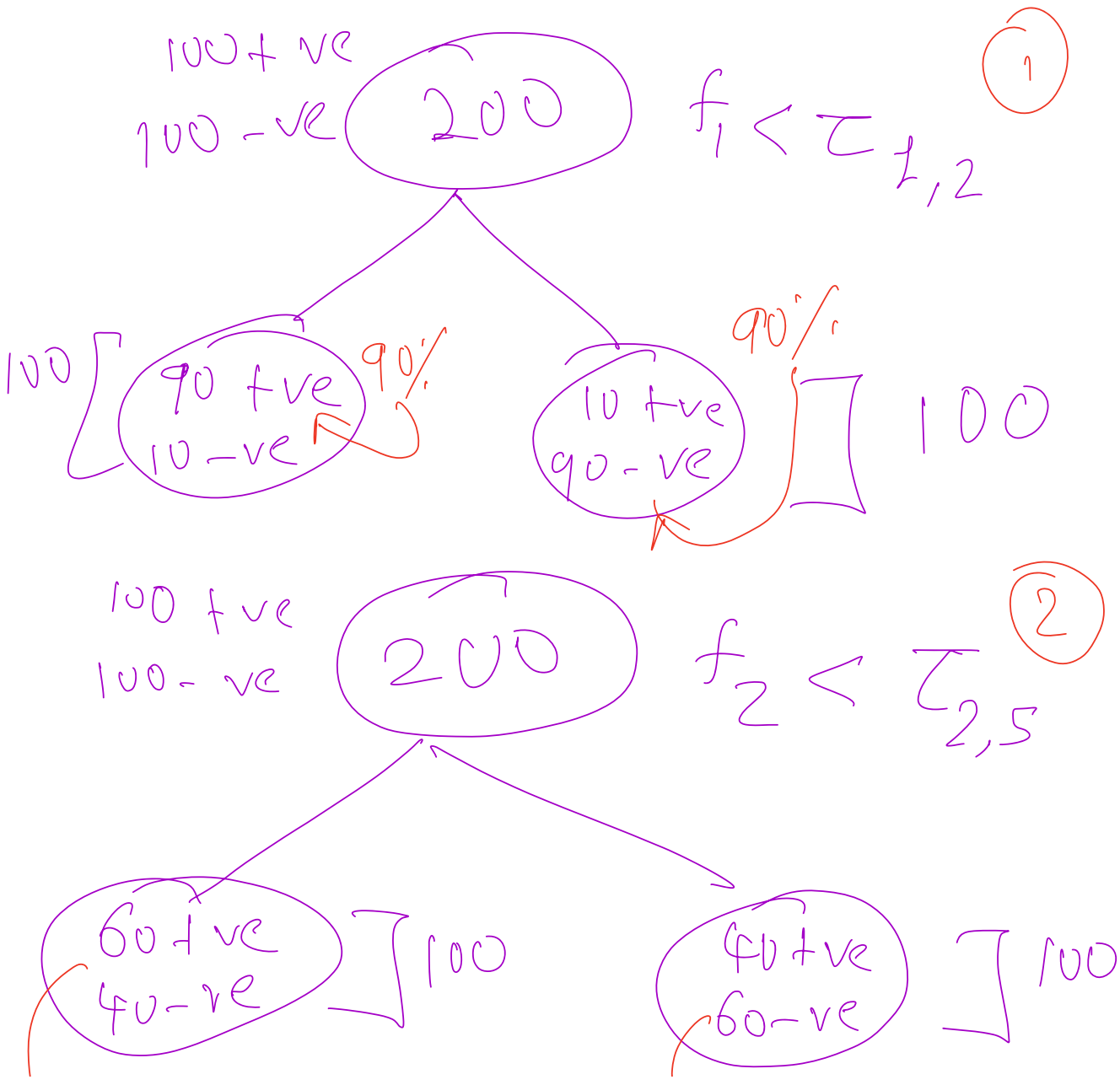
...

$\vdots$

possibilities

$f_2 > \tau_{2,1}$ <

$f_3 < \tau_{3,10}$ <



↳ 60% +ve

↳ 60% -ve