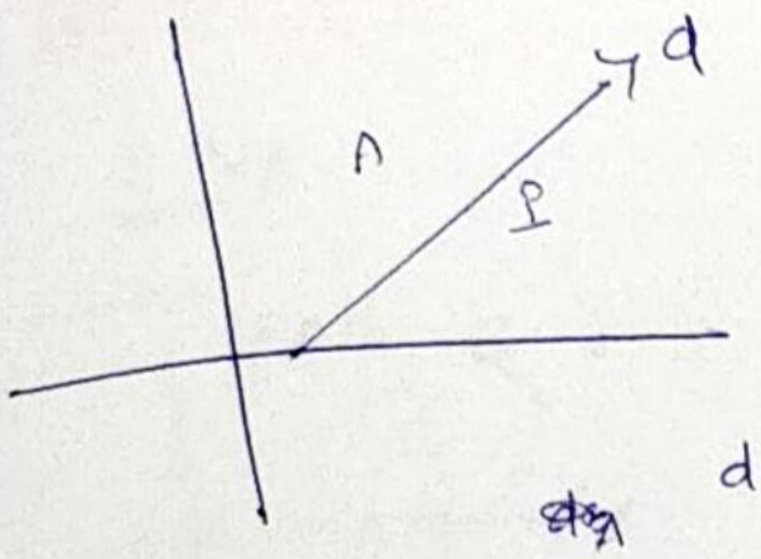
Lec-20 — Anoop

25/10/18



$d_A(x)$

$d_B(x)$

$d(x) > 0 \Rightarrow A$

$< \Rightarrow B$

2-class classifier

Linear SVM → classifier

(0) $d_A(x) > d_B(x) \Rightarrow A$

$d_B(x) > d_A(x) \Rightarrow B$

Multi class classifier using linear classifiers.

Decision Trees

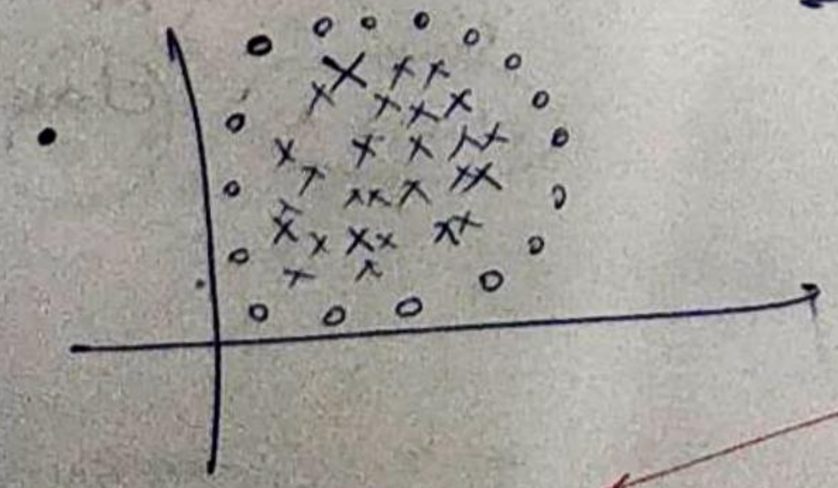
↳ A series of qns asked, fns evaluated, to determine class

eg tail:-length is a feature

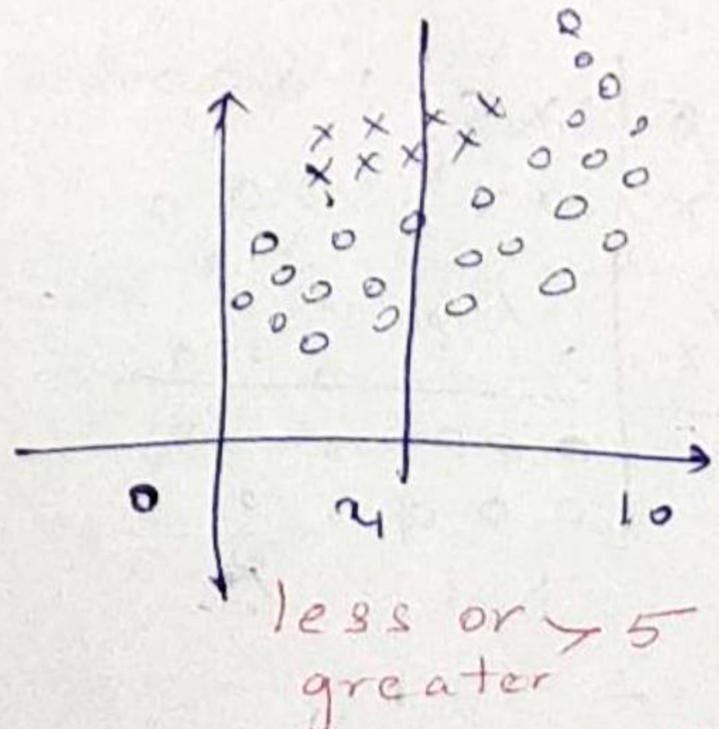
↳ tail-length > 5 ?

↳ tail ~~length~~ black ?

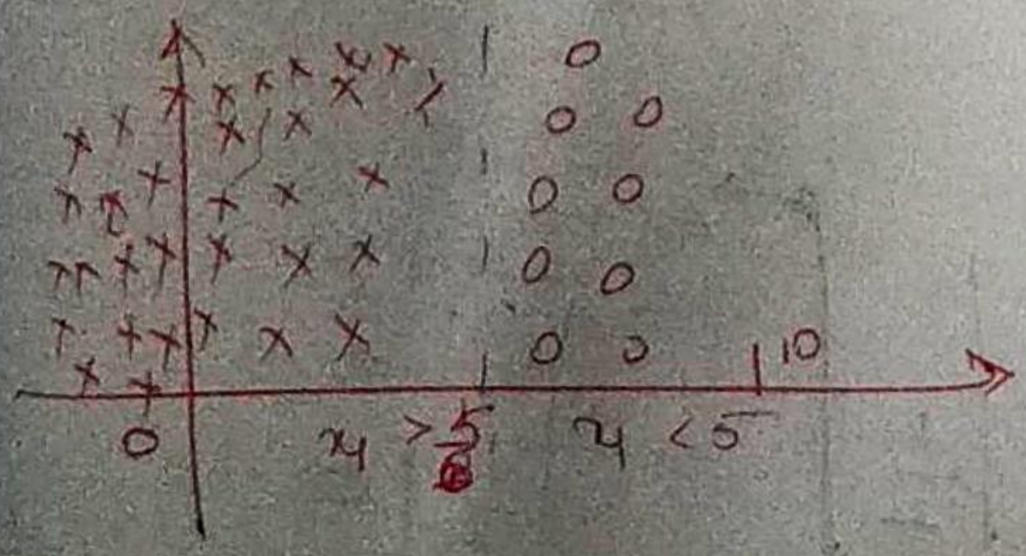
Root A dog!



The qn should reduce the number of samples to be evaluated



less or > 5 greater is not a good qn



- We want a split, so that we easily know class info

- 20 Qns → Male/Female ← will give some rdxn in space

CostaRican/not ? → If yes, huge reduction in space

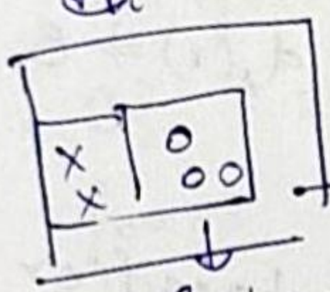
→ Else waste (most likely we get no)

- So, split so that, one class only lies on one side ! → Measure of this

Ans Entropy ↵

$H(x) = -\sum P_i \ln P_i$

Entropy ≃ Uncertainity



→ Entropy is high

Entropy is low

P_i : i is class

$P_i \ln P_i$: Take class 0

$P_1 = 0$ finding 1 in class 0

$\ln P_1 = -\infty$

$P_1 \ln P_1 = 0$

and $P_2 = 1$

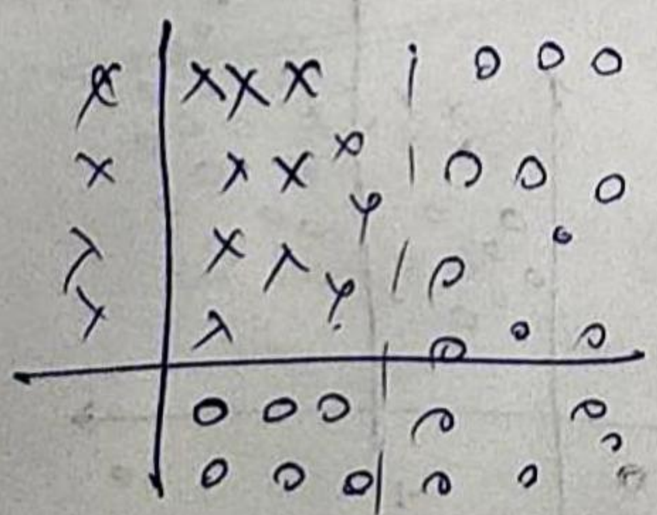
$\ln P_2 = 0$

$P_2 \ln P_2 = 0$

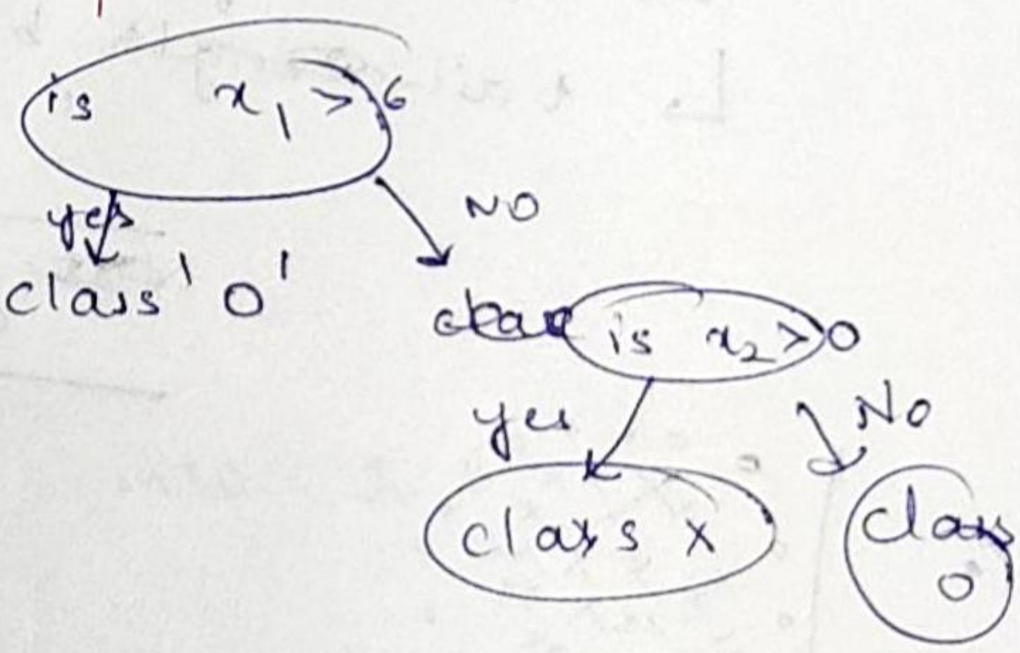
So, we want to be highly certain.

after

Level- 2



"The Tree"



Now, how do we know which qn to ask?

- Entropy → always decreases / remains same

↓

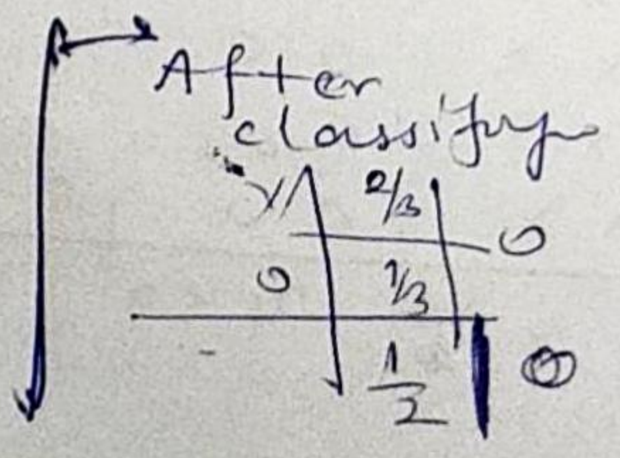
Because, uncertainity on the world never increases.

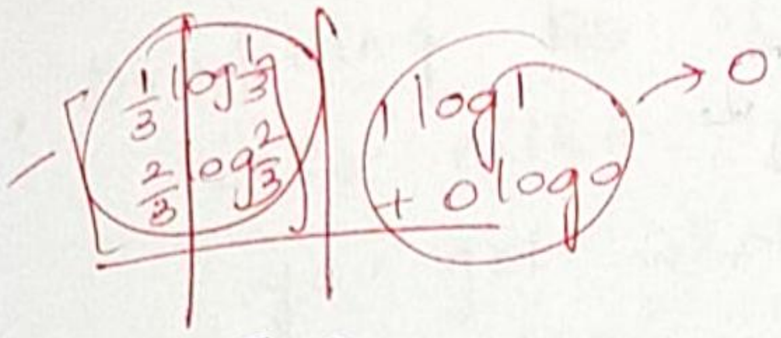
$H(x) = \frac{1}{2} P_i = \frac{1}{2}$

Entropy $= -\frac{1}{2} \log_2 \frac{1}{2}$

$= \frac{1}{2} + \frac{1}{2}$

Total = 1



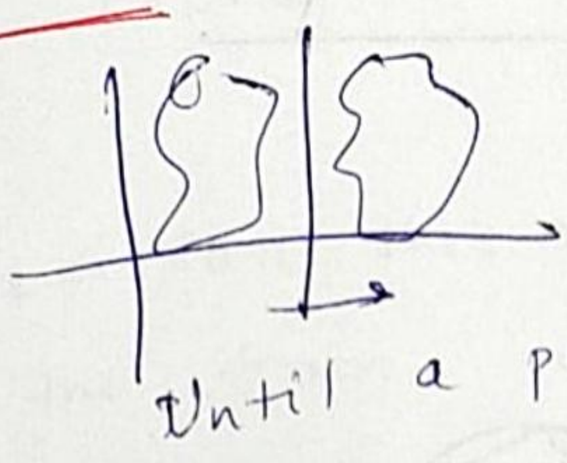


$\rightarrow 0$

$= -\frac{1}{3}\log\frac{1}{3} - \frac{2}{3}\log\frac{2}{3}$

$+ 1 \log 1 + 0 \log 0$ //

0.728



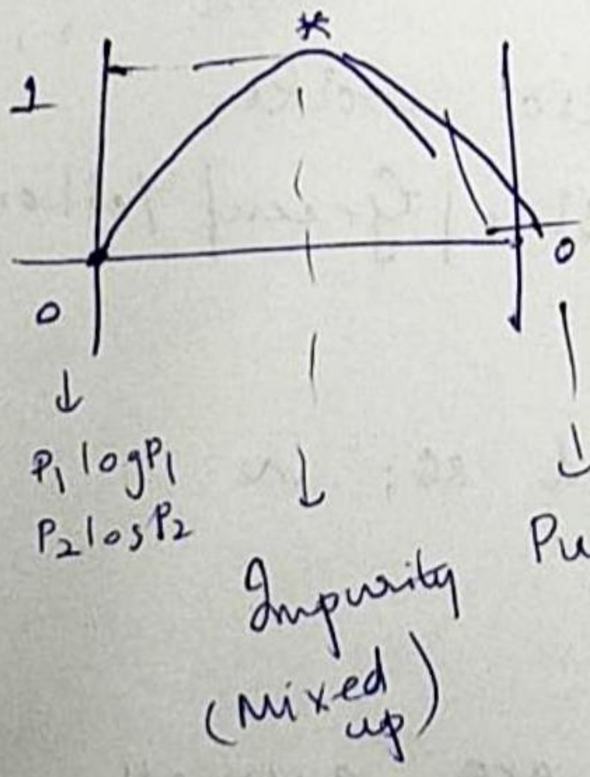
Until a point crosses over to the other side -

then,

Training a decision tree takes a lot of time, as we are bruteforcing through ~~data~~ qns.

But testing is really fast.

$$P_1 = 1 - P_2$$



- Entropy → Mixed up!
- Measure of purity

High entropy → High impurity //

So any fn that gives a fair measure of impurity & good as $\log P_i$ is time consuming.

Req $P_i \log P_i$

↓ ↓

How high is the probability, that I will mix up?

$$\sum_i \left(P_i \sum_{k \neq i} P_k \right)$$

Check → Purity → $P_1 = 1$ $P_1 \times 0 = 0$

$P_2 = \frac{1}{2}$ $\frac{1}{2} \times \frac{1}{2}$

$$= \sum_i P_i (1 - P_i)$$

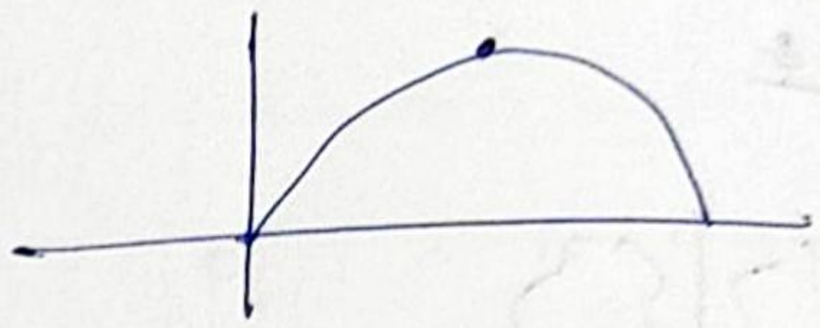
$\sum_i P_i (1-P_i)$

$\sum_i P_i - \sum_i P_i^2$

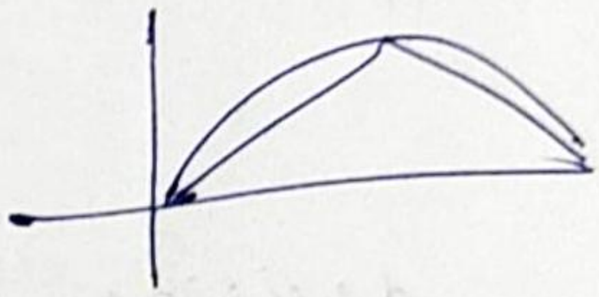
$1 - \sum_i P_i^2$ → Much simpler

$I(X)$ = Impurity of set x

An impurity metric



$1 - P_1^2 -$



Other metric:
Misclassification error

Why decis — It is like an if/else statement

Usage

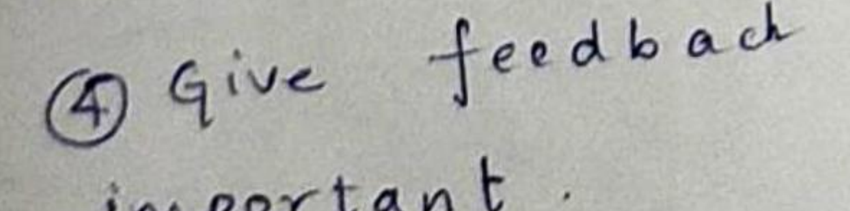
1) Decisions atree: Categorical data also works.
Male Female — Red | Blue | Green | Yellow

2) Efficient

3) It can tell, why it classified so, in english // Interpretable

4) Give feedback on which features are actually important.

Cox $\alpha_1 x_1 + \alpha_2 x_2$
↓
↳ To be learnt too
→ So more time taking



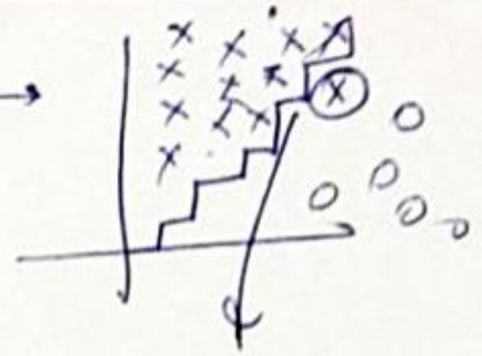
Disadv
→ Traing time >> very large
→ Categorical data only good, numerical bad
→ All splits are axis ||el
→ Can't learn curvy fns
→ Can overfit

• overfitting is possible

↳ So either cut tree after you get whole tree

(or)

↳ keep a threshold on entropy



A test sample, misclassified

• Decision tree → Missing features also solved, most of the time. until it asks for that feature

eg 20 places → what is the wt of person?

You may not know, but the other person also won't ask

• But even " not thinking of missing samples "

is bad

• Missing features treated by using

↳ Data from prev. class //