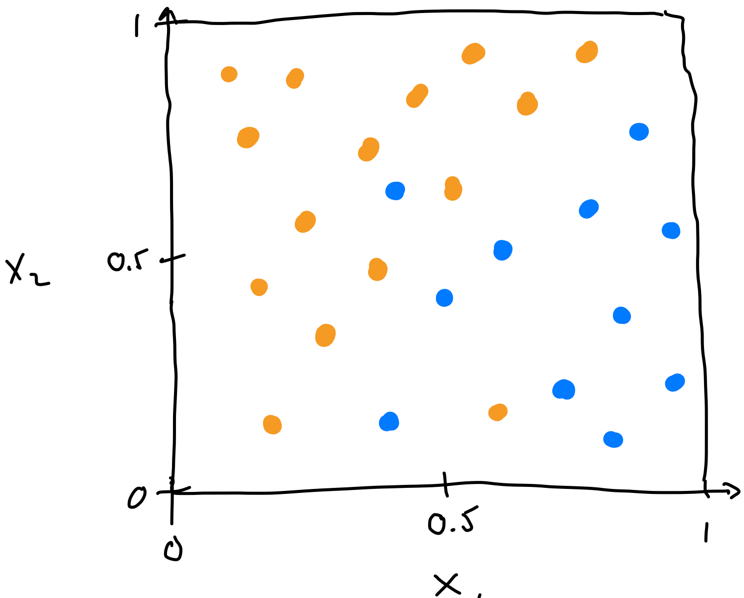
NONPARAMETRIC CLASSIFICATION

ESTIMATING $P[Y=k \mid X=x]$ WITH

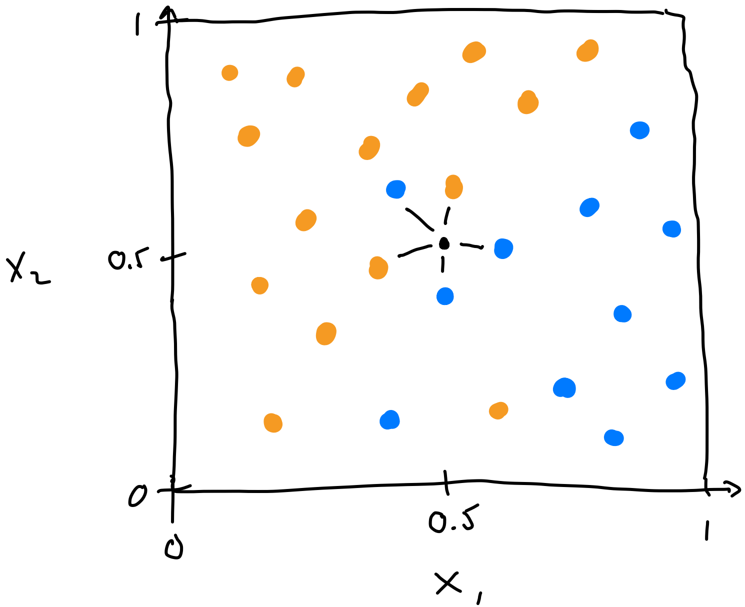
- KNN
- DECISION TREES

Setup



y	x_1	x_2
A		
A		
A		
A		
B		
B		
B		
B		
?	0.5	0.5

KNN



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

$$\text{KNN}=5 \begin{cases} \hat{P}\left[Y = \bullet \mid x_1 = 0.5,\ x_2 = 0.5\right] = 2/5 \\ \hat{P}\left[Y = \bullet \mid x_1 = 0.5,\ x_2 = 0.5\right] = 3/5 \end{cases}$$

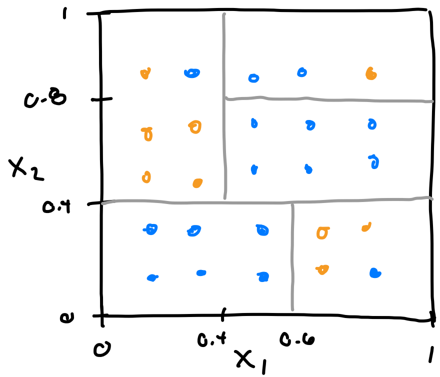
Distance?

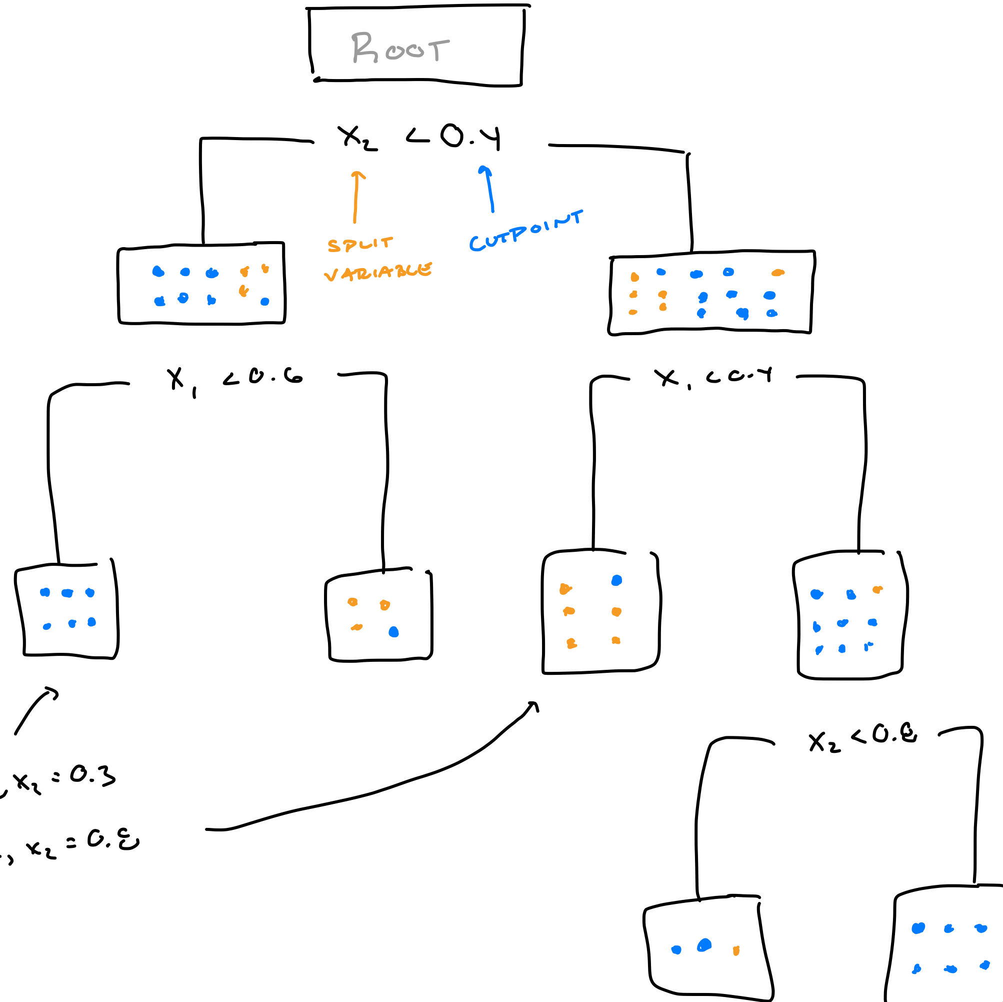
- Euclidean!
- whatever!

Binary? → use odd k

↳ Avoid ties

Decision Trees





- MIN SPLIT = 8
- $C_p = 0$

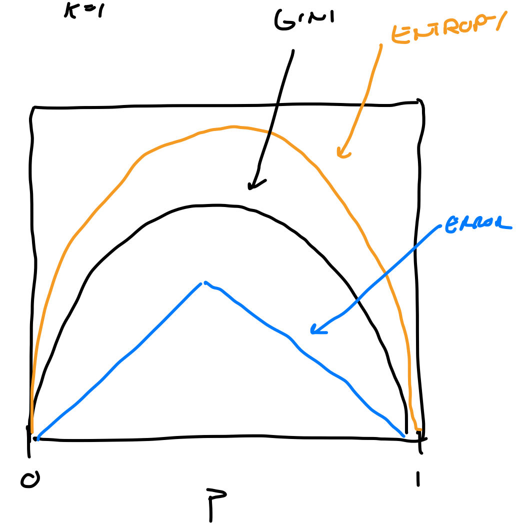
- $X_1 = 0.3, X_2 = 0.3$
- $X_1 = 0.2, X_2 = 0.8$

VARIANCE (IMPURITY) MEASURES IN CATEGORICAL DATA

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

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

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

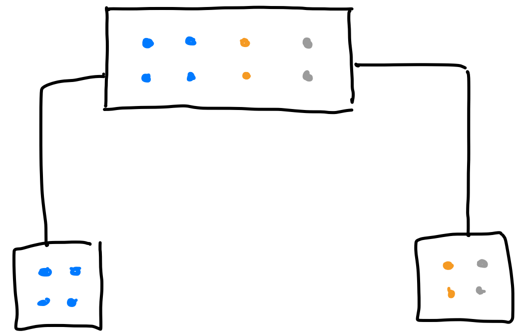


$$\hat{p}_k = \frac{\sum_i I(y_i = k)\, I(x_i \in A)}{\sum_i I(x_i \in A)}$$

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

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

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



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

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

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

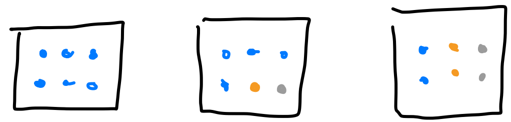
$\hat{P}_A = 0/4$

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

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

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

A:



$\hat{p}_A = 6/6$	$\hat{p}_A = 4/6$	$\hat{p}_A = 2/6$
$\hat{p}_B = 0/6$	$\hat{p}_B = 1/6$	$\hat{p}_B = 2/6$
$\hat{p}_C = 0/6$	$\hat{p}_C = 1/6$	$\hat{p}_C = 2/6$

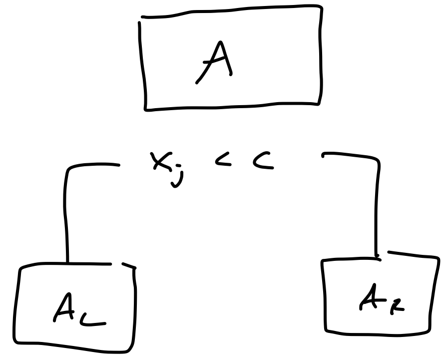
Gini(A) $\quad 0 \quad 0.5 \quad 0.66\overline{6}$

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

SPLITTING

FIND → VARIABLE x_j

FIND → CUTOFF c

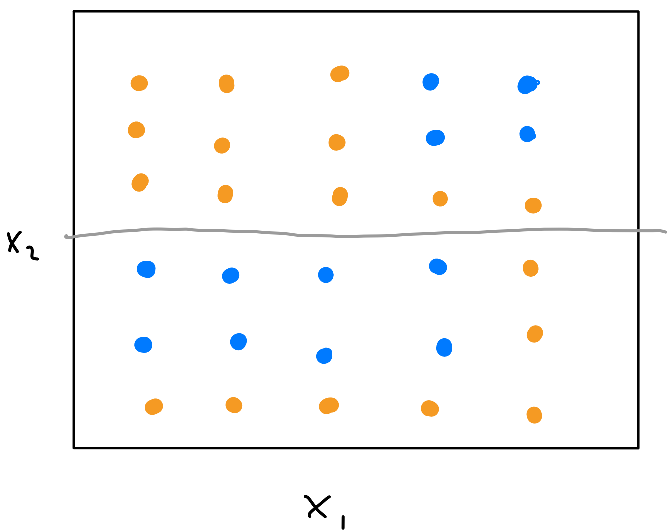


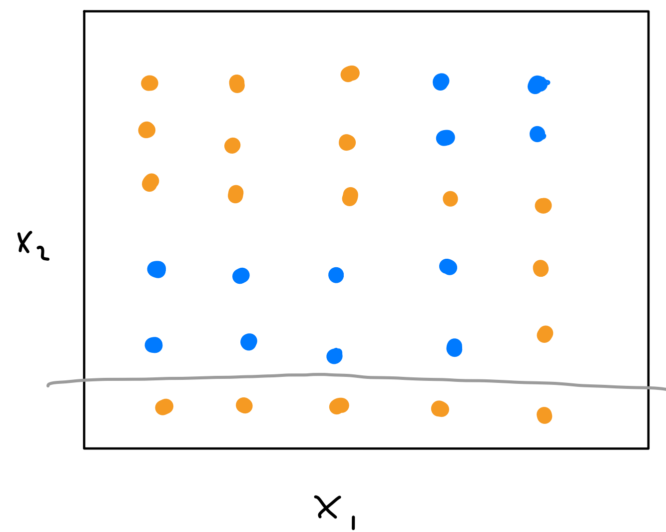
$$\min_{j,c} \left[\frac{|A_L|}{|A|} \, \mathrm{GINI}(A_L) + \frac{|A_R|}{|A|} \, \mathrm{GINI}(A_R) \right]$$

WEIGHTS

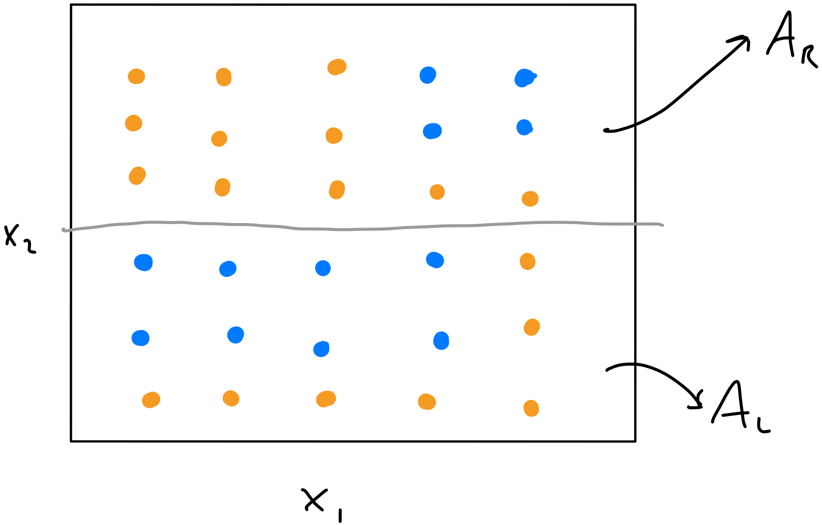
"VARIANCE"

WHICH SPLIT?





SMALLER GINI



$\hat{P}_A = 11/15$

$\hat{P}_B = 4/15$

$|A_R| = 15$

$\hat{P}_A = 7/15$

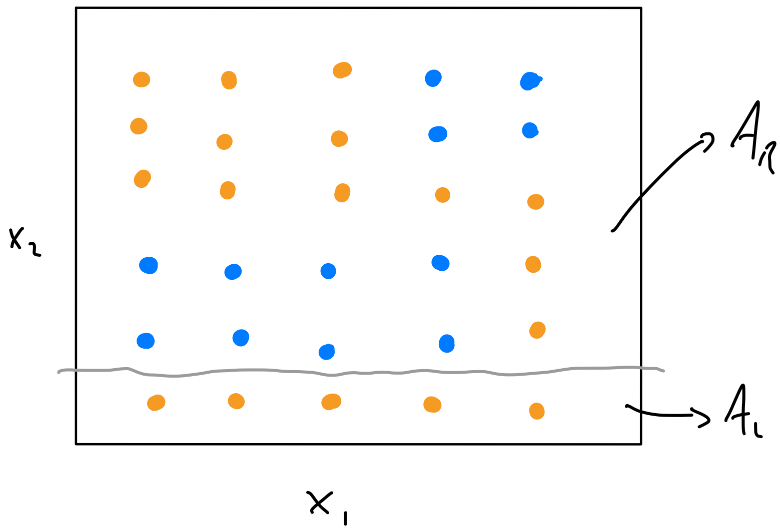
$\hat{P}_B = 8/15$

$|A_L| = 15$

$$\text{GINI}(A_R) = 1 - \left[\left(\frac{11}{15}\right)^2 + \left(\frac{4}{15}\right)^2\right] = 88/225$$

$$\text{GINI}(A_L) = 1 - \left[(7/15)^2 + (8/15)^2\right] = 112/225$$

$$\frac{|A_L|}{|A|}\text{GINI}(A_L) + \frac{|A_R|}{|A|}\text{GINI}(A_R) = \frac{15}{30}\left(\frac{112}{225}\right) + \frac{15}{30}\left(\frac{88}{225}\right) = 0.4\overline{4}$$



$\hat{p}_A = 13/25$

$\hat{p}_B = 12/25$

$|A_R| = 25$

$\hat{p}_A = 5/5$

$\hat{p}_B = 0/5$

$|A_L| = 5$

$$\text{GINI}(A_R) = 1 - \left[\left(\tfrac{13}{25}\right)^2 + \left(\tfrac{12}{25}\right)^2\right] = 312/625$$

$$\text{GINI}(A_L) = 1 - \left[\left(5/5\right)^2 + \left(0/5\right)^2\right] = 0$$

$$\frac{|A_L|}{|A|}\,\text{GINI}(A_L) + \frac{|A_R|}{|A|}\,\text{GINI}(A_R) = \frac{5}{30}\left(0\right) + \frac{25}{30}\left(\frac{312}{625}\right) = 0.416$$

Tree Questions

Categorical variables?

Missing data?

Stopping rules?

Pruning? Complexity?