Decision Trees

For Classification

CS 307

NON PARAMETRIC CLASSIFICATION

ESTIMATING $P\left[Y = k \mid X = x\right]$ WITH
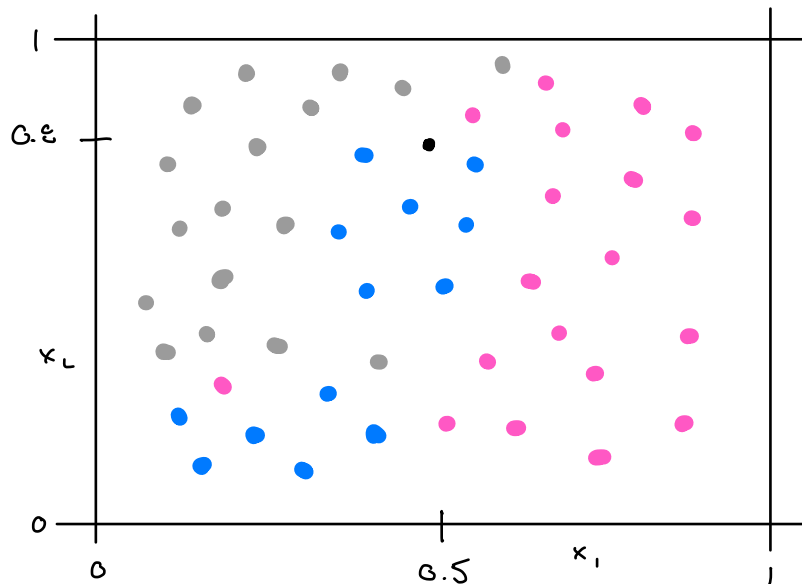
- KNN
- TREES

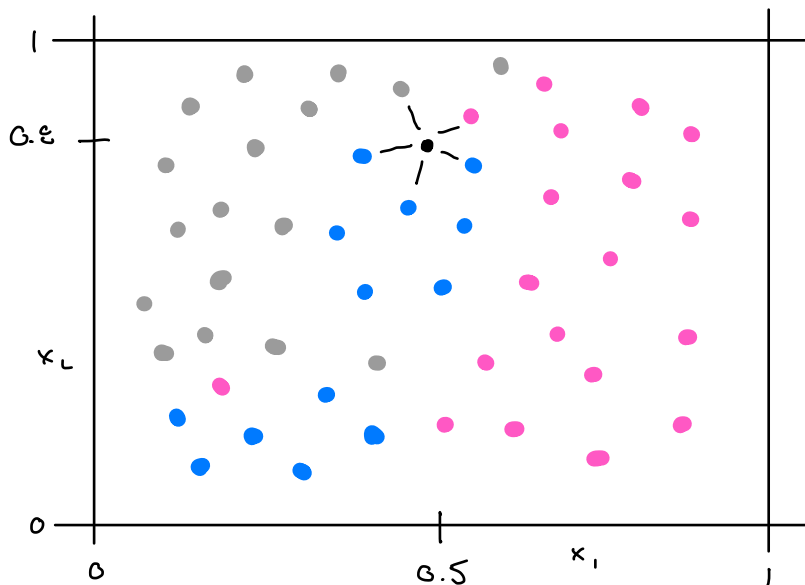PROBABILITY OF CLASS k
GIVEN THE FEATURE INFORMATION

S ETUP

y	x_1	x_2
A		
⋮		
A		
B		
⋮		
B		
C		
⋮		
C		
?	0.5	0.8

KNN

$$\hat{P}\left[Y = j \mid X = x\right] = \frac{1}{K} \sum_{\{i \,:\, x_i \in N_k(x, D)\}} I(y_i = j)$$

WITH $K = 5$, AND $x = (0.5, 0.8)$

$$\hat{P}\left[Y = A \mid X = x\right] = 3/5 \longleftarrow$$

$$\hat{P}\left[Y = B \mid X = x\right] = 1/5$$

$$\hat{P}\left[Y = C \mid X = x\right] = 1/5$$

IF BINARY → USE ODD K

↳ AVOID TIES

Decision Trees

ROOT

← YES $X_2 < 0.3$ NO →

$X_1 < 0.5$

$X_1 < 0.3$

$X_2 < 0.8$
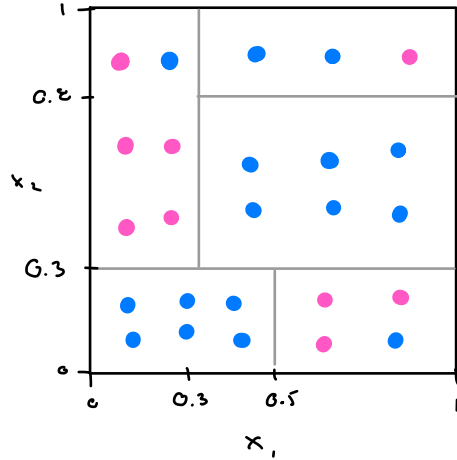
$$\hat{P}\left[Y = \bullet \mid x_1 = 0.9,\ x_2 = 0.1\right] = 1/4$$
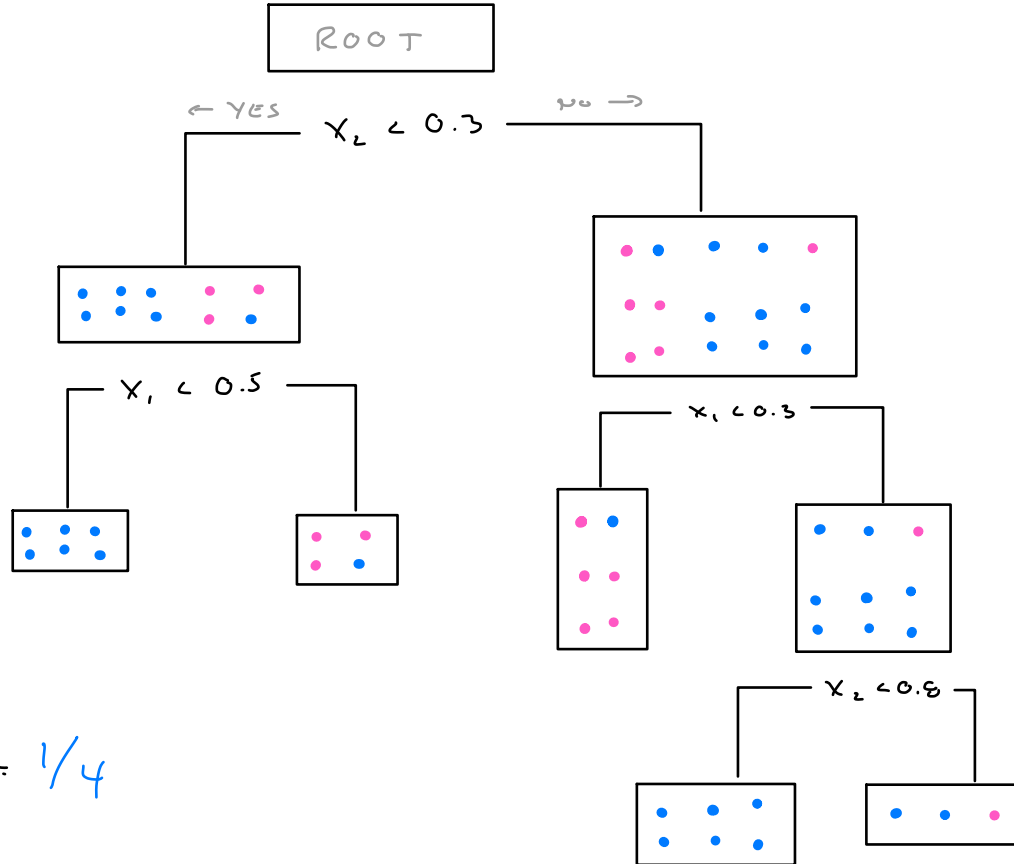
$\mathcal{N}_{ODE}$ $\mathcal{P}_{ROBABILITIES}$

$$\hat{P}_k = \frac{\sum_i I(y_i = k) I(x_i \in A)}{\sum_i I(x_i \in A)}$$

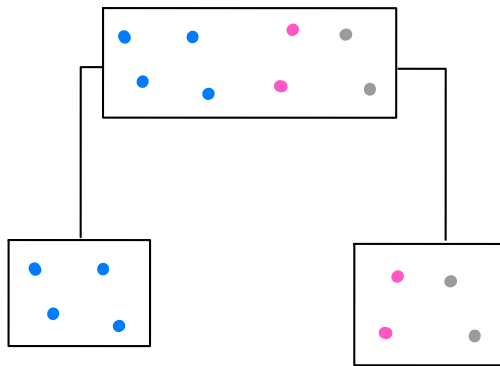$$\downarrow$$

$$\hat{P}[Y = k \mid x \in \mathcal{N}]$$

$\hat{P}_A = 4/8$

$\hat{P}_B = 2/8$

$\hat{P}_C = 2/8$

$\hat{P}_A = 4/4$ $\hat{P}_A = 0/4$

$\hat{P}_B = 0$ $\hat{P}_B = 2/4$

$\hat{P}_C = 0$ $\hat{P}_C = 2/4$

IMPURITY MEASURES FOR CATEGORICAL DATA
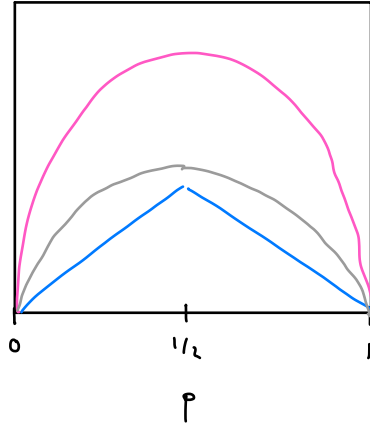
"VARIANCE"

\# CATEGORIES

$$\mathrm{G_{INI}}\left(A\right) = \sum_{k=1}^{K} \hat{p}_k \left(1 - \hat{p}_k\right) = 1 - \sum_{k=1}^{K} \hat{p}_k^{\,2}$$

$$\mathrm{E_{NTROPY}}\left(A\right) = - \sum_{k=1}^{K} \hat{p}_k \log\left(\hat{p}_k\right)$$

$$\mathrm{E_{RROR}}\left(A\right) = 1 - \max_k \left(\hat{p}_k\right)$$

CALCULATING GINI

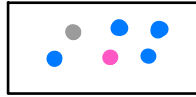$$\text{GINI}(A) = \sum_{k=1}^{K} \hat{p}_k (1 - \hat{p}_k) = 1 - \sum_{k=1}^{K} \hat{p}_k^2$$

A :
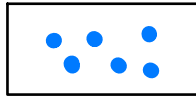
$\hat{p}_A = 6/6$

$\hat{p}_B = 0/6$
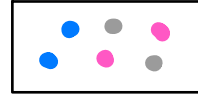
$\hat{p}_C = 0/6$

$\hat{p}_A = 4/6$

$\hat{p}_B = 1/6$

$\hat{p}_C = 1/6$

$\hat{p}_A = 2/6$

$\hat{p}_B = 2/6$

$\hat{p}_C = 2/6$

GINI(A)

0

0.5

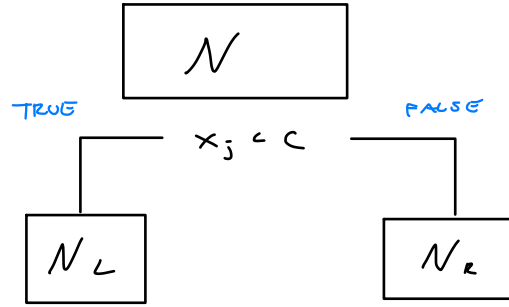$0.6\overline{6}$

$$\hookrightarrow = 1 - \left[(4/6)^2 + (1/6)^2 + (1/6)^2 \right]$$

SPLITTING

$$\text{FIND} \begin{cases} \text{FEATURE } x_j \\ \text{CUTOFF } c \end{cases}$$

TRUE   FALSE

$$x_j < c$$

N

N_L N_R

$$\min_{j,c} \left[\frac{|N_L|}{|N|} \, \text{GINI}\left(N_L\right) + \frac{|N_R|}{|N|} \, \text{GINI}\left(N_R\right) \right]$$

(WEIGHT)

VARIANCES

WHICH SPLIT ?

x_2

x_1

0.44

x_2
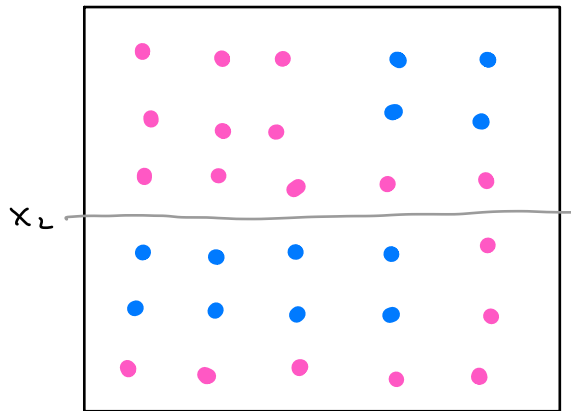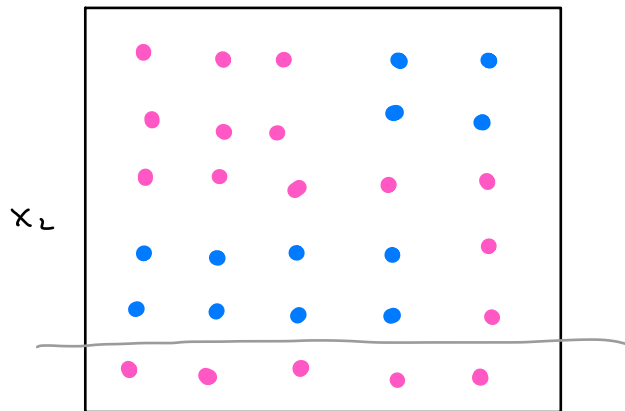
x_1

0.416

TREES ARE INDIFFERENT TO SCALING

X_1

1000   2000   3000

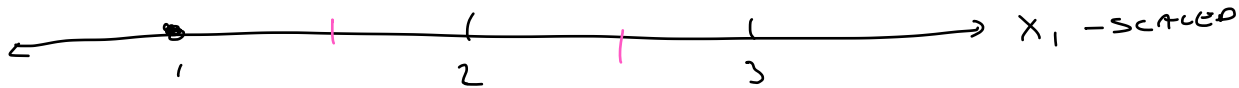X_1 - SCALED

1   2   3
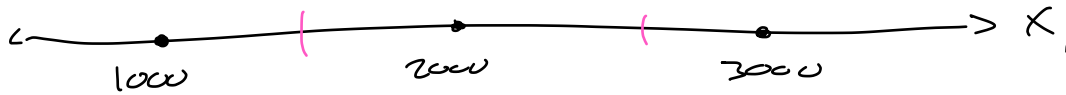
CUTOFFS WILL RESULT IN SAME
OBSERVATIONS IN NEWLY SPLIT NODES