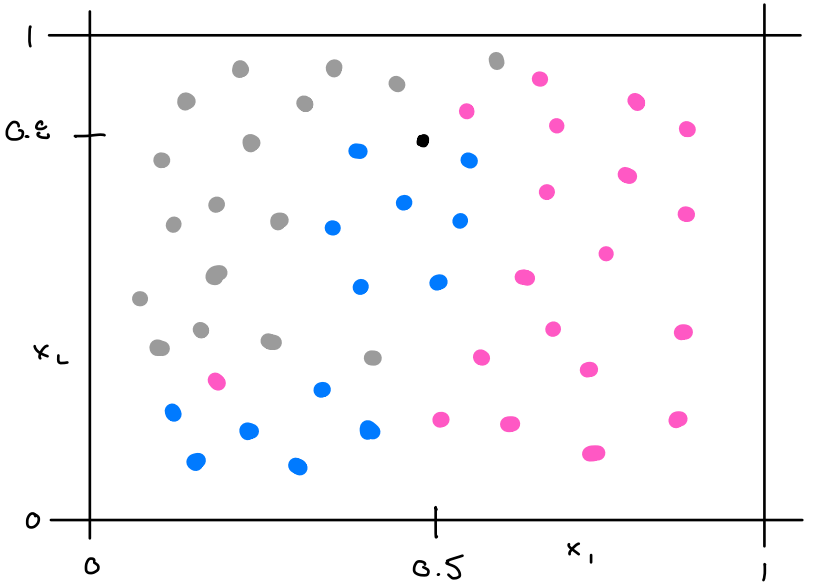
Non Parametric Classification

Estimating $P[Y = k \mid X = x]$ with

- KNN
- TREES

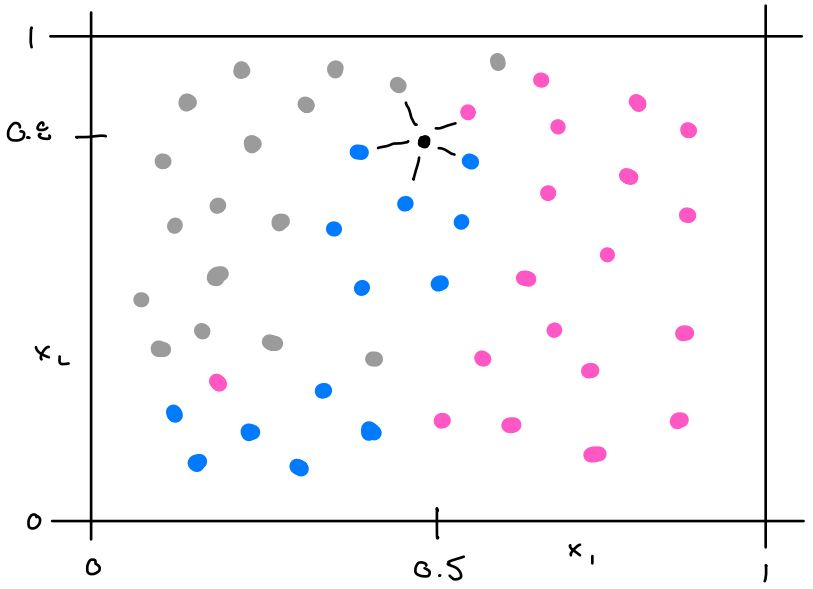
Setup

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



KNN

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



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

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

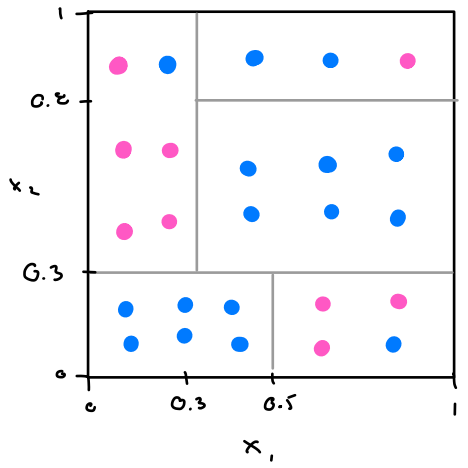
$$\hat{P}[Y = B \mid X = x] = 1/5$$

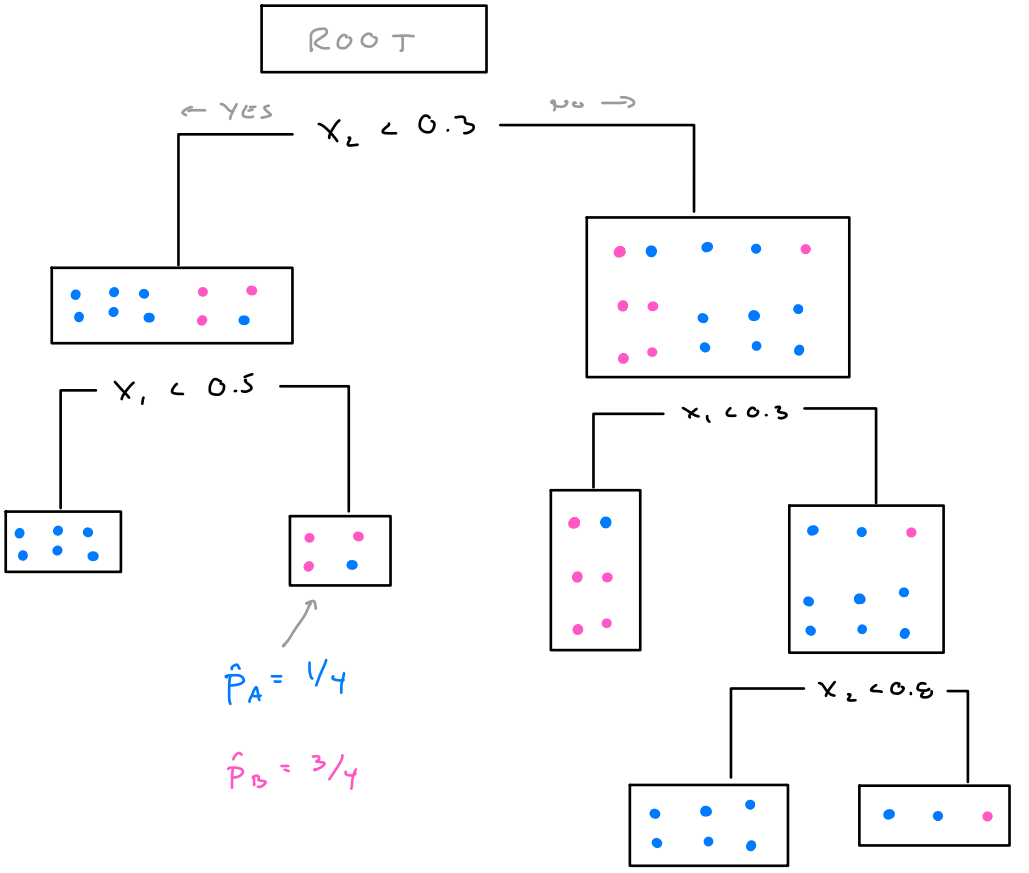
$$\hat{P}[Y = C \mid X = x] = 1/5$$

IF BINARY → USE ODD k

↳ AVOID TIES

Decision Trees





minsplit = 8

cp = 0

↳ DIFFERENT INTERPRETATION
SAME USAGE

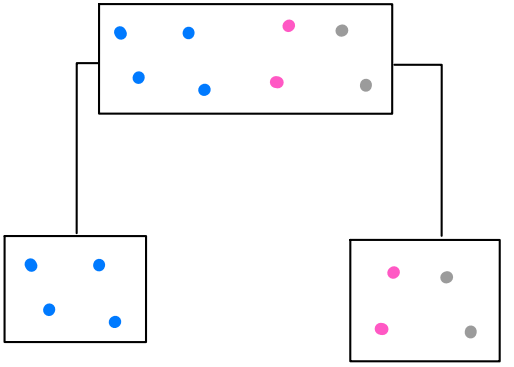
Node Probabilities

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

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

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

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



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

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

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

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

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

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

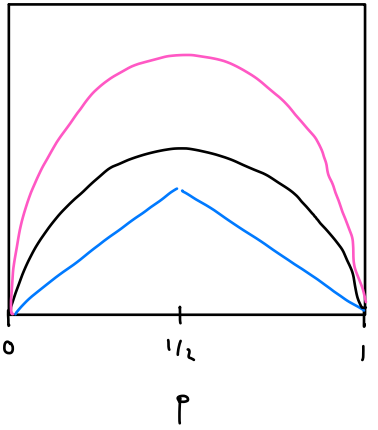
Impurity Measures For Categorical Data

"Variance"

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

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

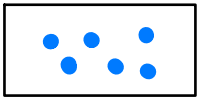
$$\text{Error}(A) = 1 - \max_k (\hat{p}_k)$$



Calculating Gini

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

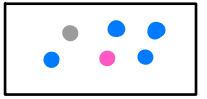
A:



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

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

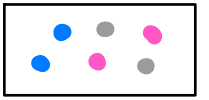
$\hat{p}_C = 0/6$



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

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

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



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

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

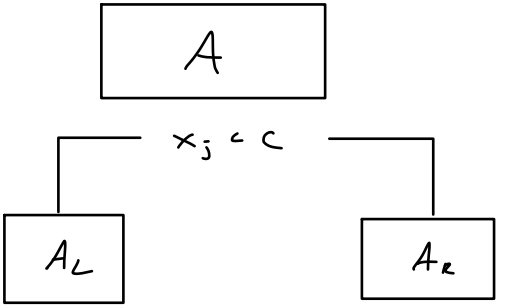
$\hat{p}_C = 2/6$

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

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

Splitting

Find → feature x_j
Find → cutoff c

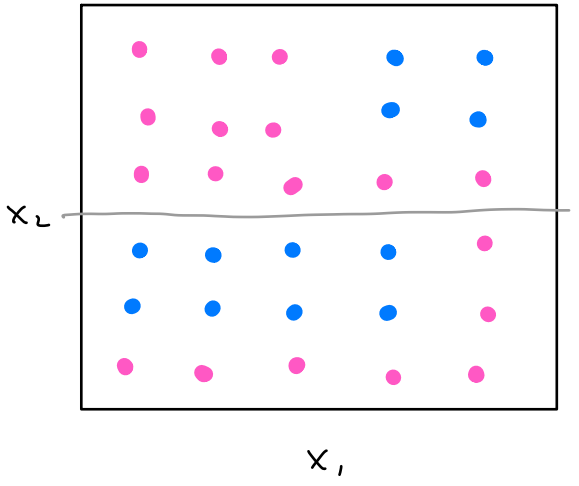


$$\min_{j,c} \left[\frac{|A_L|}{|A|} \text{Gini}(A_L) + \frac{|A_R|}{|A|} \text{Gini}(A_R) \right]$$

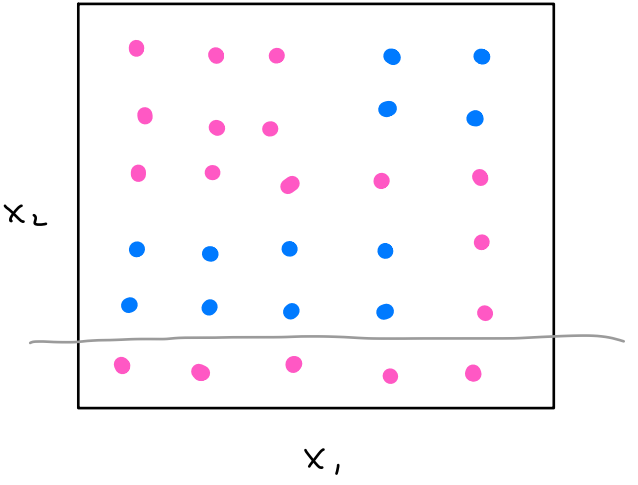
weights

"variance"

Which Split?



$0.4\overline{4}$



Smaller Gini

0.416

In R

KNN `caret::knn3()`

TREES `rpart::rpart()`