Chp 6 → Decision Trees

* Entropy

$P = [P_1, P_2 \ldots P_k]$

$p_i \geq 0$

$\Sigma p_i = 1$

$$\text{entropy}(P) = -\sum_{i=1}^{k} p_i \cdot \log_k p_i$$

↳ non negative number

0 single entry with P=1

$P_j = 1$

$[\frac{1}{k} \ldots \frac{1}{k}]$

1 evenly distribute

14 Points

$$P = \left[\frac{9}{14}, \frac{5}{14}\right]$$

↓ Yes's

↘ No's

entropy of a feature

Outlook

Sunny → how many have Yes and No

↳ 5

2 3

$$\left(\frac{2}{5}, \frac{3}{5}\right)$$

→ Entropy(D | Outlook = Sunny) = E(P1)

→ Entropy (D | Outlook = Rain) = E(P2)

→ Entropy (D | Outlook = Overcast) = E(P3)

avg entropy

$$E(\text{Outlook}) = \frac{5}{14} E(P_1) + \frac{5}{14} E[P_2] + \frac{4}{14} E(P_3)$$

all features.

E[Outlook	0.7	
E[temperature]	0.2	→ smallest → becomes first question
E[humidity]	0.3	
E[wind.]	0.6	

Hyperparameter

Entropy (P) $P = [P_1 \dots . P_u]$

↪ gini index(p)

① $$\text{gini}(p) = 1 - \sum_{i=1}^{n} p_i^2 \Big]_{0}^{1-\frac{k}{k^2}}$$ } max value } default

Whenerer we have high gini index
↓
Its a good feature

② When to stop asking question?

Good to ask not too many questions.

10k days → Large
↓
10 days → Overcast (Very Small Subset)
↓
doesn't have much value
↓
if we ask qustions about it doesn't make sense

→ Small SubDataset − Size of min dataset

→ Depth of the tree (we can restrict it)

if reach dept x → stop

*Regression

Calculate mean of numerical →

Variance/MSE → $\sum_{j=1}^{n} (y_j - \bar{y})^2/n$

$MSE(D) = 0 \mid y_j = \bar{y}$

entire Dataset (how good?)

labels

avg

labels in output - same

↳ stop asking QS

* Trees on numerical features

-100° ..... 100° | human intuition

bucketisation | automatic

num feature → k binary feature → hyperparameter or decided by algorithm

$i : X_i$ — $[0, \ldots 1]$

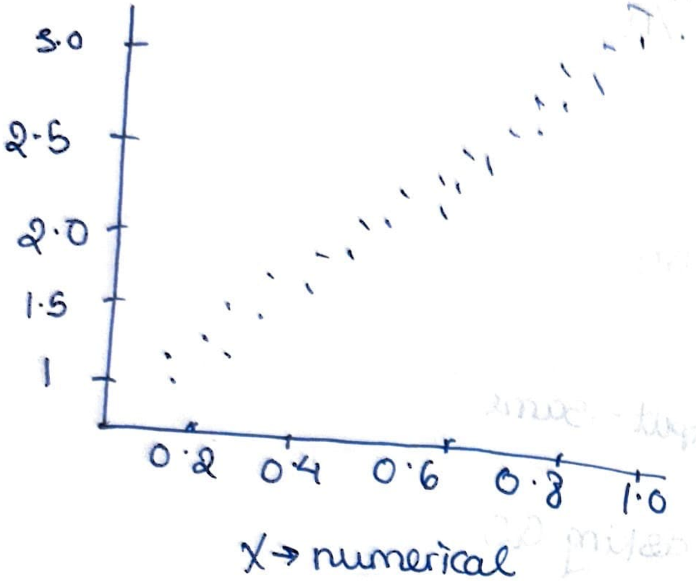
0.33   0.66

divide into sub interval

thresholds $\left\{ \begin{array}{l} (X_i > 0.33) \\ (X_i > 0.66) \end{array} \right\}$ if 0.5 which is numerical example → convert to 1

if $X_i > 0.33$ 1 → assign

Example → fully numerical.



y → numerical

X > 0.25

X > 0.5

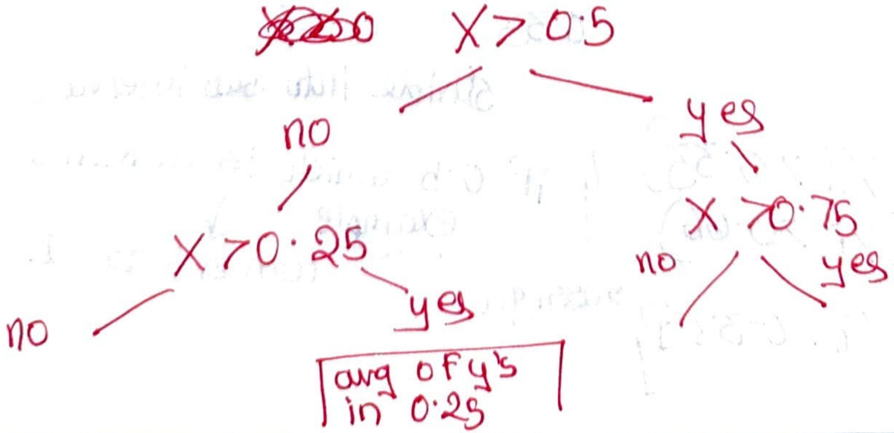
0 > 0.75

→ X>0.5 → first question

0.25 ———— High Variance

0.5 —/———— → Great → first question (not high variance)

0.75 ————/—— High Variance



Random Forests

forest (multiple decision trees)

TRAINING

subdataset

X — X'_1 $n' \times d'$

$n \times d$ \ $\vdots$

X_k

(random selected n' with replacement (able to reselect)

→ reselect rw (repetition)

↓

select d' attributes (randomly not replacemen

Train T_1 ... T_n } over datasets X_1, X_2, ..., X_k }

EVALUATION / PREDICTION

X → $T_1(x)$ ... $T_n(x)$

y_1 [0, 1, 2] ... y_k } majority which label comes more often → output

k diff labels

Early stopping questions] 10 decisio trees → 10 protocols

↓

overfitting will cancel out.