HUDK 4050: CORE METHODS IN EDM

11/15/17 4:39 PM

Due

- Assignment 4: Today
- Assignment 5: 11/21
- Assignment 6: 11/28
- Assignment 7: 12/5
- Formative Test: 12/5
- Assignment 8: 12/21
- Stack Overflow: 12/21
- Notes: 12/21

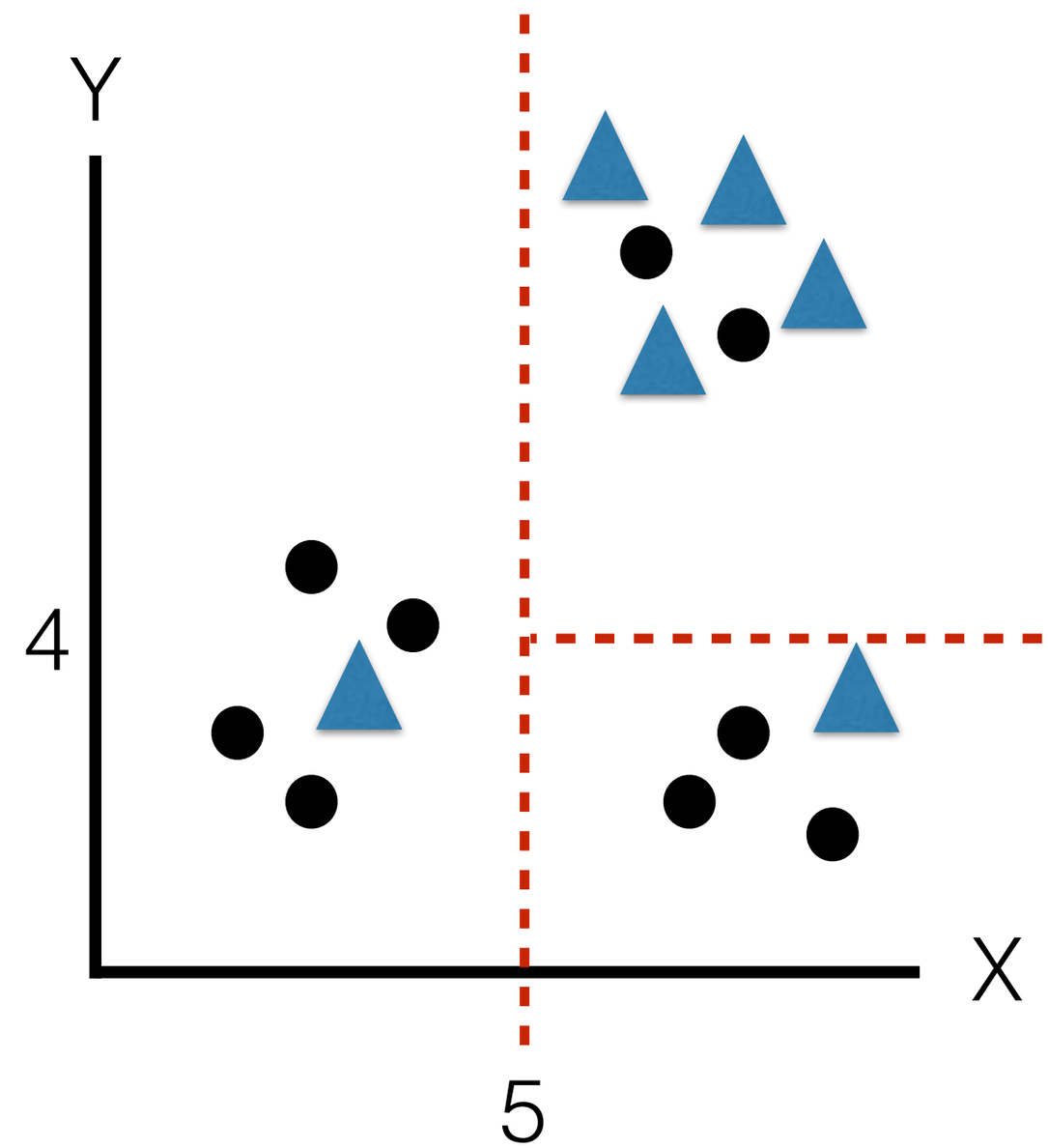
Prediction

CART Trees

- Leo Breiman invented in the 1970s
- Non-parametric model
- Designed to deal with data that has too many interaction effects
- Trademarked CART (classification & regression trees) so is called rpart (recursive partitioning and regression trees)

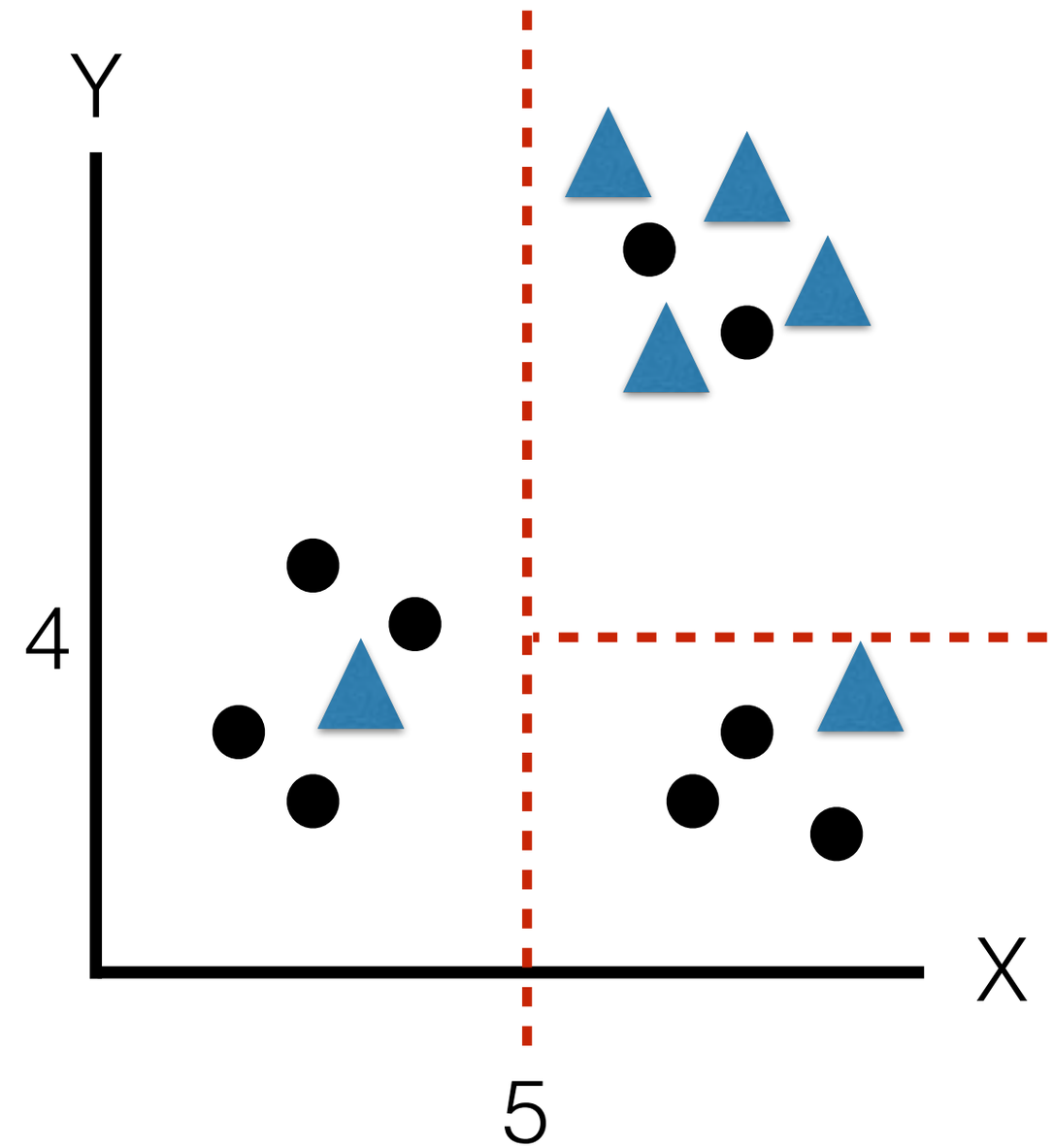
RPART

- Recursive (splits the leaves until you tell it to stop)
- At each step, the split is made based on the independent variable that results in the largest possible <u>reduction in heterogeneity</u> of the dependent (predicted) variable



Heterogeneity

- Impurity/homogeneity
- leaf has only 1 class, impurity = 0
- Entropy (information gain)
- Gini index
- Classification error



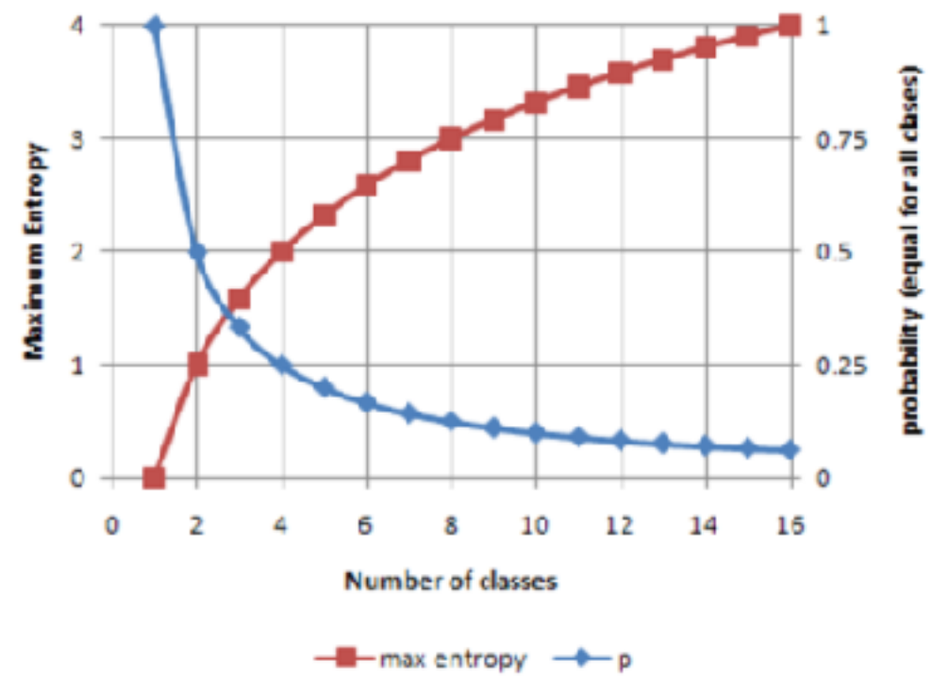
RPART

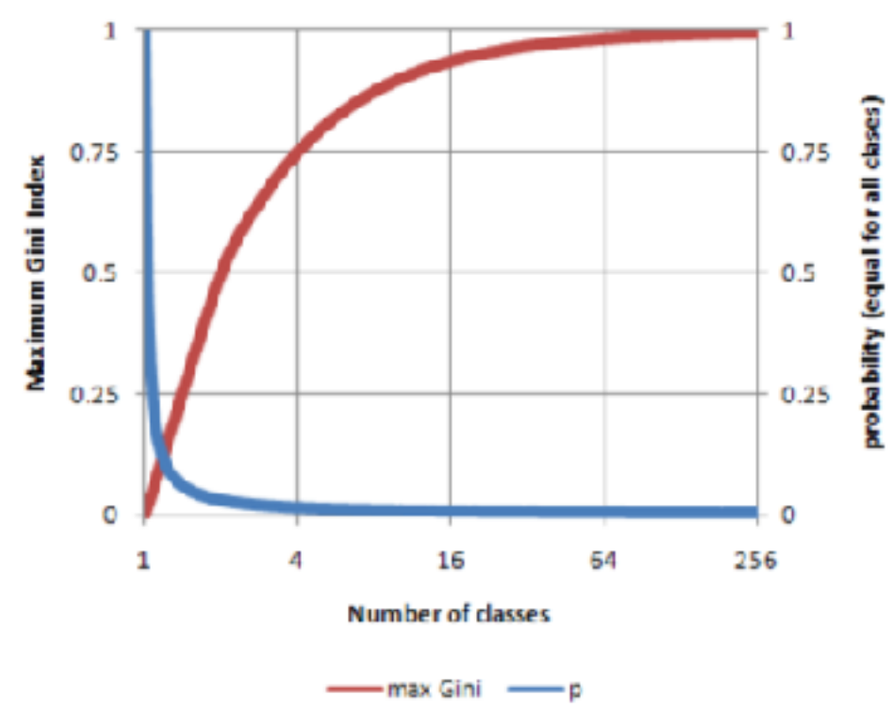
- `parms`
- Entropy (information gain)

$$Entropy = \sum_j -p_j log_2 p_j$$

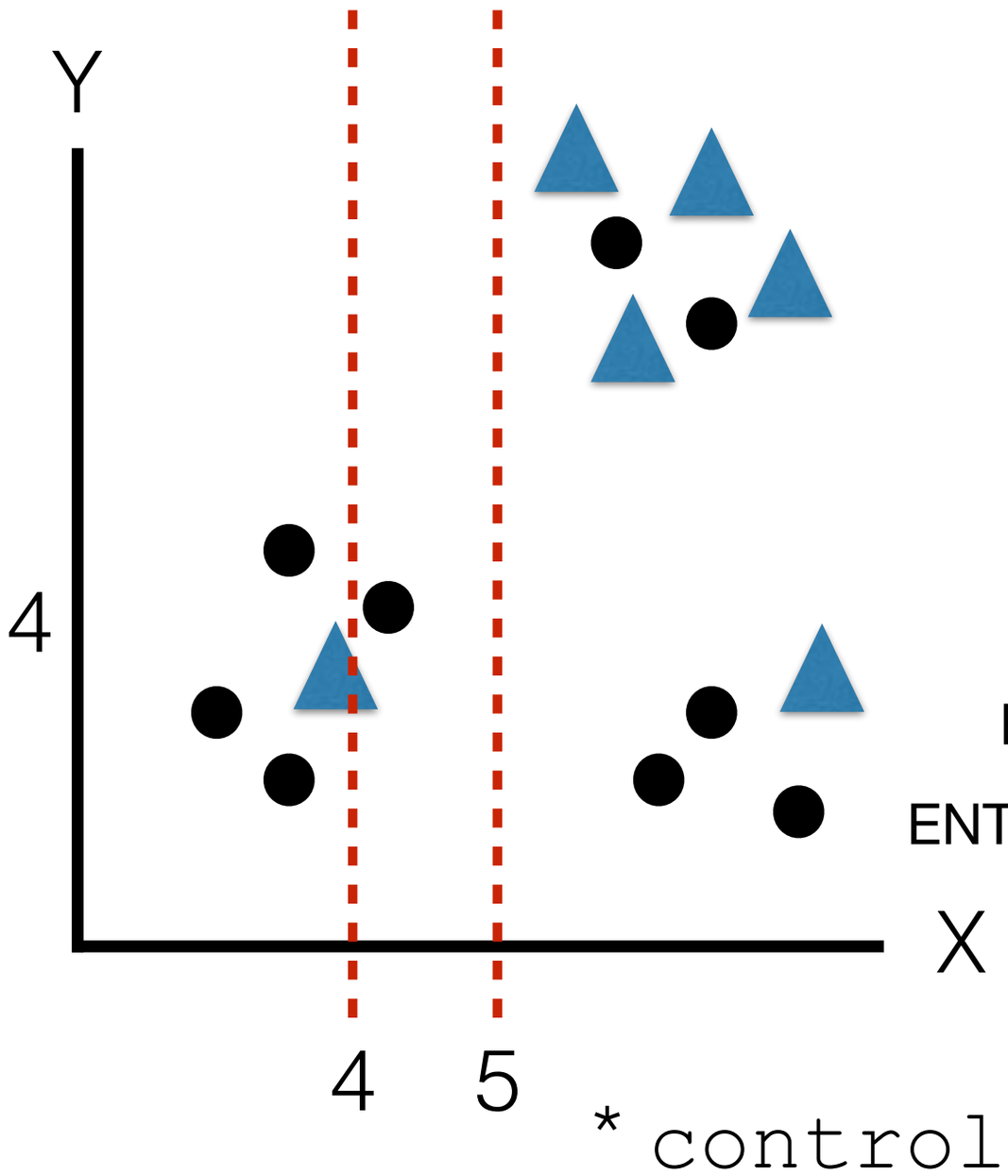
- Gini index

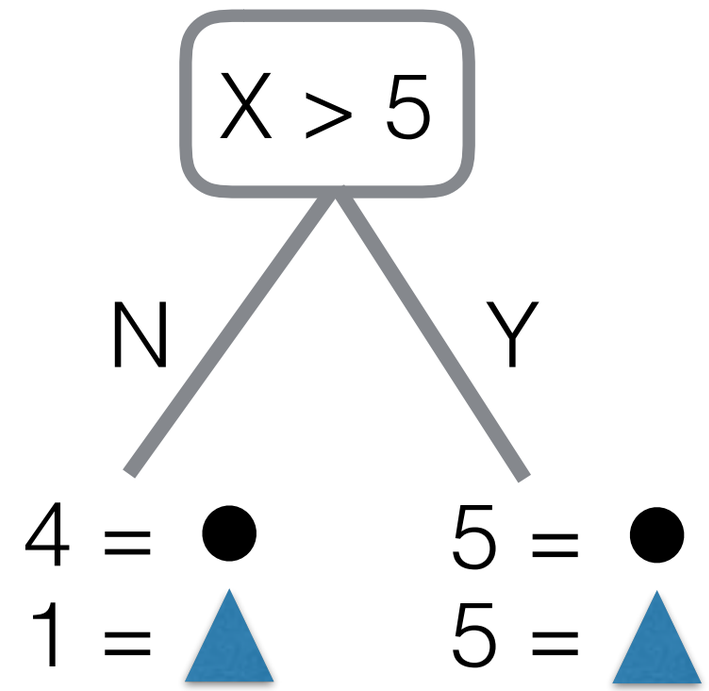
$$Gini\ Index = 1 - \sum_j p_j^2$$





RPART



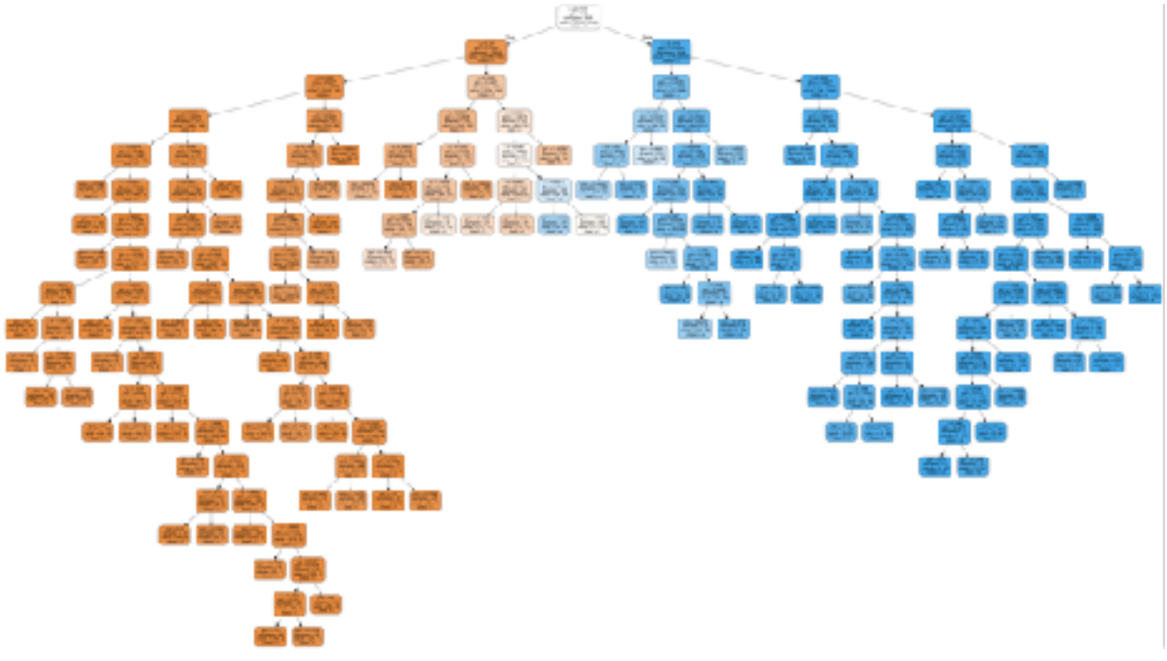


$ENT_5 = -0.8*\log_2(0.8) + -0.5*\log_2(0.5) = 0.75$

$ENT_4 = -0.75*\log_2(0.75) + -0.55*\log_2(0.55) = 0.76$

RPART

- Tree chooses the optimal fit at each leaf - NOT the overall best fit for the data
- Therefore, there is a danger of <u>overfitting</u> the tree
- Tree is too specific to training data to be able to predict new data
- Therefore: stop the tree at a certain number of nodes OR prune



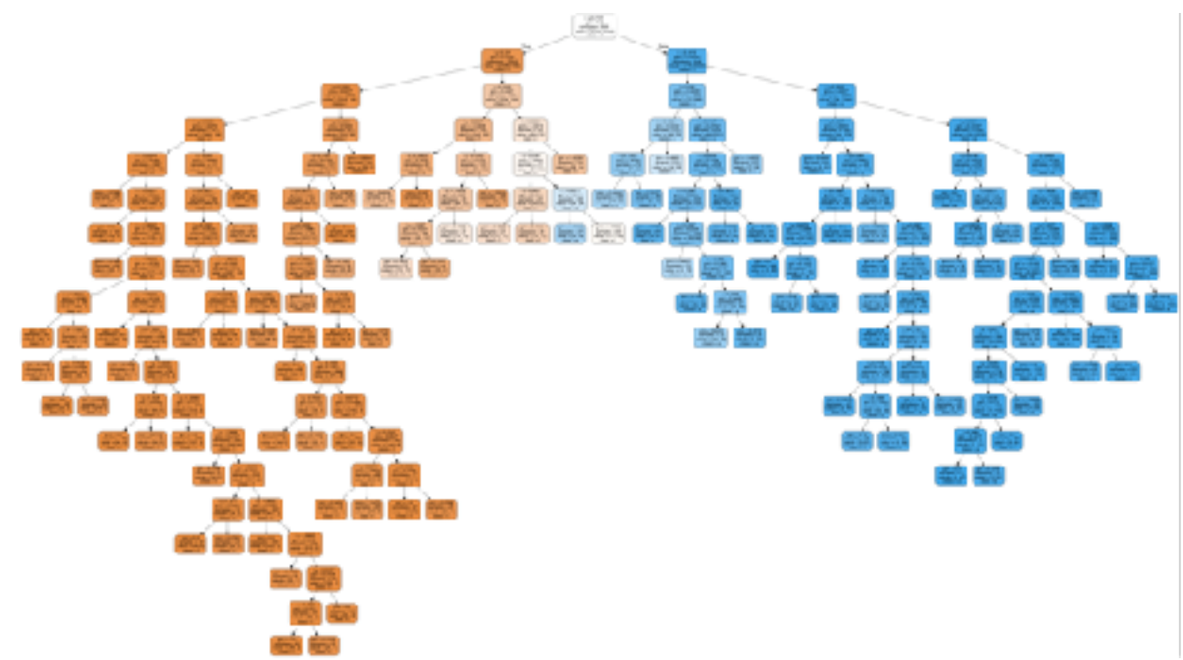
RPART

- RPART prunes
- Uses a cost function:

Number of leaves

$$C_\alpha(T) = R(T) + \alpha|\tilde{T}|$$

Misclassified instances



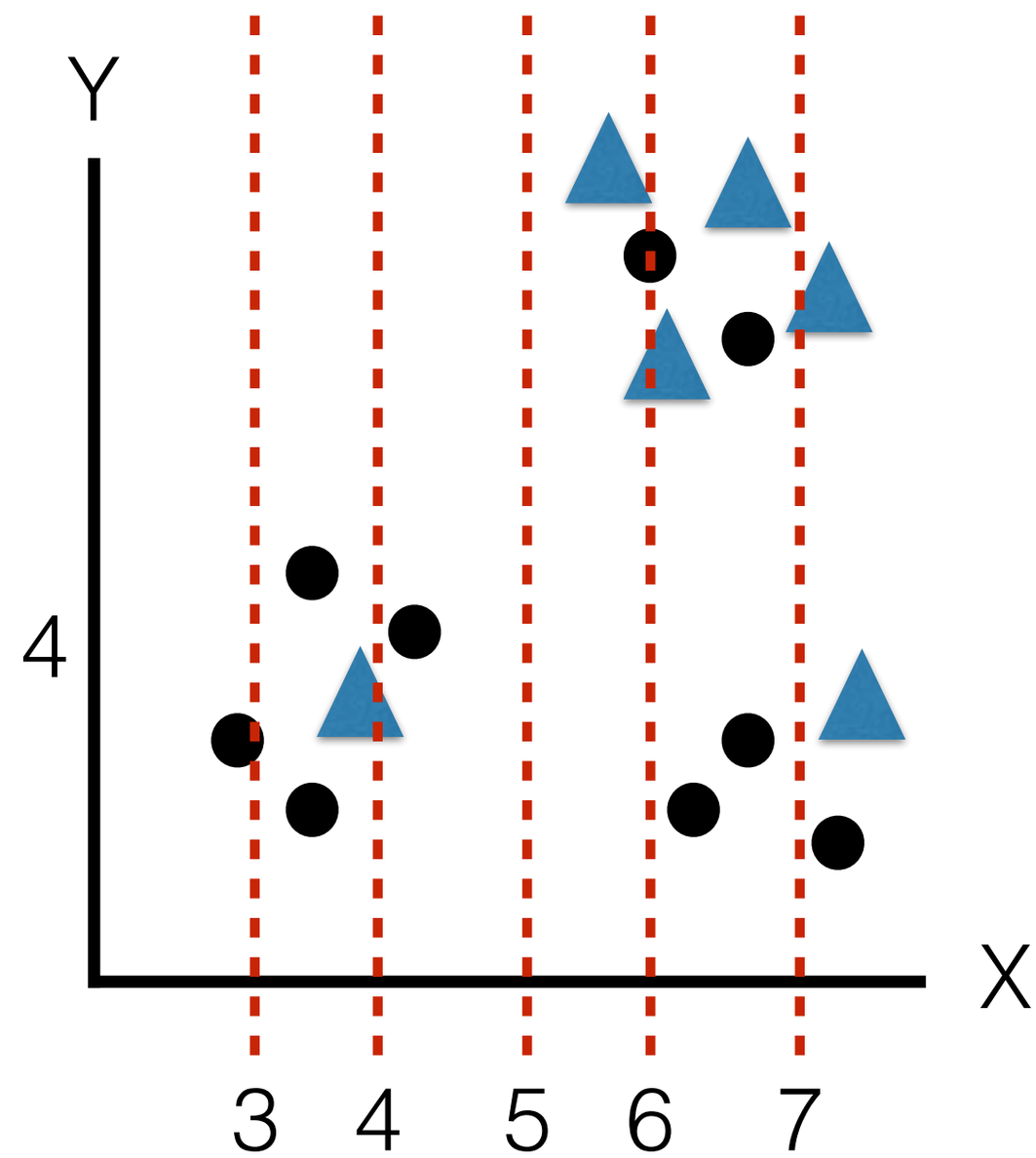
RPART

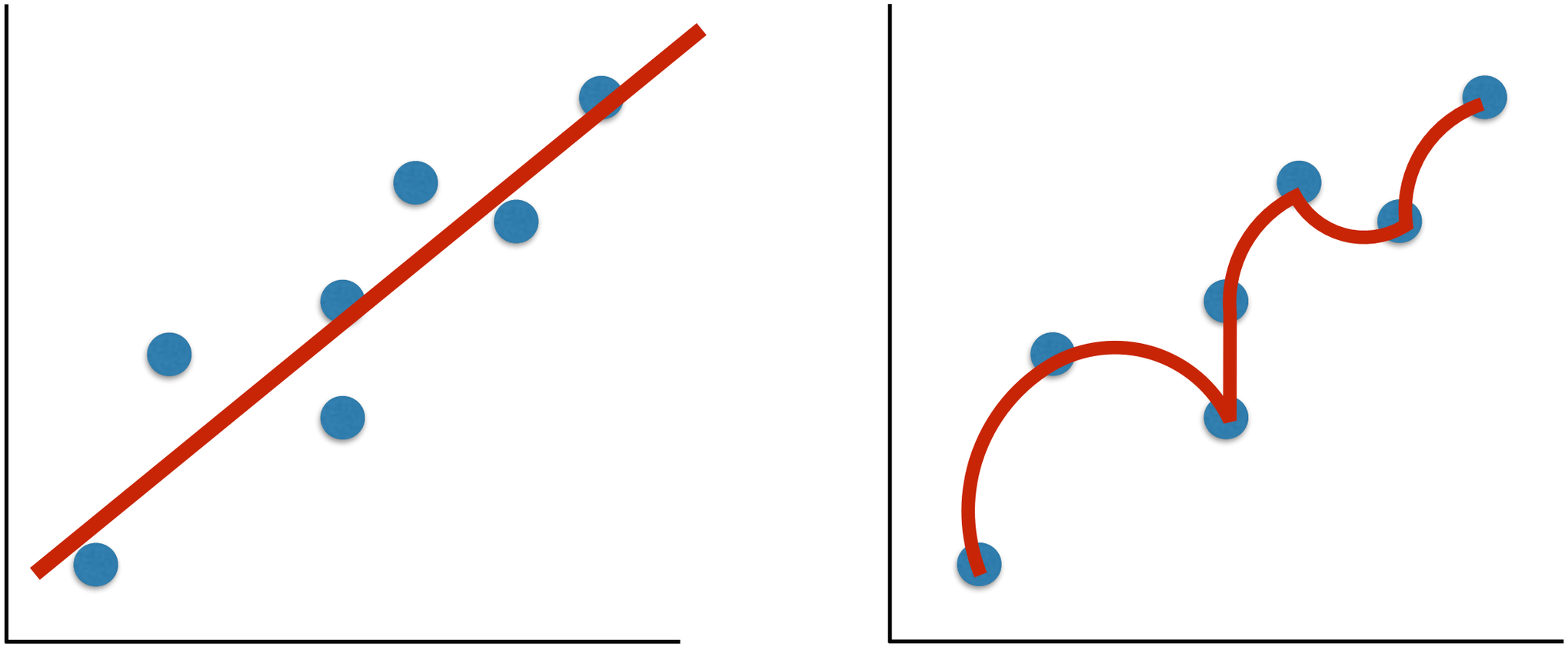
<u>Gotchas</u>

- Overfitting
- Local overfitting
- Sensitive to test data
- Selection bias toward covariates with many possible splits

PARTY

- “part(y)itioning” 😎
- Conditional Inference Tree
- Look at correlation between X and shape and Y and shape
- Statistically test H_0: there is no relationship
- Choose the variable with the highest correlation
- Split on that variable
- Stop when H_0 cannot be rejected





Which is more “accurate”?

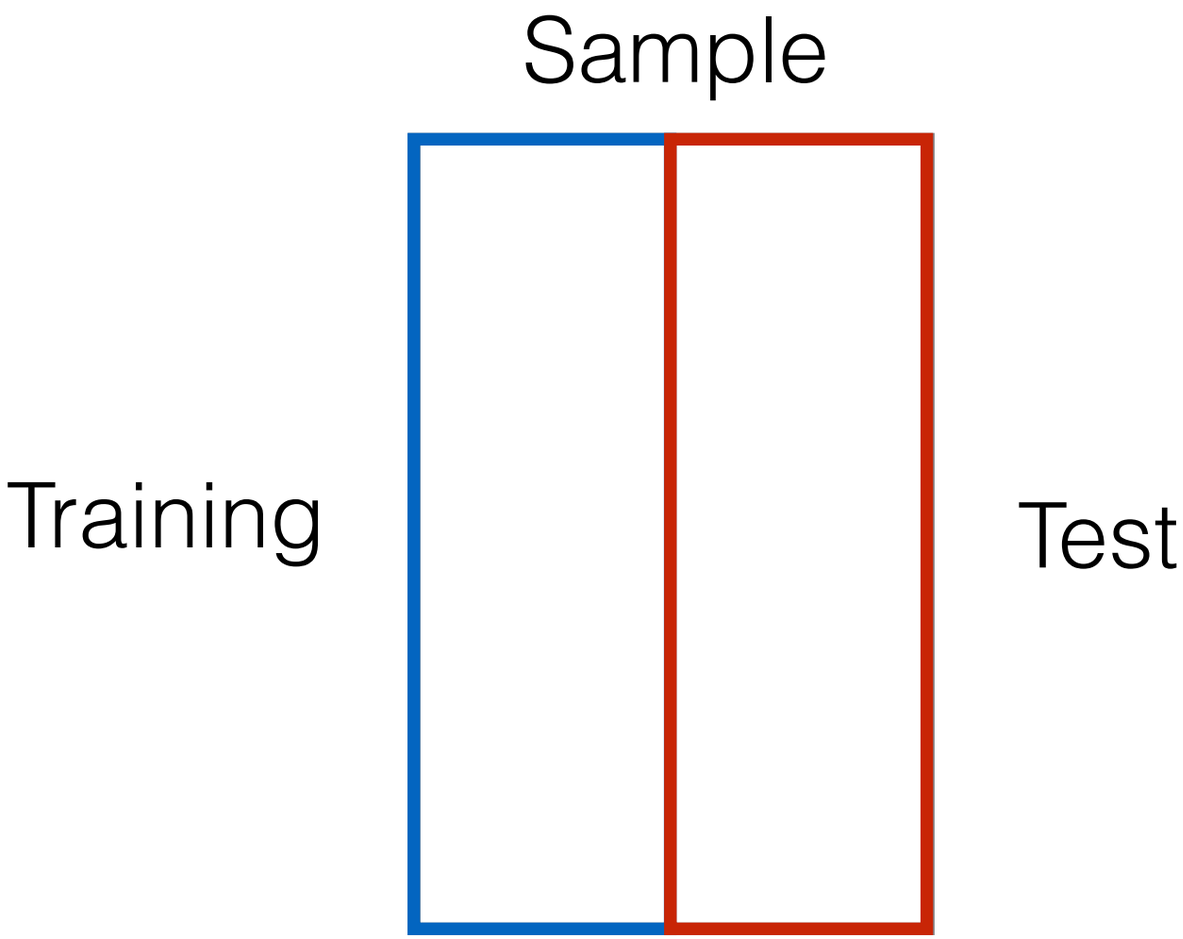
Which is more “useful”?

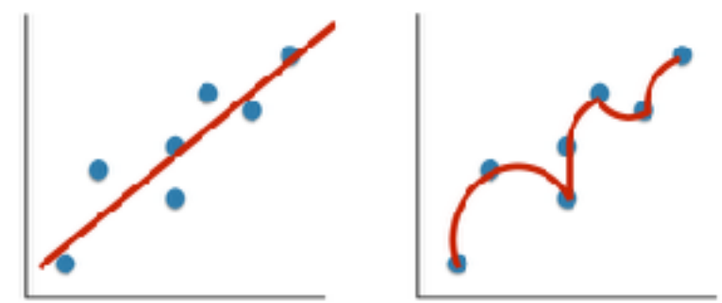
How can we tell?

Cross Validation

- Estimate how accurately a predictive model will perform in practice
- Give an insight on how the model will generalize to an independent dataset

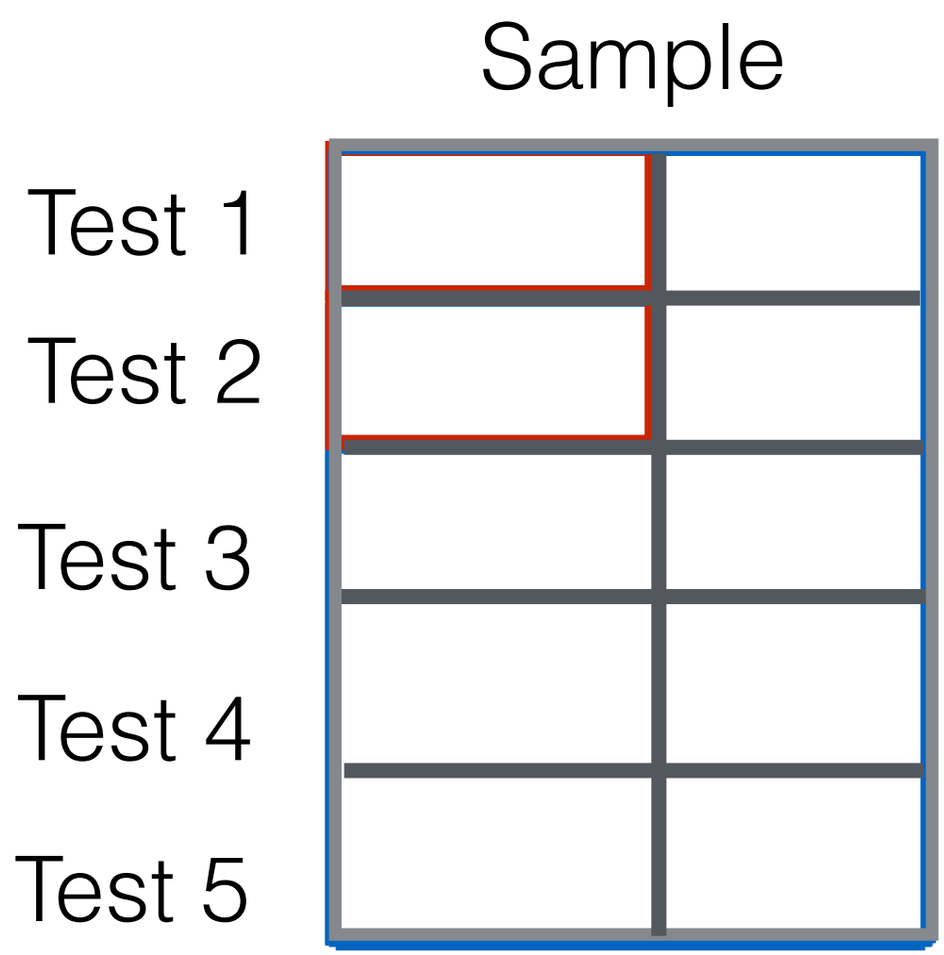
Hold-out Validation

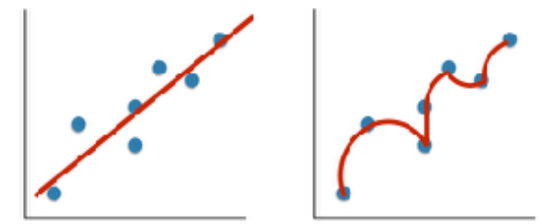




Problem: very dependent on which data are in each group

K-Fold Cross Validation





Training 1	5	2
Training 2	4	2
Training 3	3	1
Training 4	5	4
Training 5	4	2
	4.2	2.2

Calculate how accurate we are in each “fold” and average the answer

http://bit.ly/cmedma6