Last Class (Sep 04)

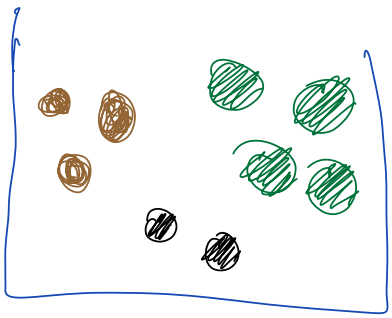
1) KNN with categorical features
2) Different distance metrics
3) LSH for KNN
4) Missing data with KNN — imputation
5) Employee Attrition Problem Statement
6) Decision Tree Intuition
7) How to split the nodes
8) Purity and Impurity of the nodes

Today's class

0) Quizzes ✓
1) Recap
2) Entropy
3) Information Gain
4) How to split a node — categorical & Numerical
5) Issue with Entropy
6) Gini Impurity
7) Feature Scaling
8) Overfit vs Underfit
9) Hyper-parameter Tuning (Time Permits)

numerical F_1	F_2 Categorical
0.09	GREEN
0.1	RED
−1.5	BLUE
11.7	YELLOW

All distinct values



$$E = -\left[\tfrac{1}{3}\log_2\left(\tfrac{1}{3}\right) + \tfrac{4}{9}\log\left(\tfrac{4}{9}\right) + \tfrac{2}{9}\log\left(\tfrac{2}{9}\right) \right]$$

$P(\text{brown}) = 1/3$

$P(\text{green}) = 4/9$

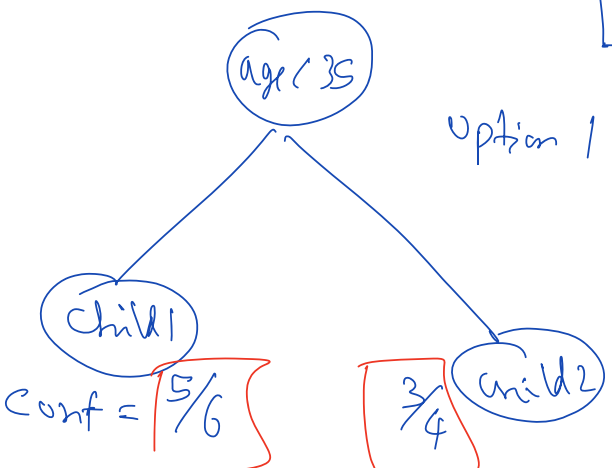
$P(\text{black}) = 2/9$

$$E = -\left[p_0 \log(p_0) + p_1 \log(p_1) \right]$$

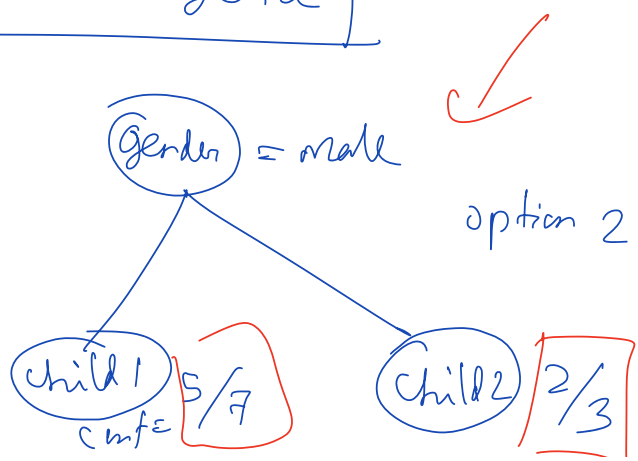
$$\text{log-loss} = -\left[y_0 \log(p_0) + y_1 \log(p_1) \right]$$

y_0, y_1: ground truths

$\boxed{\text{KL Divergence}}$



option 1



option 2

→ no! Confidence desired

proportion of majority class = confidence of predictions

$H[\text{parent}] = 0.97$

$H[C1] = 0.65$ $\quad C_1 = 60$ samples

$H[C2] = 0.81$ $\quad P(C_1) = \frac{60}{100} = 0.6$ ✓

$C_2 = 40$ samples

$P(C_2) = \frac{40}{100} = 0.4$ ✓

weighted E for children $= WH[C]$

$= P(C_1) \times H[C1] + P(C_2) \times H[C_2]$

$= 0.6 \times 0.65 + 0.4 \times 0.81$

$= 0.714$

$\boxed{\text{Information Gain}}$ $(IG) = \boxed{H[\text{parent}]} - WH[C]$

$= 0.97 - 0.714$

$= \boxed{0.256}$ ✓

Gender condition

$H[P] = 0.97$

$H[C_1] = 0.86$

$H[C_2] = 0.91$

$P[C_1] = 7/10$

$P[C_2] = 3/10$

$$WH[C] = P[C_1] \times H[C_1] + P[C_2] \times H[C_2]$$

$$= \frac{7}{10} \times 0.86 + \frac{3}{10} \times 0.91$$

$$= 0.875$$

$$IG = 0.97 - 0.875$$

$$= \boxed{0.095}$$

$\boxed{0.251} > \boxed{0.095}$ → option 2

↓

Option 1 which is age < 35

condition

Information Gain ∝ Reduction in impurity

$\boxed{1000 \text{ conditions}}$

↓

$\boxed{2 \text{ classes}}$

$f_j > \tau_1$

$f_i \geq \tau_2$

$f_3 < \tau_3$

⋮

every condition out of 1000 ⇒ some info gain

↓

best / max info gain

$O(1000) \rightarrow O(n)$

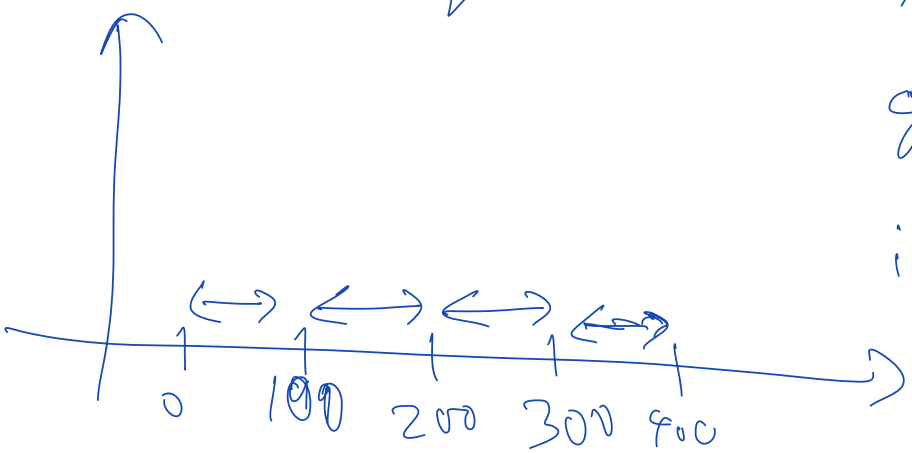
$\boxed{\text{number of conditions}}$

age = [1 to 1000]

range

90, 994, 70, ...

grouped values into bins



1000 conditions

$\rightarrow$ $\boxed{\text{10 conditions}}$

1, 2, 3, 4, ...

0 — 100, 100 – 200; 200 – 300, ...

$$1 - \left[\left(\tfrac{1}{2}\right)^2 + \left(\tfrac{1}{2}\right)^2 \right]$$

$$1 - \left[0.5^2 + 0.5^2 \right]$$

$$1 - \left[0.25 + 0.25 \right]$$

$$1 - \left[0.5 \right] = 0.5$$

$f_1 \rightarrow$ 0 to 100

↳ best split → 60

$f_2 \rightarrow$ 0 to 1000

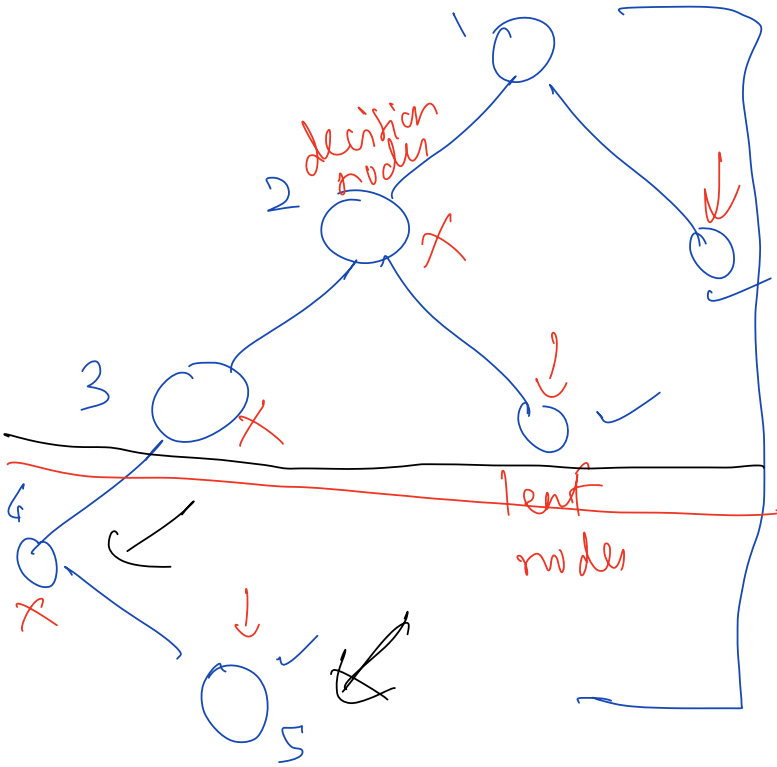
↳ best split → 700

$f_1' \rightarrow$ 0 to 1

↳ best split → 0.6

$f_2' \rightarrow$ 0 to 1

↳ best split → 0.7

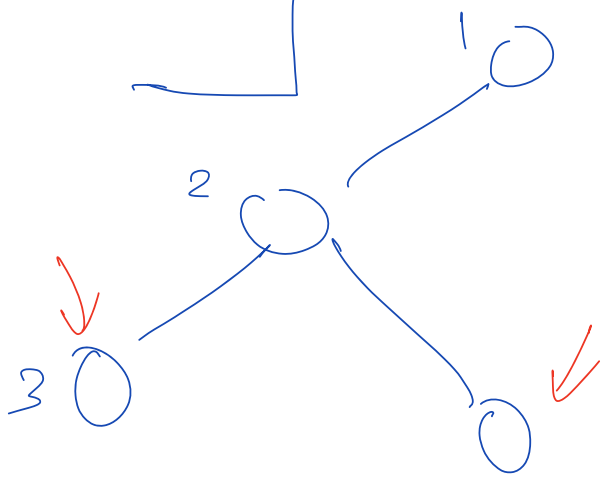


Tree 1

depth = 5

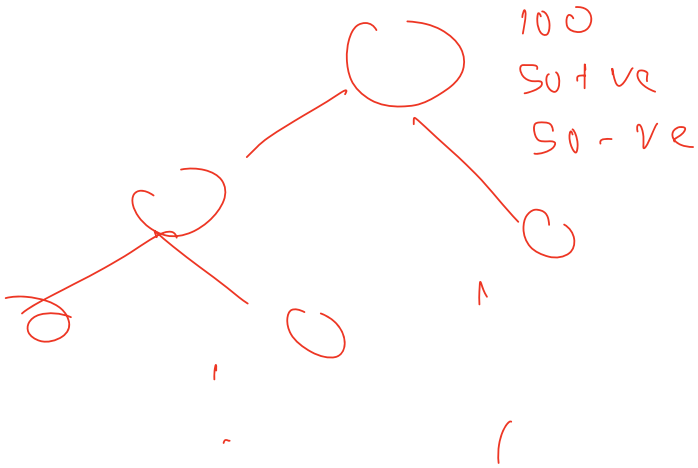
max_depth = 3

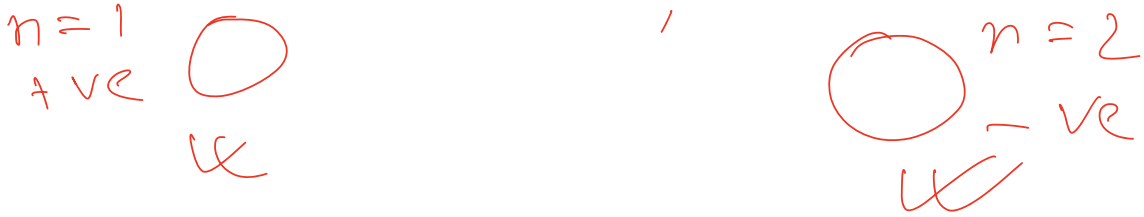
Tree 2



depth = 3

max - depth

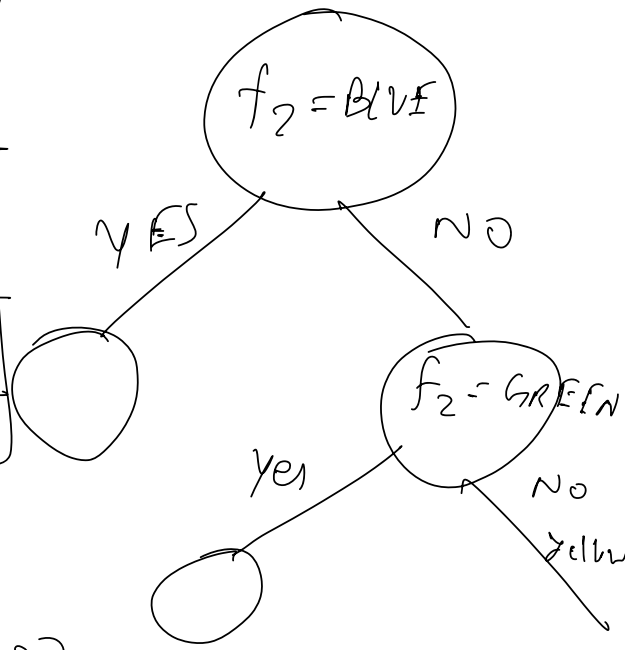




increase max_depth → overfit

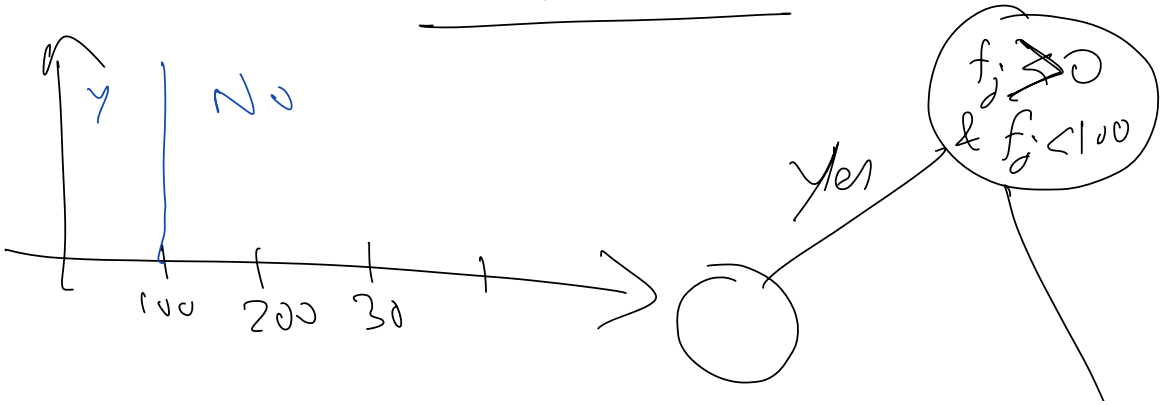
decrease " → underfit

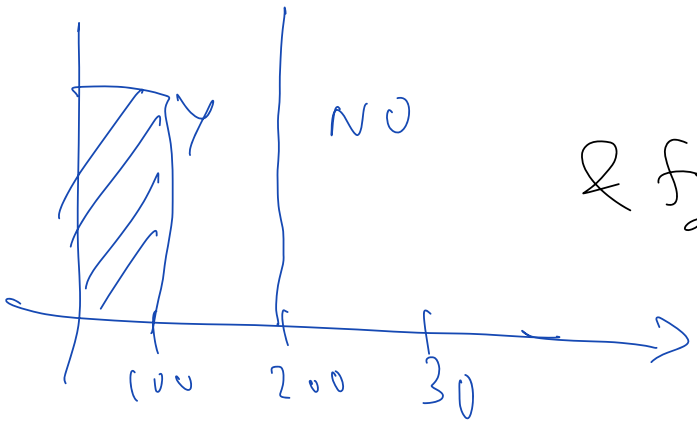
f_1	f_2
	BLUE
	YELLOW
	GREEN



Numerical :

1 to 1000





$f_j > 100$

& $f_j < 200$