

Decision Trees

For Classification

CS 307

NON PARAMETRIC CLASSIFICATION

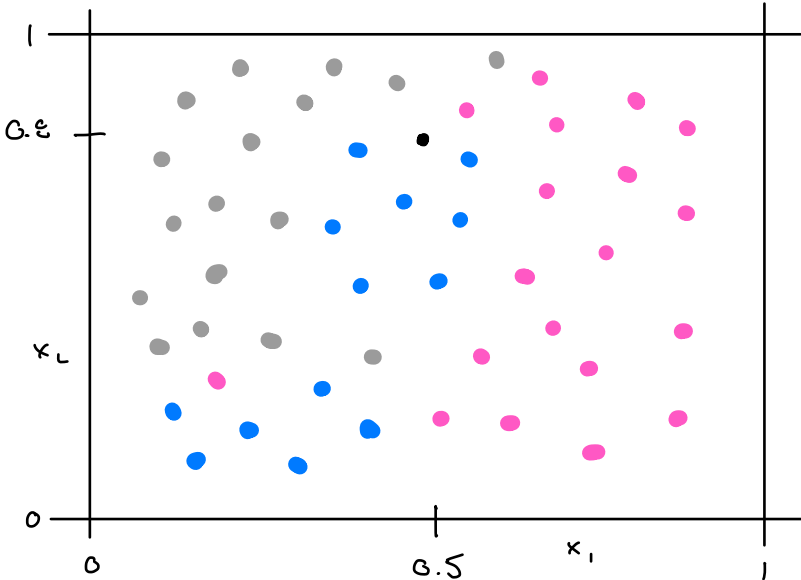
ESTIMATING $P[Y = k | X = x]$ WITH

- KNN
- TREES

PROBABILITY OF CLASS k
GIVEN THE FEATURE INFORMATION

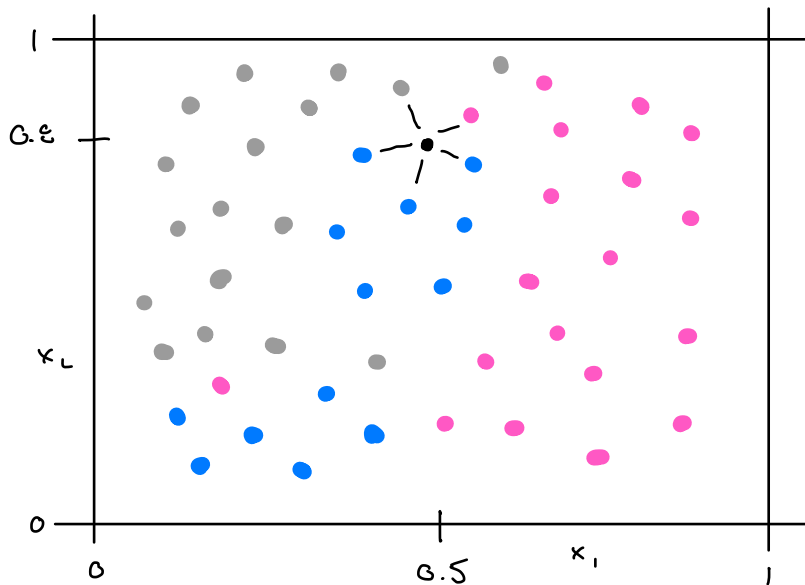
SETUP

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



KNN

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



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

$$\hat{P}[Y=A | X=x] = 3/5 \leftarrow$$

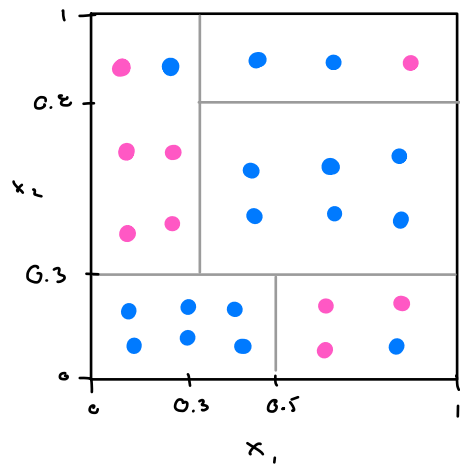
$$\hat{P}[Y=B | X=x] = 1/5$$

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

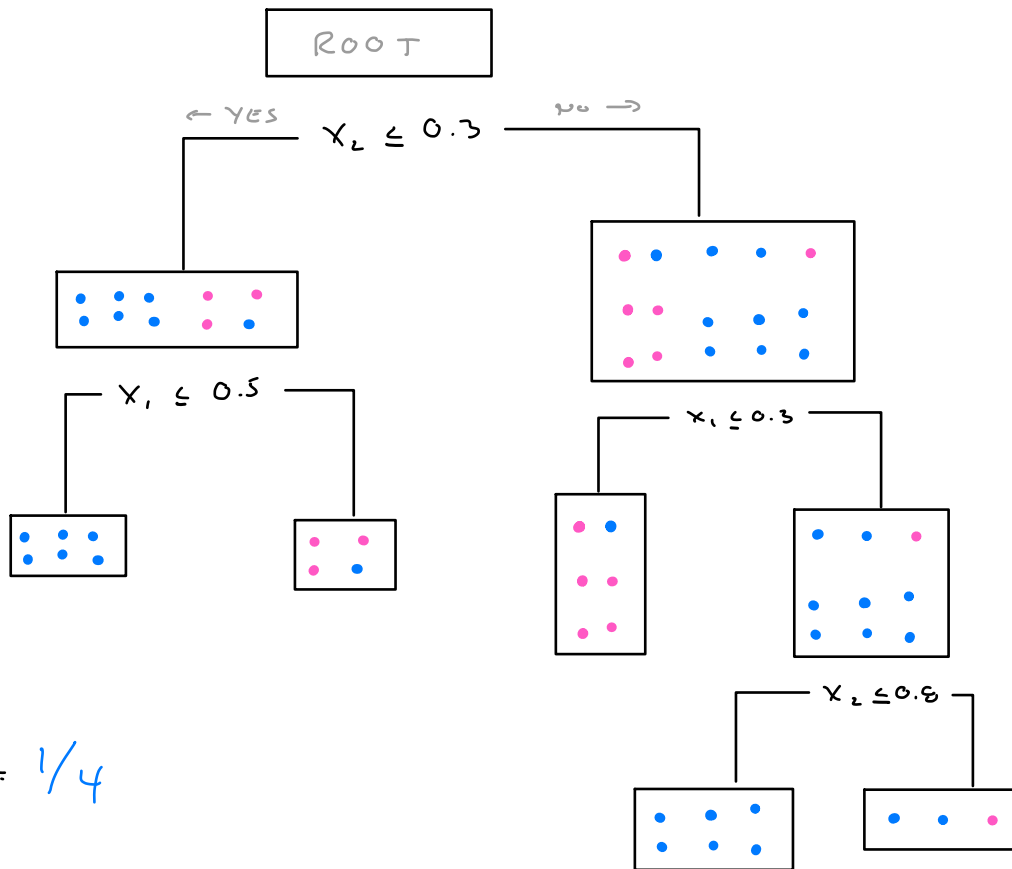
IF BINARY $\rightarrow$ USE ODD K

$\hookrightarrow$ AVOID TIES

DECISION TREES



$$\hat{P}[Y = \bullet \mid x_1 = 0.9, x_2 = 0.1] = 1/4$$



Node Probabilities

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

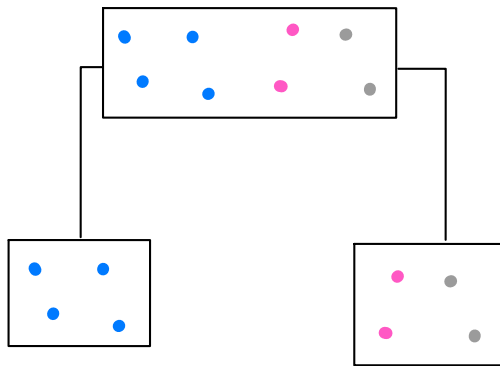


$$\hat{p}[Y = k | X \in \mathcal{N}]$$

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

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

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



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

$$\hat{p}_B = 0$$

$$\hat{p}_C = 0$$

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

$$\hat{p}_B = 2/4$$

$$\hat{p}_C = 2/4$$

IMPURITY MEASURES FOR CATEGORICAL DATA

"ERROR" $\swarrow$

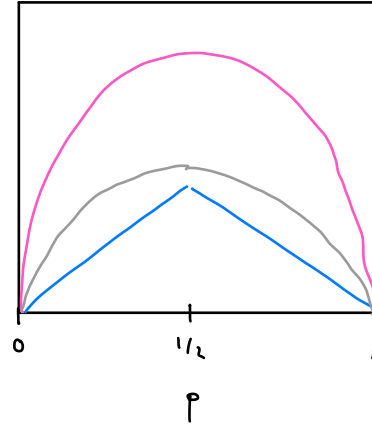
"VARIANCE" $\swarrow$

$\nwarrow$ # CATEGORIES

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

$$\underline{ENTROPY}(A) = - \sum_{k=1}^K \hat{p}_k \log(\hat{p}_k)$$

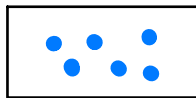
$$\underline{ERROR}(A) = 1 - \max_k (\hat{p}_k)$$



CALCULATING GINI

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

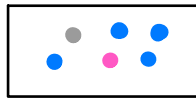
A:



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

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

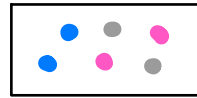
$$\hat{p}_C = 0/6$$



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

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

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



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

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

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

$$G_{INI}(A)$$

$$0$$

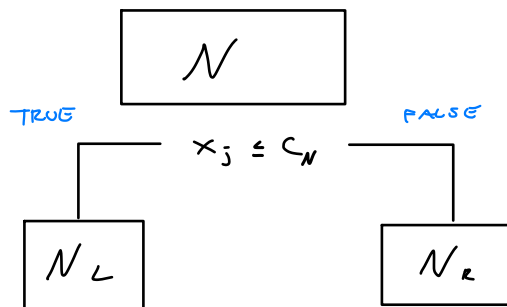
$$0.5$$

$$0.6\bar{6}$$

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

SPLITTING

FIND $\rightarrow$ FEATURE x_j
 $\rightarrow$ CUTOFF c_N



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

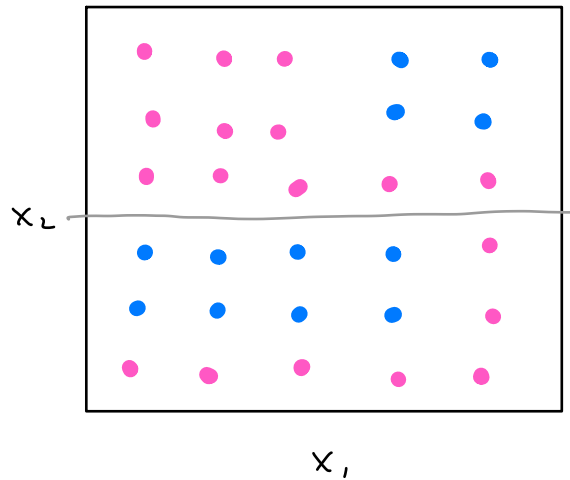
Diagram illustrating the Gini index formula for splitting a dataset N into two subsets N_L and N_R . The formula is:

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

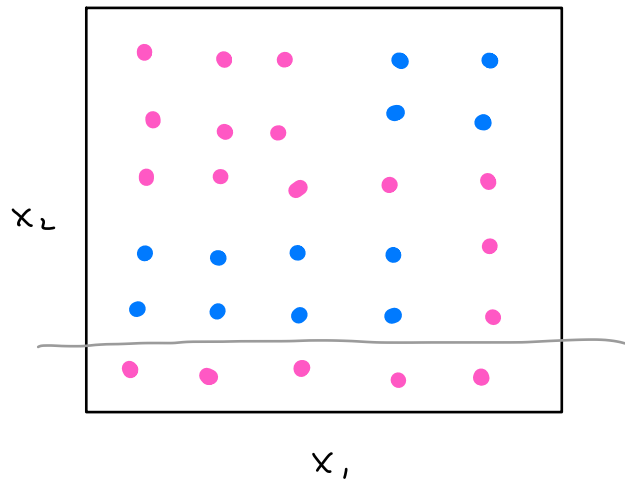
The diagram includes annotations:

- WEIGHTS** (pink arrow) points to the fractions $\frac{|N_L|}{|N|}$ and $\frac{|N_R|}{|N|}$.
- VARIANCES** (blue arrow) points to the $\text{GINI}(N_L)$ and $\text{GINI}(N_R)$ terms.

Which Split?



0.44



0.416