# Decision Trees for Classification

CS 307

# NONPARAMETRIC CLASSIFICATION

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

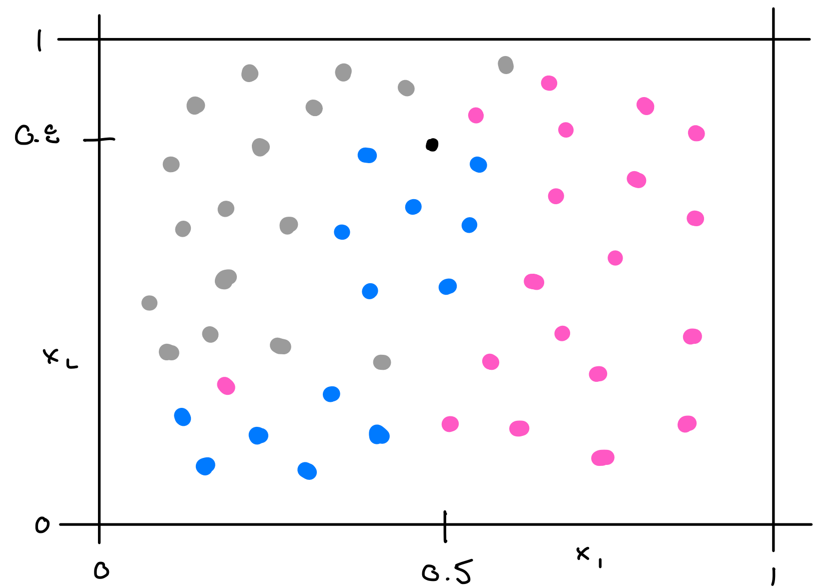
- KNN
- TREES

PROBABILITY OF CLASS $k$

GIVEN THE FEATURE INFORMATION

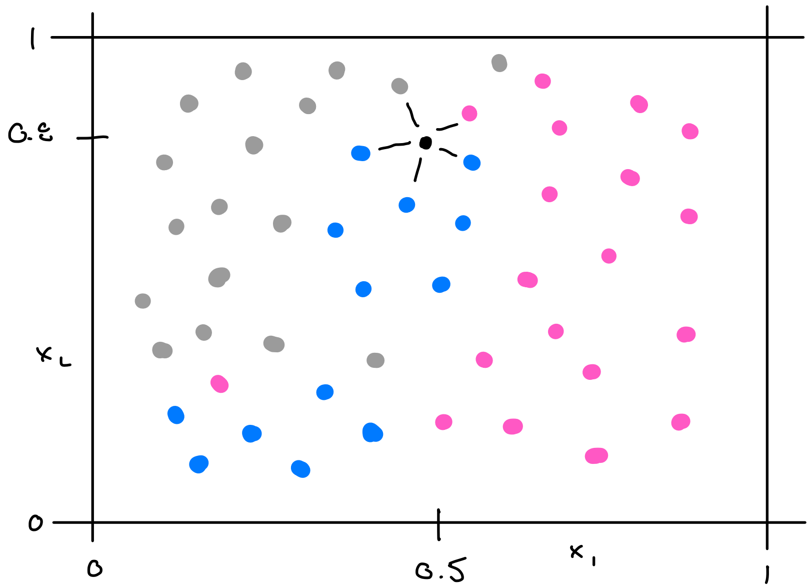
# Setup

| $y$ | $x_1$ | $x_2$ |
|---|---|---|
| A | | |
| ⋮ | | |
| A | | |
| B | | |
| ⋮ | | |
| B | | |
| C | | |
| ⋮ | | |
| C | | |
| ? | 0.5 | 0.8 |



# KNN

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




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

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

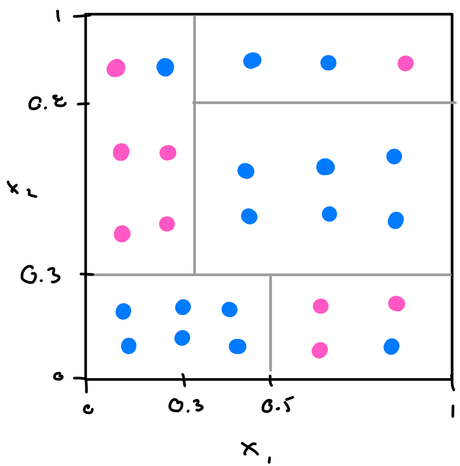
$$\hat{P}[Y = B \mid X = x] = 1/5$$

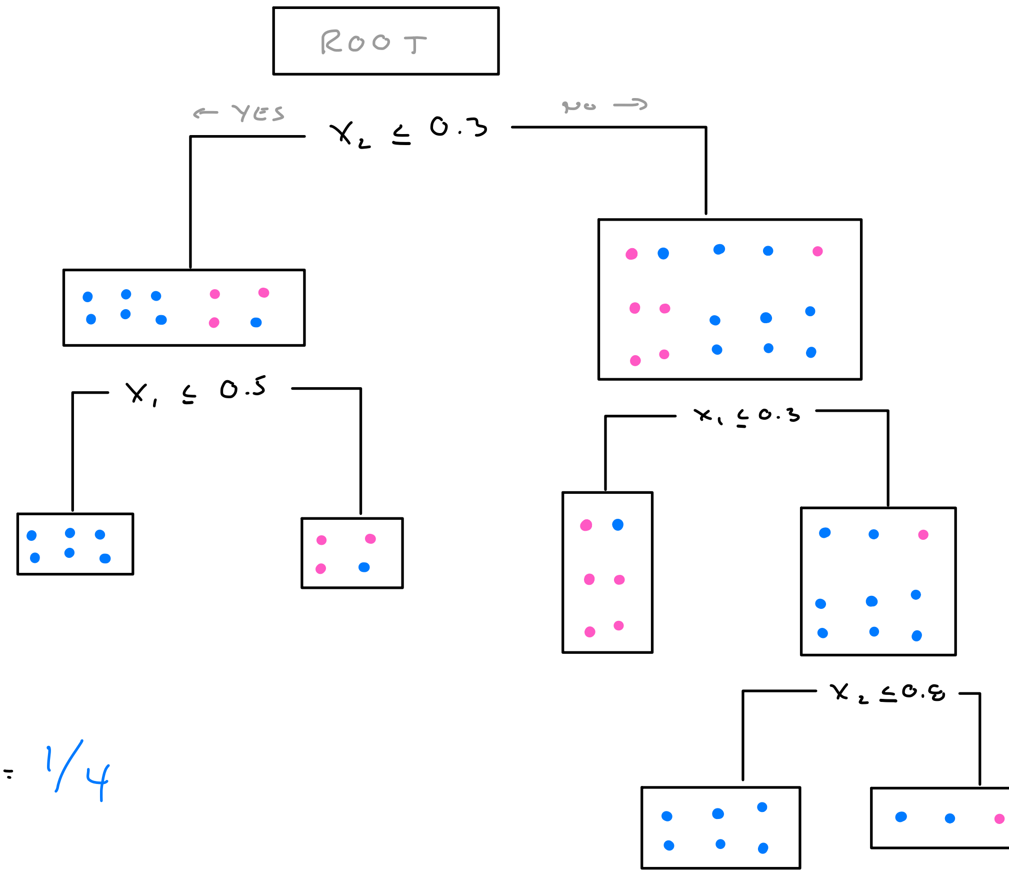
$$\hat{P}[Y = C \mid X = x] = 1/5$$

IF BINARY → USE ODD $k$

└→ AVOID TIES

# Decision Trees




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

# Node Probabilities

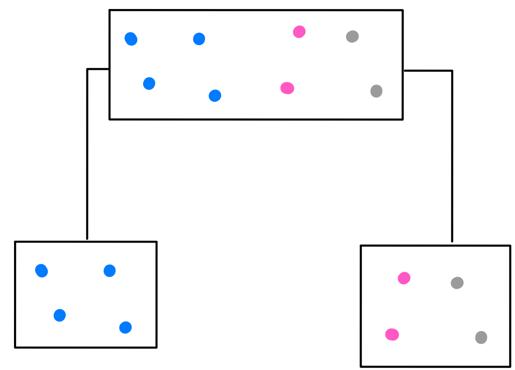
$$\hat{p}_k = \frac{\sum_i I(y_i = k)\, I(x_i \in A)}{\sum_i I(x_i \in A)}$$

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

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

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

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



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

$\hat{p}_B = 0$ | $\hat{p}_B = 2/4$

$\hat{p}_C = 0$ | $\hat{p}_C = 2/4$

"Error"

# Impurity Measures for Categorical Data

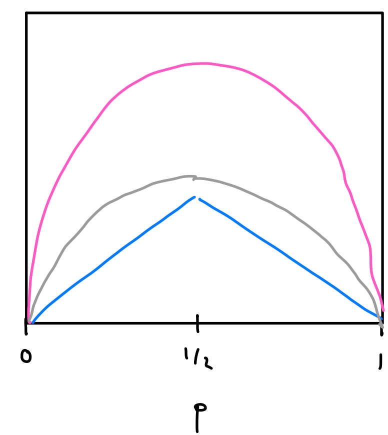
"Variance"

\# categories

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

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

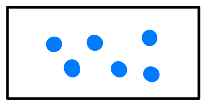
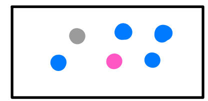
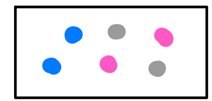
$$\underline{\text{Error}}(A) = 1 - \max_{k} (\hat{p}_k)$$



# 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:

|  |  |  |
|---|---|---|
| $\hat{P}_A = 6/6$ | $\hat{P}_A = 4/6$ | $\hat{P}_A = 2/6$ |
| $\hat{P}_B = 0/6$ | $\hat{P}_B = 1/6$ | $\hat{P}_B = 2/6$ |
| $\hat{P}_C = 0/6$ | $\hat{P}_C = 1/6$ | $\hat{P}_C = 2/6$ |

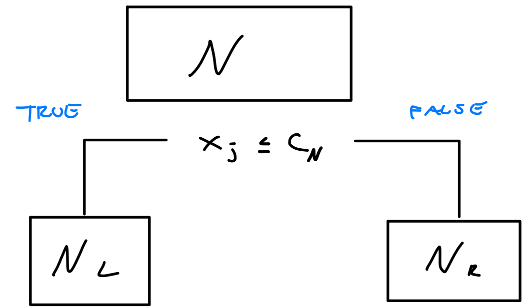
Gini(A): $0$    $0.5$    $0.6\overline{6}$

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

# SPLITTING

FIND → FEATURE $x_j$

FIND → CUTOFF $c_N$

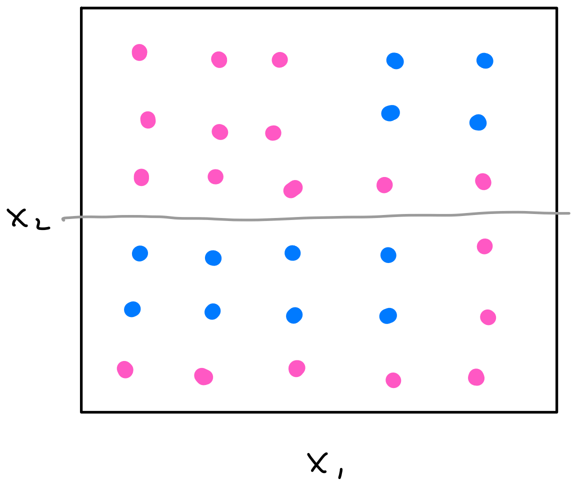


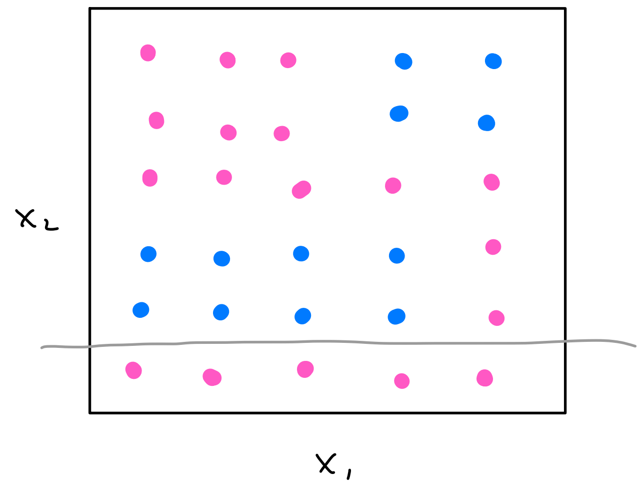

$$\min_{j,c} \left[ \underbrace{\frac{|N_L|}{|N|}} \underbrace{\mathrm{GINI}(N_L)} + \underbrace{\frac{|N_R|}{|N|}} \underbrace{\mathrm{GINI}(N_R)} \right]$$

WEIGHTS

VARIANCES

# Which Split?




0.44      0.416