

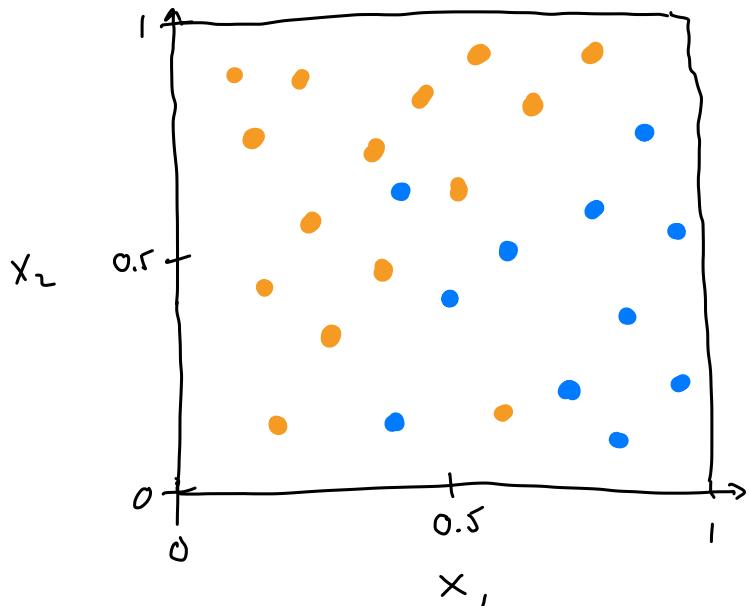
NONPARAMETRIC

CLASSIFICATION

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

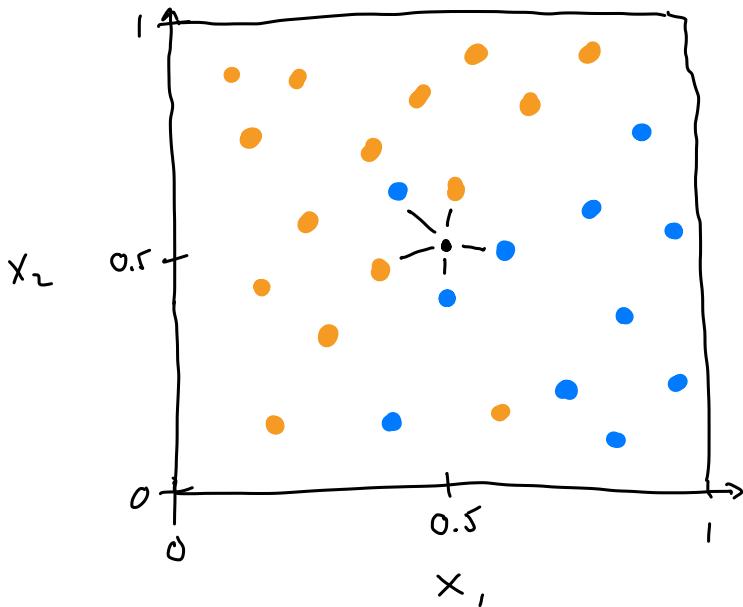
- KNN
- DECISION TREES

SETUP



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

KNN



DI STANCE ?

- EUCLIDEAN !
- WHATEVER !

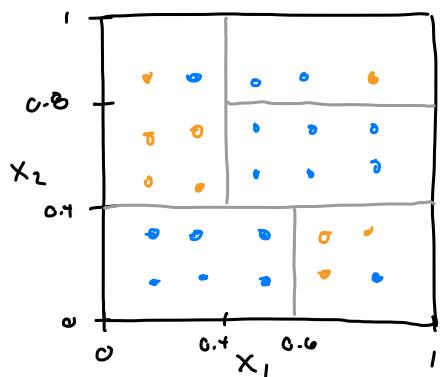
$$\hat{P}[x = j \mid X = x] = \frac{1}{k} \sum_{i \in N_k(x, D)} I(y_i = j)$$

KNN=5

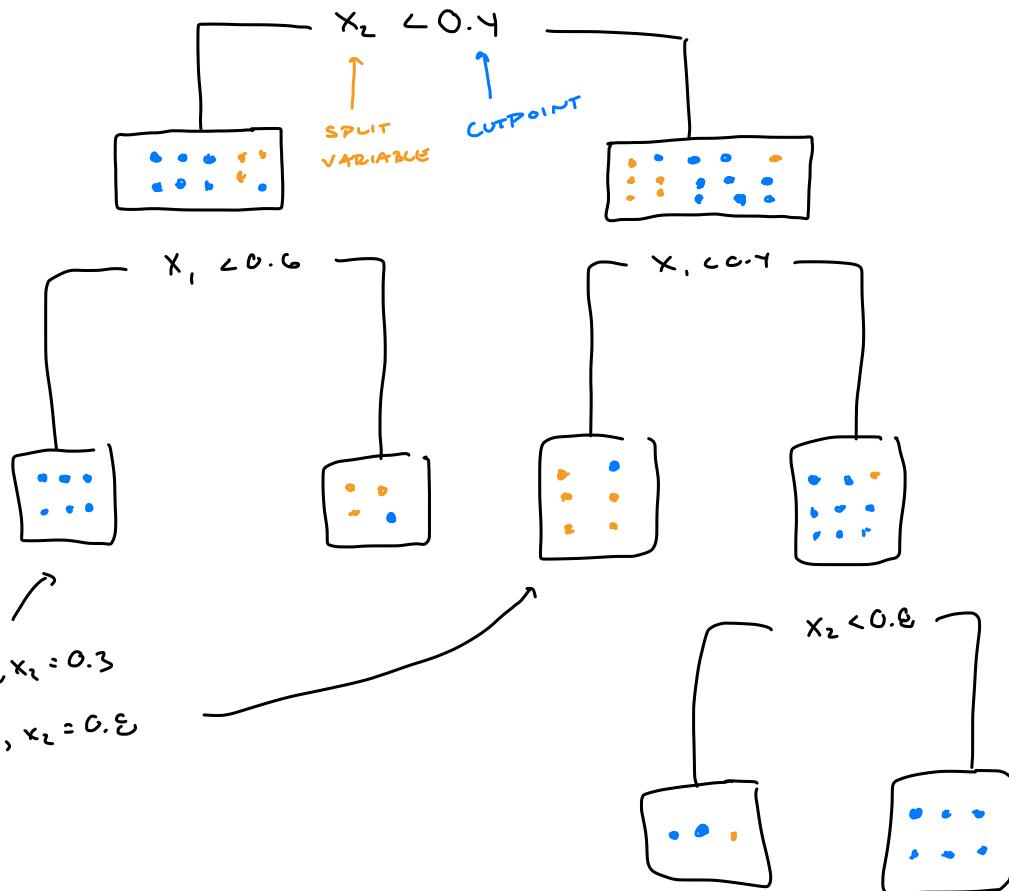
$$\begin{cases} \hat{P}[y = \text{orange} \mid x_1 = 0.5, x_2 = 0.5] = 2/5 \\ \hat{P}[y = \text{blue} \mid x_1 = 0.5, x_2 = 0.5] = 3/5 \end{cases}$$

BINARY ? \rightarrow USE ODD K
 ↳ AVOID TIES

DECISION TREES

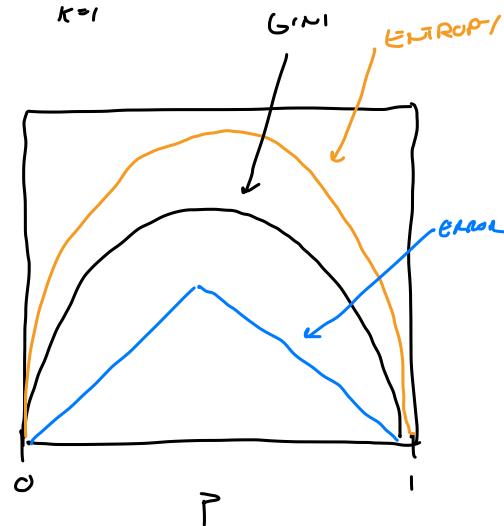


ROOT



VARIANCE (IM PURITY) MEASURES IN CATEGORICAL DATA

- • $G_{INI}(A) = \sum_{k=1}^K \hat{p}_k (1 - \hat{p}_k) = 1 - \sum_{k=1}^K p_k^2$
- $ENTROPY(A) = - \sum_{k=1}^K \hat{p}_k \log(\hat{p}_k)$
- $ERROR(A) = 1 - \max_k (\hat{p}_k)$

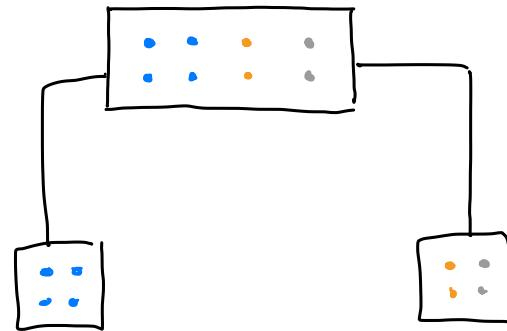


$$\hat{P}_A = 4/8$$

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

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

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



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

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

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

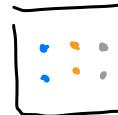
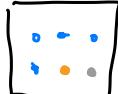
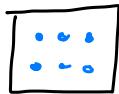
$$\hat{P}_D = 2/4$$

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

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

$$G_{IN}(A) = \sum_{k=1}^K \hat{P}_k (1 - \hat{P}_k) = 1 - \sum_{k=1}^K P_k^2$$

A:



$$\hat{P}_1 = 6/9$$

$$\hat{P}_4 = 4/9$$

$$\hat{P}_5 = 2/9$$

$$\hat{P}_3 = 0/9$$

$$\hat{P}_6 = 1/9$$

$$\hat{P}_7 = 2/9$$

$$\hat{P}_c = 0/9$$

$$\hat{P}_8 = 1/9$$

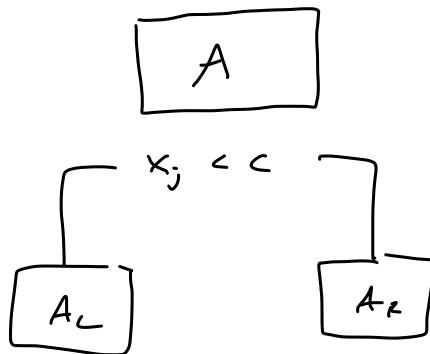
$$\hat{P}_9 = 2/9$$

$$G_{IN}(A) \quad 0 \quad 0.5 \quad 0.666\bar{6}$$

$$= 1 - \left[\left(\frac{4}{9} \right)^2 + \left(\frac{1}{9} \right)^2 + \left(\frac{1}{9} \right)^2 \right]$$

SPLITTING

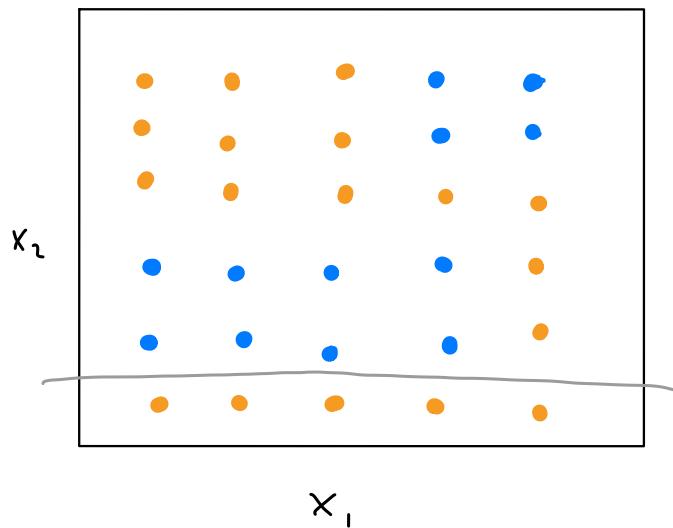
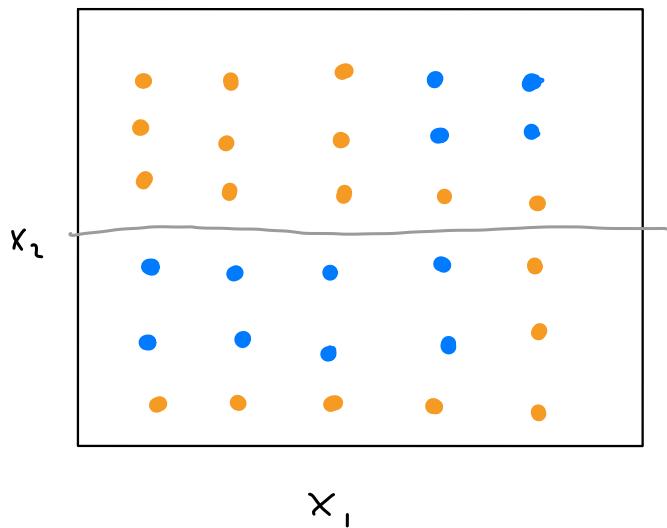
FIND
→ VARIABLE x_j
→ CUTOFF c



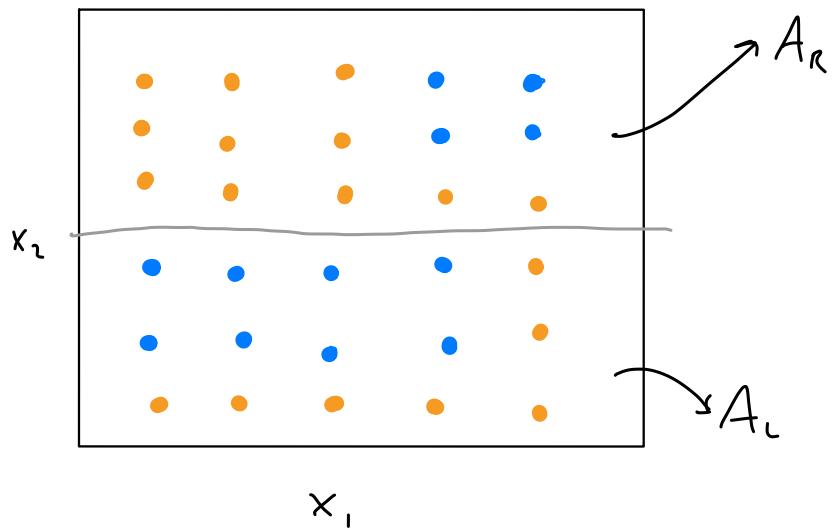
$$\min_{j,c} \left[\frac{|A_L|}{|A|} G_{\text{INE}}(A_L) + \frac{|A_R|}{|A|} G_{\text{INE}}(A_R) \right]$$

The equation shows the formula for finding the best variable j and cutoff c . It consists of a weighted sum of two terms. The first term is $\frac{|A_L|}{|A|} G_{\text{INE}}(A_L)$ and the second term is $\frac{|A_R|}{|A|} G_{\text{INE}}(A_R)$. The word "WEIGHTS" is written in blue below the fraction lines, and the word "VARIANCE" is written in orange below the terms $G_{\text{INE}}(A_L)$ and $G_{\text{INE}}(A_R)$.

WHICH SPLIT?



SMALLER GINI



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

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

$$|A_R| = 15$$

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

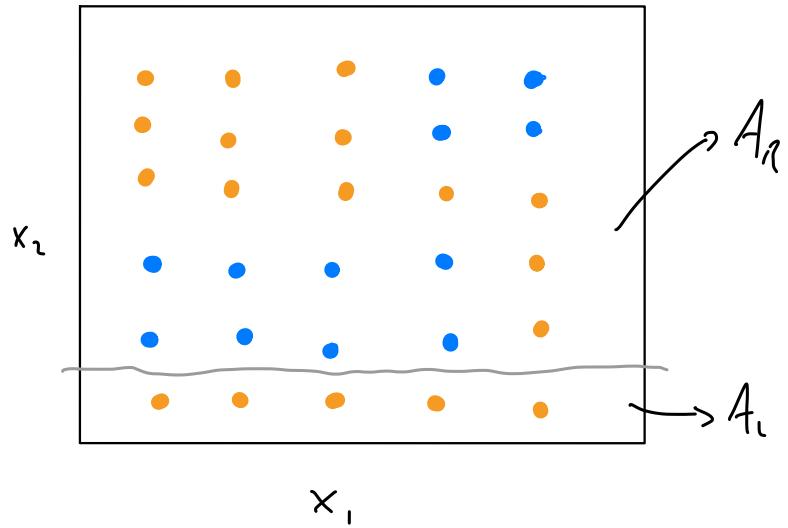
$$\hat{P}_B = 8/15$$

$$|A_L| = 15$$

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

$$G_{INI}(A_L) = 1 - \left[\left(\frac{7}{15}\right)^2 + \left(\frac{8}{15}\right)^2 \right] = \frac{112}{225}$$

$$\frac{|A_L|}{|A|} G_{INI}(A_L) + \frac{|A_R|}{|A|} G_{INI}(A_R) = \frac{15}{30} \left(\frac{112}{225} \right) + \frac{15}{30} \left(\frac{88}{225} \right) = 0.44$$



$$\hat{P}_A = \frac{13}{25}$$

$$\hat{P}_B = \frac{12}{25}$$

$$|A_R| = 25$$

$$\hat{P}_A = \frac{5}{5}$$

$$\hat{P}_B = \frac{0}{5}$$

$$|A_L| = 5$$

$$G_{INI}(A_R) = 1 - \left[\left(\frac{13}{25} \right)^2 + \left(\frac{12}{25} \right)^2 \right] = \frac{312}{625}$$

$$G_{INI}(A_L) = 1 - \left[\left(\frac{5}{5} \right)^2 + \left(\frac{0}{5} \right)^2 \right] = 0$$

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

Tree Questions

CATEGORICAL VARIABLES ?

MISSING DATA ?

STOPPING RULES ?

PRUNING ? COMPLEXITY ?