DS 4400

Machine Learning and Data Mining I
Spring 2024

David Liu

Khoury College of Computer Science

Northeastern University

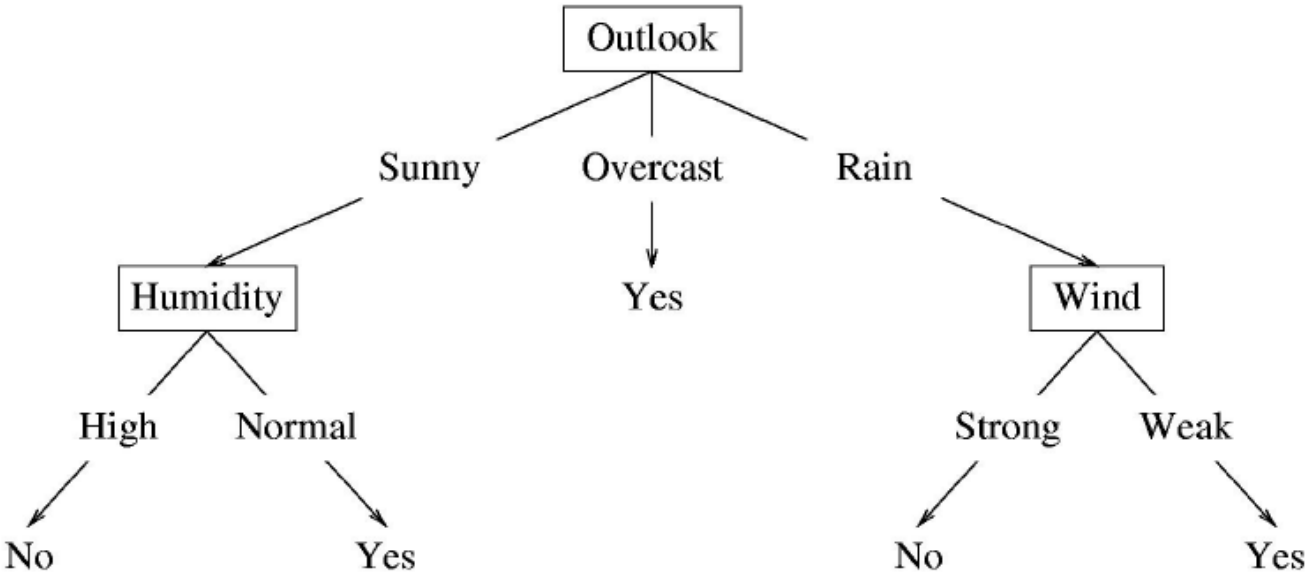
March 1 2024

Outline

- Decision trees
  - Information gain / entropy measures
  - Training algorithm
  - Example
- Ensemble models
  - Bagging
  - Boosting

Decision Tree

- A possible decision tree for the data:

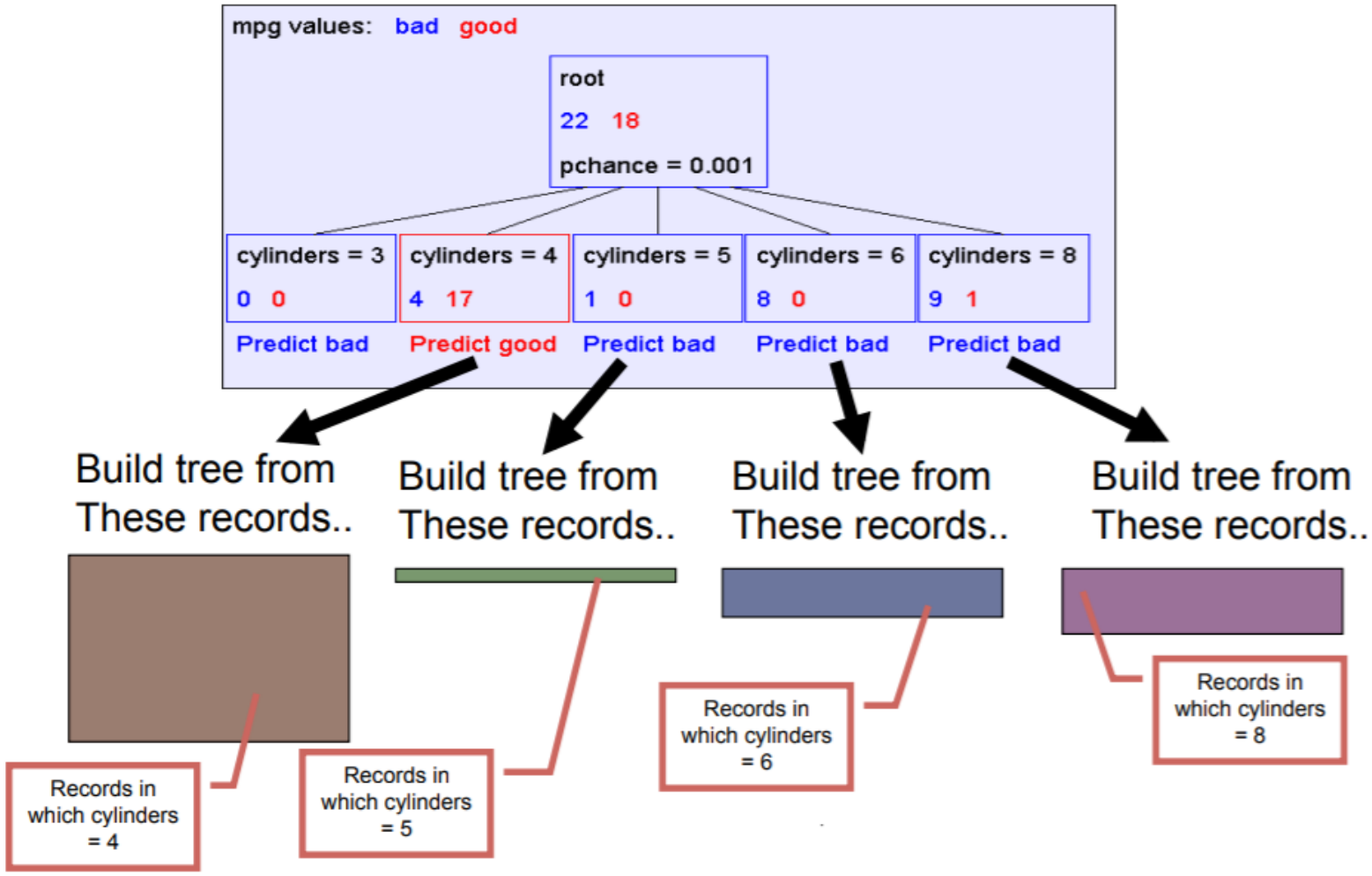


- Each internal node: test one attribute X_i
- Each branch from a node: selects one value for X_i
- Each leaf node: predict Y (or $p(Y \mid x \in \text{leaf})$)

Learning Decision Trees

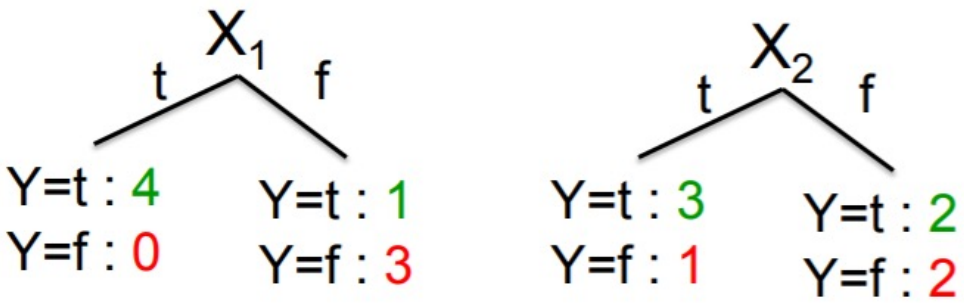
- Learning the simplest (smallest) decision tree is an NP-complete problem [Hyafil & Rivest '76]
- Resort to a greedy heuristic:
  - Start from empty decision tree
  - Split on **next best attribute (feature)**
  - Recurse

Key Idea: Use Recursion Greedily



Splitting

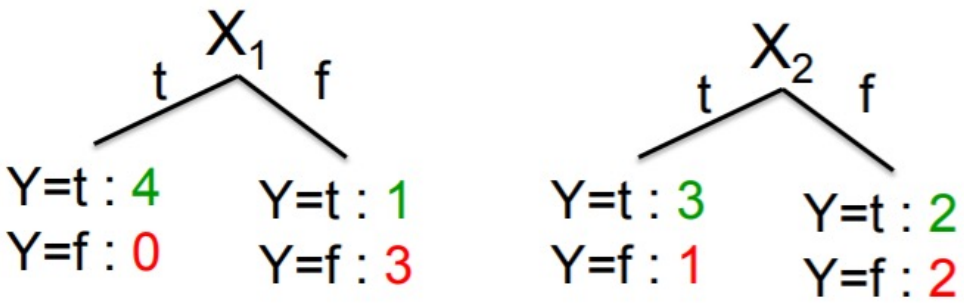
Would we prefer to split on X_1 or X_2?



X_1	X_2	Y
T	T	T
T	F	T
T	T	T
T	F	T
F	T	T
F	F	F
F	T	F
F	F	F

Splitting

Would we prefer to split on X_1 or X_2?



X_1	X_2	Y
T	T	T
T	F	T
T	T	T
T	F	T
F	T	T
F	F	F
F	T	F
F	F	F

Idea: use counts at leaves to define probability distributions, so we can measure uncertainty!

Use entropy-based measure (Information Gain)

Entropy

Suppose X can have one of *m* values... $V_1, V_2, \ldots V_m$

$P(X=V_1) = p_1$	$P(X=V_2) = p_2$	....	$P(X=V_m) = p_m$

What's the smallest possible number of bits, on average, per symbol, needed to transmit a stream of symbols drawn from X's distribution? It's

$$H(X) = -p_1 \log_2 p_1 - p_2 \log_2 p_2 - \ldots - p_m \log_2 p_m$$

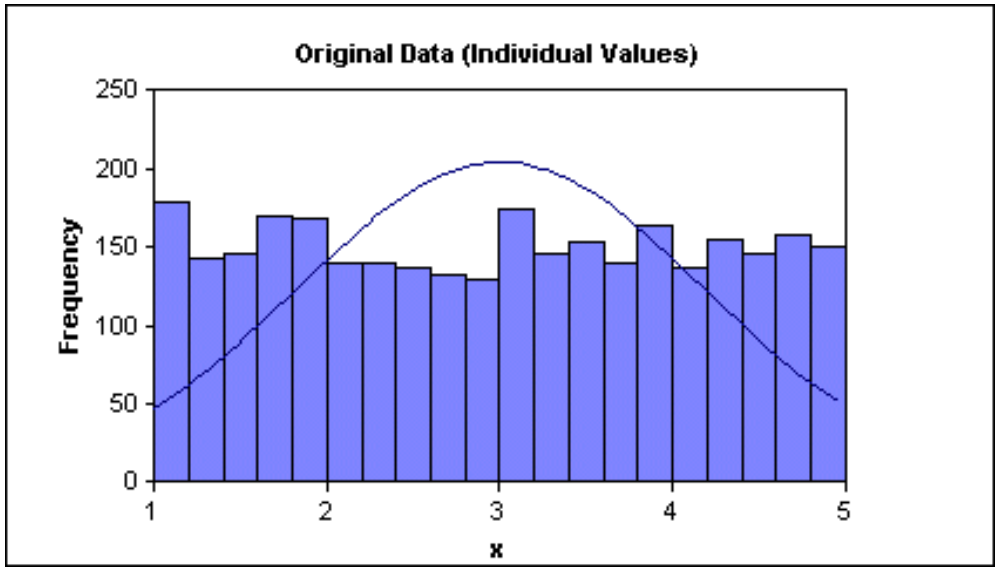
$$= -\sum_{j=1}^{m} p_j \log_2 p_j$$

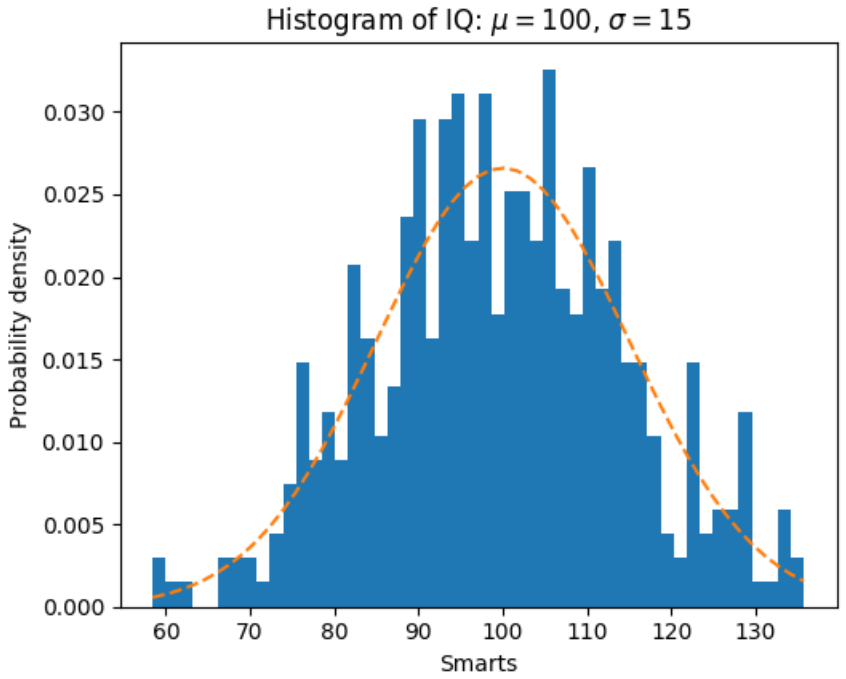
H(X) = The entropy of X

- "High Entropy" means X is from a uniform (boring) distribution
- "Low Entropy" means X is from varied (peaks and valleys) distribution

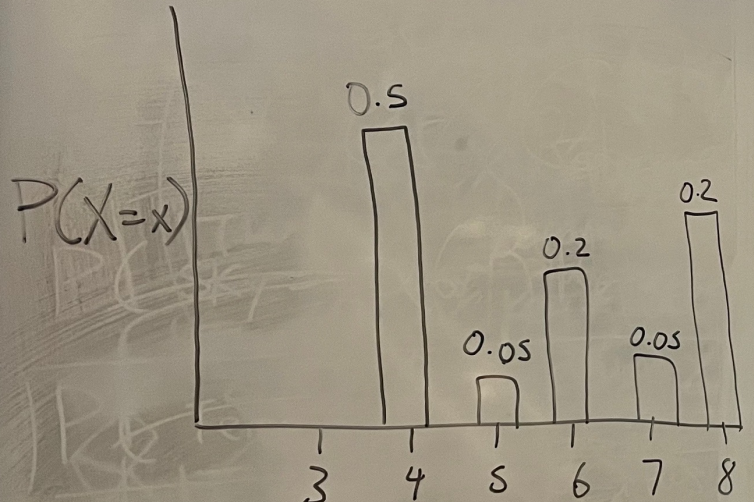
High/Low Entropy

Which distribution has high entropy?





Entropy



$$H(x) = -\sum_{j=1}^{m} P_j \log_2(P_j)$$

"Uncertainty" — Weight — # of bits

Uniform → High Uncertain

$0.5 \log(0.5) + 0.05 \log(0.05) + 0.2 \log(0.2)$

$+ 0.05 \log(0.05) + 0.2 \log(0.2)$

Conditional Entropy

Suppose I'm trying to predict output Y and I have input X

X = College Major

Y = Likes "Gladiator"

X	Y
Math	Yes
History	No
CS	Yes
Math	No
Math	No
CS	Yes
History	No
Math	Yes

Let's assume this reflects the true probabilities

Conditional Entropy

Suppose I'm trying to predict output Y and I have input X

X = College Major

Y = Likes "Gladiator"

X	Y
Math	Yes
History	No
CS	Yes
Math	No
Math	No
CS	Yes
History	No
Math	Yes

Let's assume this reflects the true probabilities

E.G. From this data we estimate

- *P(LikeG = Yes) = 0.5*
- *P(Major = Math & LikeG = No) = 0.25*
- *P(Major = Math) = 0.5*
- *P(LikeG = Yes | Major = History) = 0*

Note:

- *H(X) = 1.5*
- *H(Y) = 1*

Conditional Entropy

X = College Major

Y = Likes "Gladiator"

X	Y
Math	Yes
History	No
CS	Yes
Math	No
Math	No
CS	Yes
History	No
Math	Yes

Definition of Specific Conditional Entropy:

$H(Y|X=v)$ = **The entropy of Y among only those records in which X has value v**

Example:

- $H(Y|X=Math) =$
- $H(Y|X=History) =$
- $H(Y|X=CS) =$

Conditional Entropy

X = College Major

Y = Likes "Gladiator"

X	Y
Math	Yes
History	No
CS	Yes
Math	No
Math	No
CS	Yes
History	No
Math	Yes

Definition of Specific Conditional Entropy:

$H(Y|X=v)$ = **The entropy of** Y **among only those records in which** X **has value** v

Example:

- $H(Y|X=Math) = 1$
- $H(Y|X=History) = 0$
- $H(Y|X=CS) = 0$

Conditional Entropy

X = College Major

Y = Likes "Gladiator"

X	Y
Math	Yes
History	No
CS	Yes
Math	No
Math	No
CS	Yes
History	No
Math	Yes

Definition of Conditional Entropy:

$H(Y|X)$ = The average specific conditional entropy of Y

= if you choose a record at random what will be the conditional entropy of Y, conditioned on that row's value of X

= Expected number of bits to transmit Y if both sides will know the value of X

$= \Sigma_j Prob(X=v_j)\, H(Y \mid X = v_j)$

Conditional Entropy

X = College Major

Y = Likes "Gladiator"

X	Y
Math	Yes
History	No
CS	Yes
Math	No
Math	No
CS	Yes
History	No
Math	Yes

Definition of Conditional Entropy:

$H(Y|X)$ = The average conditional entropy of Y

$= \Sigma_j Prob(X=v_j)\, H(Y \mid X = v_j)$

Example:

v_j	$Prob(X=v_j)$	$H(Y \mid X = v_j)$
Math		
History		
CS		

Conditional Entropy

X = College Major

Y = Likes "Gladiator"

X	Y
Math	Yes
History	No
CS	Yes
Math	No
Math	No
CS	Yes
History	No
Math	Yes

Definition of Conditional Entropy:

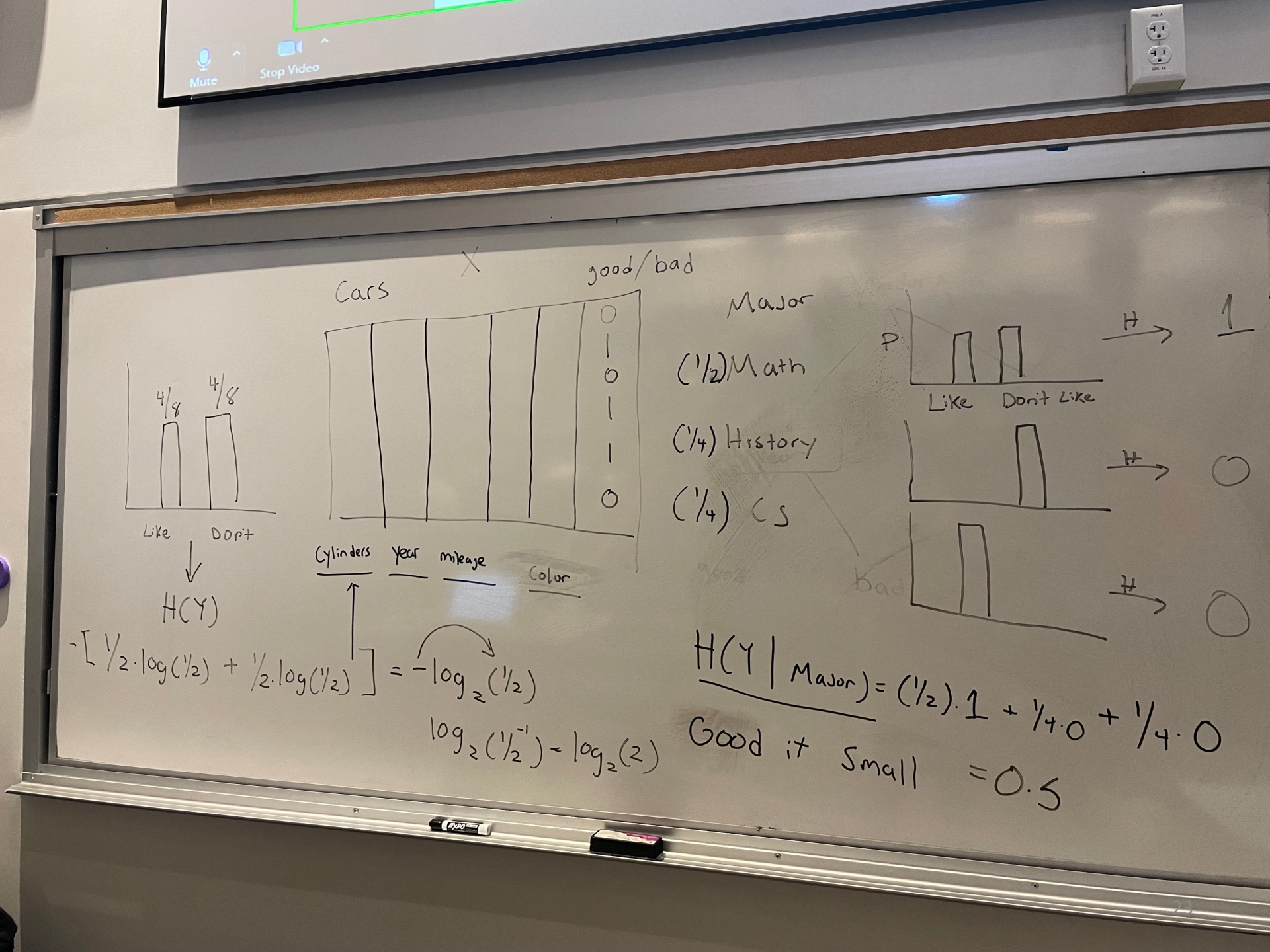
$H(Y|X)$ = The average conditional entropy of Y

$= \Sigma_j Prob(X=v_j)\, H(Y \mid X = v_j)$

Example:

v_j	$Prob(X=v_j)$	$H(Y \mid X = v_j)$
Math	0.5	1
History	0.25	0
CS	0.25	0

$H(Y|X) = 0.5 * 1 + 0.25 * 0 + 0.25 * 0 = 0.5$


Mute
Stop Video
Cars
X
good/bad
Major
(1/2) Math
(1/4) History
(1/4) CS
4/8
4/8
Like
Don't
Cylinders
Year
Mileage
Color
H(Y)
$-[\frac{1}{2}\cdot\log(\frac{1}{2}) + \frac{1}{2}\cdot\log(\frac{1}{2})] = -\log_2(\frac{1}{2})$
$\log_2(\frac{1}{2}^{-1}) = \log_2(2)$
P
H
1
Like
Don't Like
0
bad
0
$H(Y \mid \text{Major}) = (\frac{1}{2})\cdot 1 + \frac{1}{4}\cdot 0 + \frac{1}{4}\cdot 0$
Good it Small
$= 0.5$

Information Gain

X = College Major

Y = Likes "Gladiator"

X	Y
Math	Yes
History	No
CS	Yes
Math	No
Math	No
CS	Yes
History	No
Math	Yes

Definition of Information Gain:

$IG(Y|X)$ = **I must transmit** Y. **How many bits on average would it save me if both ends of the line knew** X**?**

$IG(Y|X) = H(Y) - H(Y|X)$

Example:

- **H(Y) =**
- **H(Y|X) =**
- **Thus IG(Y|X) =**

Information Gain

X = College Major

Y = Likes "Gladiator"

X	Y
Math	Yes
History	No
CS	Yes
Math	No
Math	No
CS	Yes
History	No
Math	Yes

Definition of Information Gain:

$IG(Y|X)$ = **I must transmit** Y. **How many bits on average would it save me if both ends of the line knew** X**?**

$$IG(Y|X) = H(Y) - H(Y|X)$$

Example:

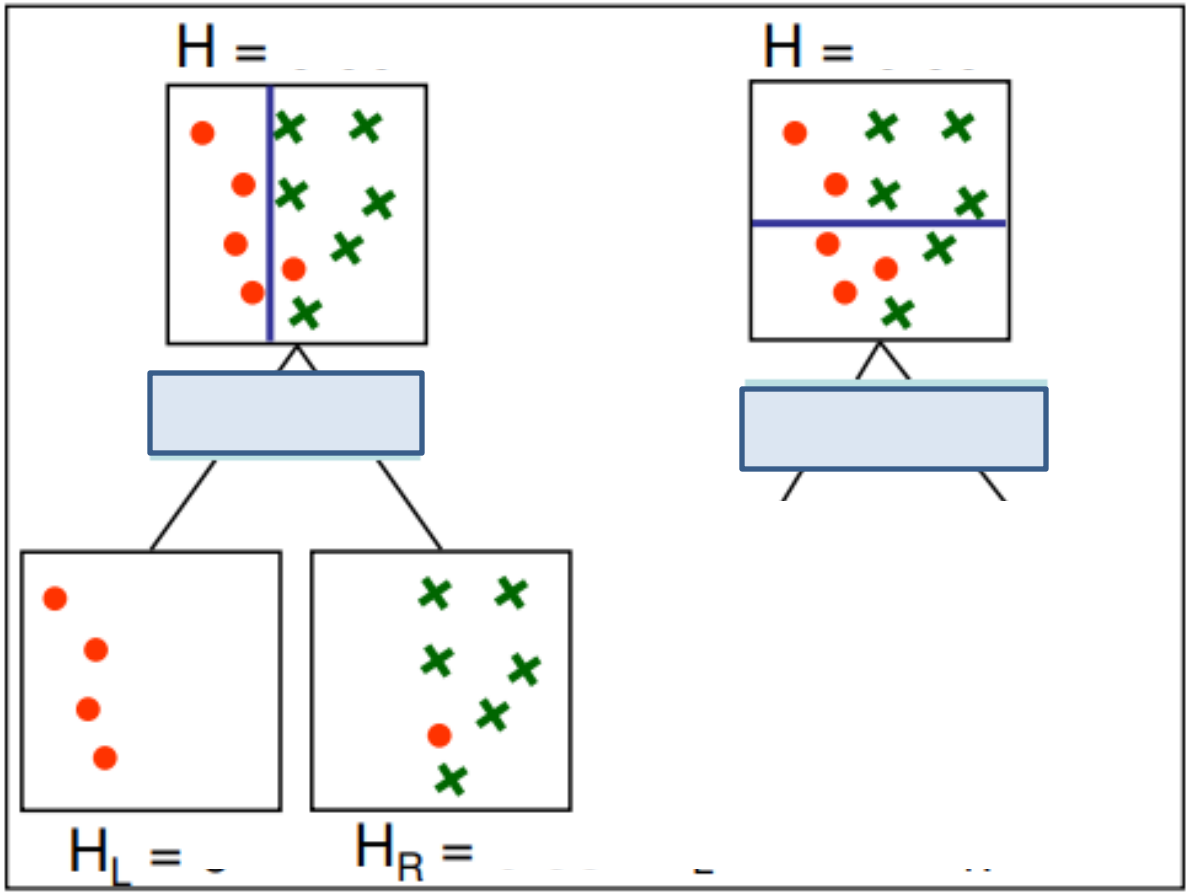
- **H(Y) = 1**
- **H(Y|X) = 0.5**
- **Thus IG(Y|X) = 1 – 0.5 = 0.5**

Relevance for decision trees

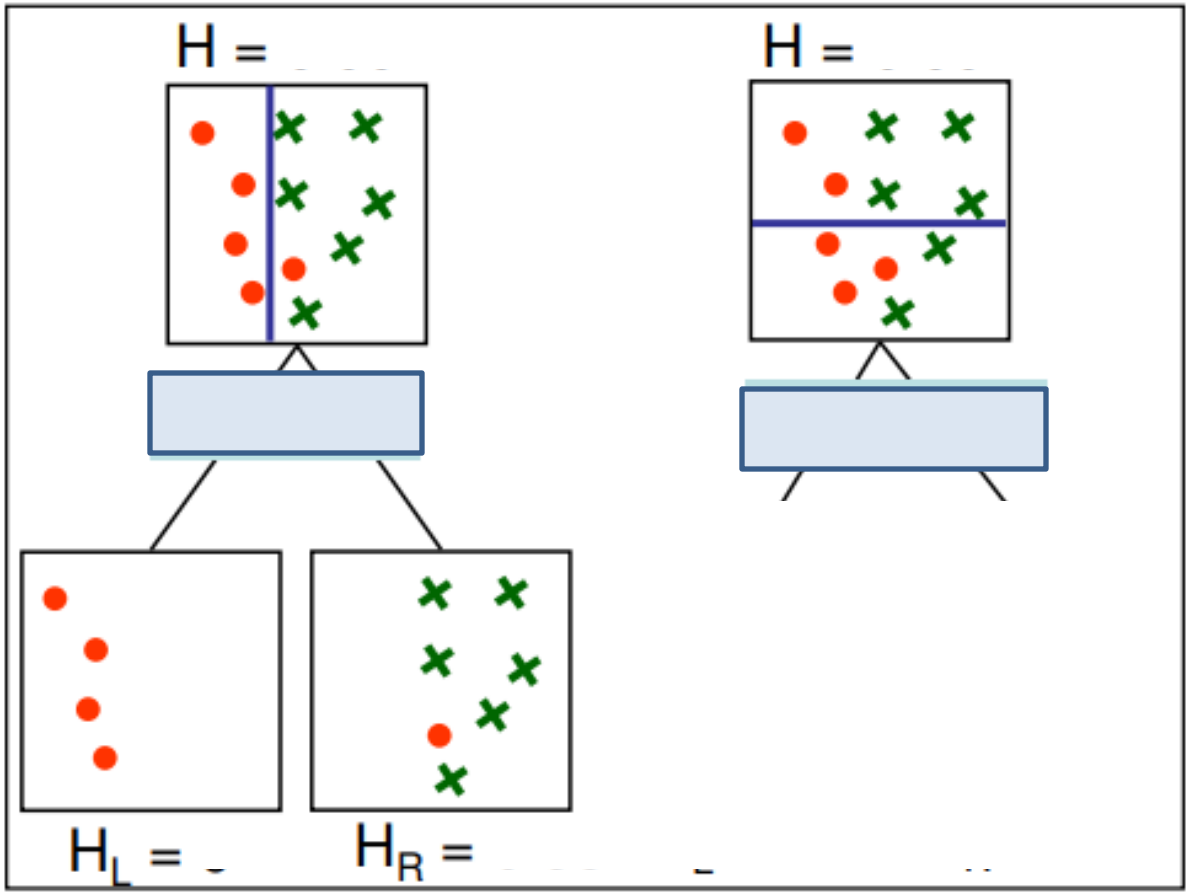
- Multiple features $X_1, \ldots, X_d$
- Label Y: Initial entropy $H(Y)$
- How much each feature X_i helps explain uncertainty in Y
  - Compute Information gain

$$IG(Y|X_i) = H(Y) - H(Y|X_i)$$

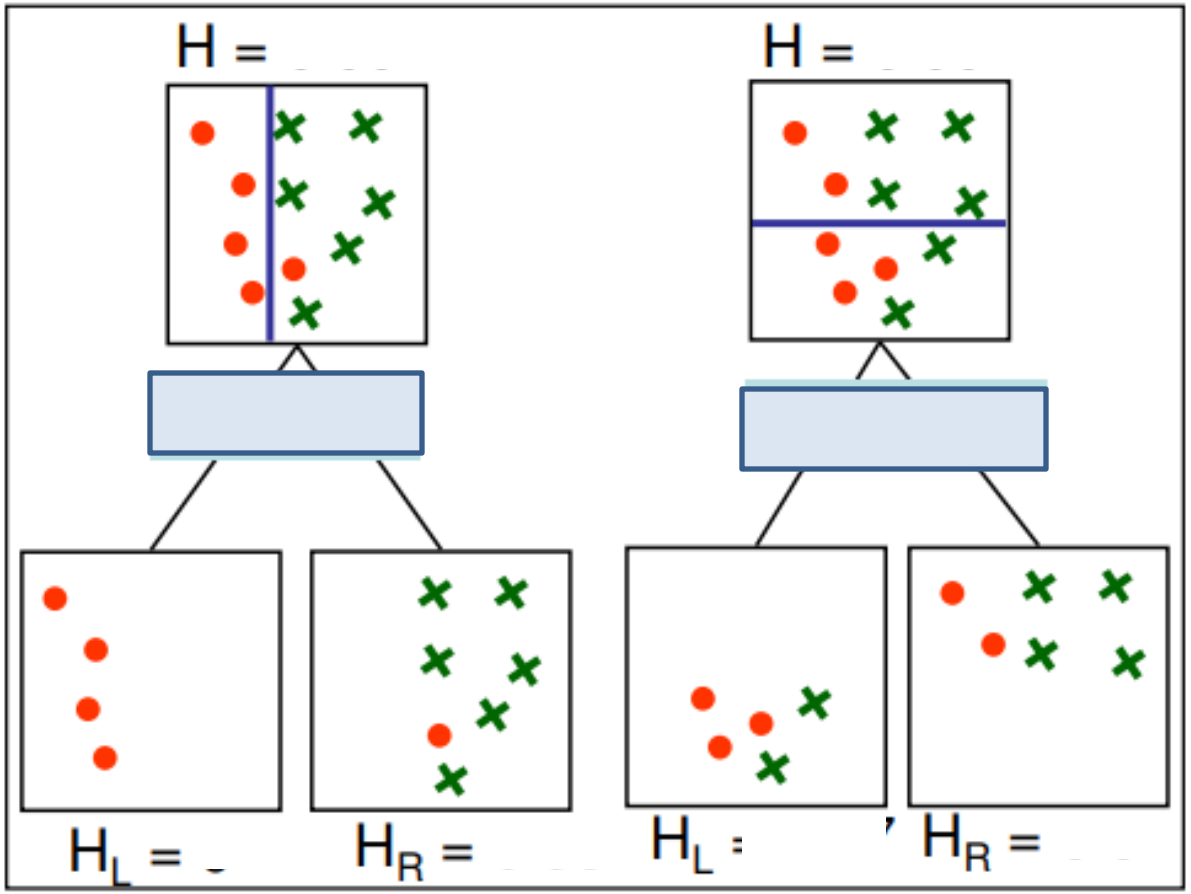
Example Information Gain



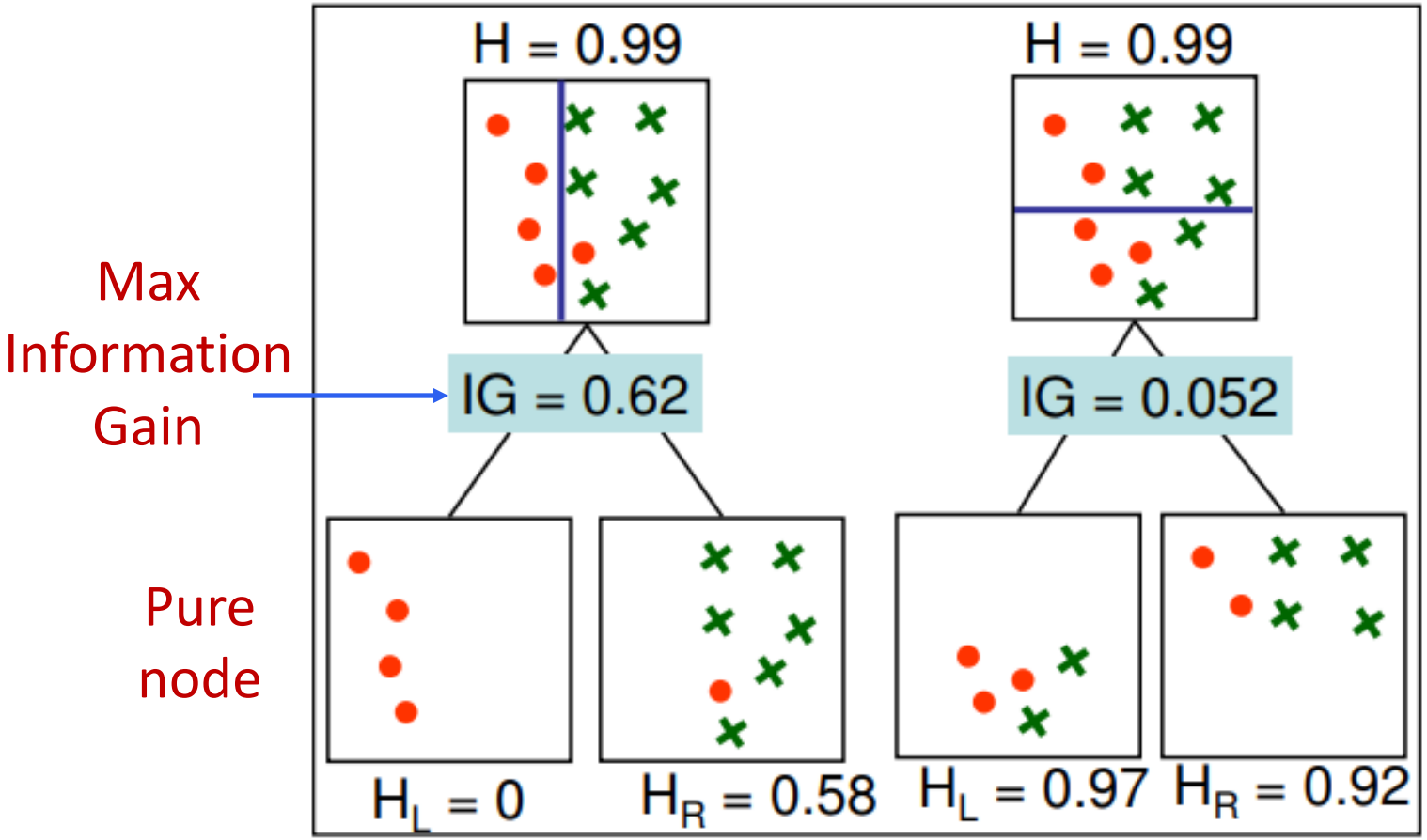
Example Information Gain



Example Information Gain



Example Information Gain



Learning Decision Trees

- Start from empty decision tree
- Split on **next best attribute (feature)**
  - Use, for example, information gain to select attribute:

$$\arg\max_i IG(X_i) = \arg\max_i H(Y) - H(Y \mid X_i)$$

- Recurse

ID3 algorithm uses Information Gain
Information Gain reduces uncertainty on Y

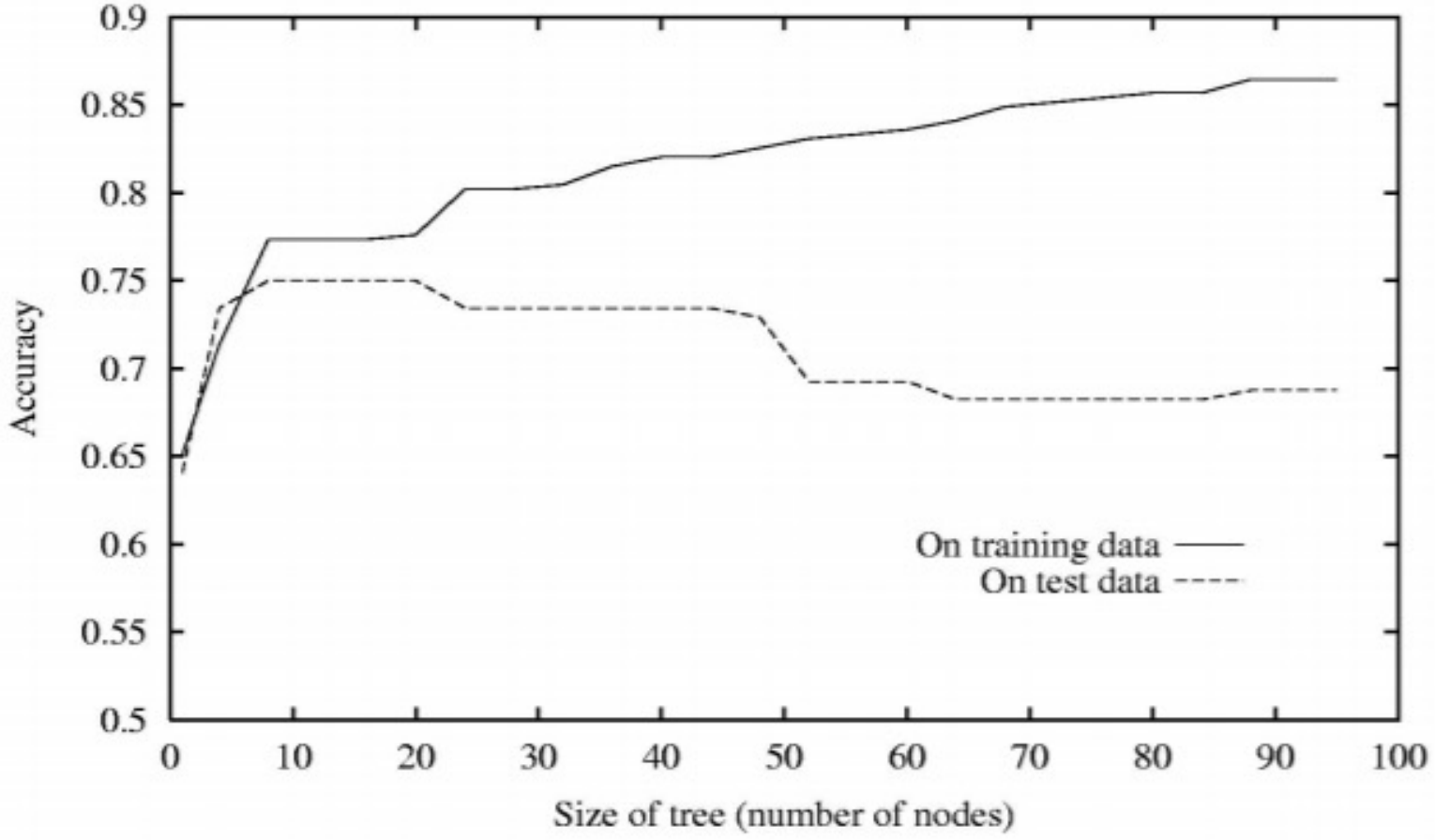
Impurity Metrics

Split a node according to max reduction of impurity

1. Entropy
2. Gini Index
   - For binary case with prob p_0, p_1:
     $$I(p_0, p_1) = 2p_0p_1 = 2p_0(1 - p_0)$$
   - For multi-class with prob $p_1, \dots, p_K$:
     $$I(p_1, \dots p_K) = \sum_{i=1}^{K} p_i \ (1 - p_i)$$

- Properties
  - Impurity metrics have value 0 for pure nodes
  - Impurity metrics are maximized for uniform distribution (nodes with most uncertainty)

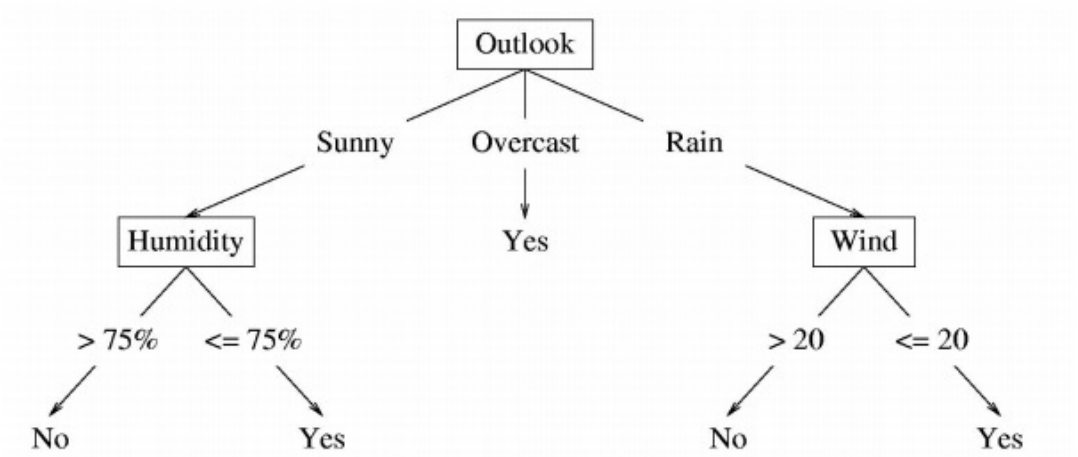
Overfitting



Solutions against Overfitting

- Standard decision trees have no learning bias
  - Training set error is always zero!
    - (If there is no label noise)
  - Lots of variance
  - Must introduce some bias towards simpler trees
- Many strategies for picking simpler trees
  - Fixed depth
  - Minimum number of samples per leaf
- Pruning
  - Remove branches of the tree that increase error using cross-validation

Real-valued Features



- Change to binary splits by choosing a threshold
- One method:
  - Sort instances by value, identify adjacencies with different classes

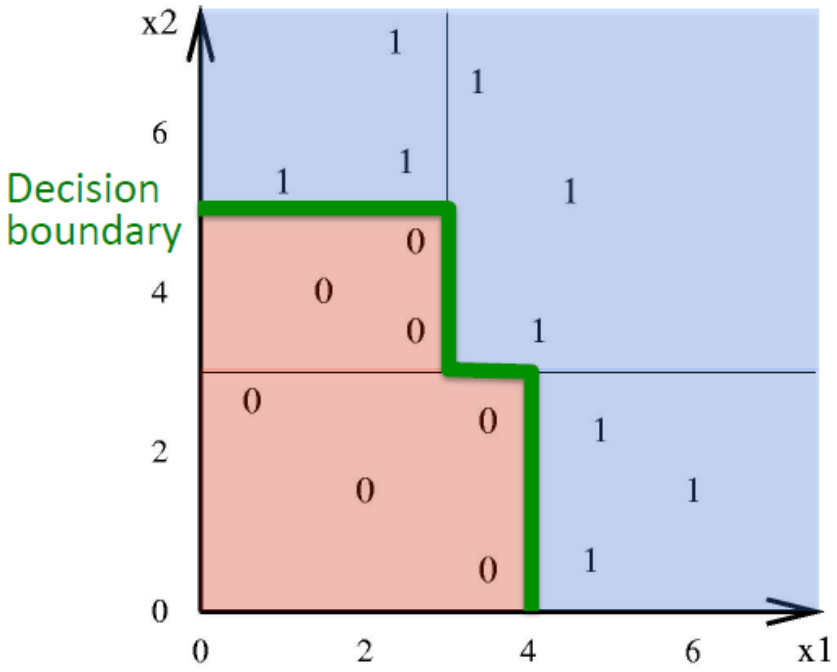
Humidity	40	48	60	72	80	90
PlayTennis:	No	No	Yes	Yes	Yes	No

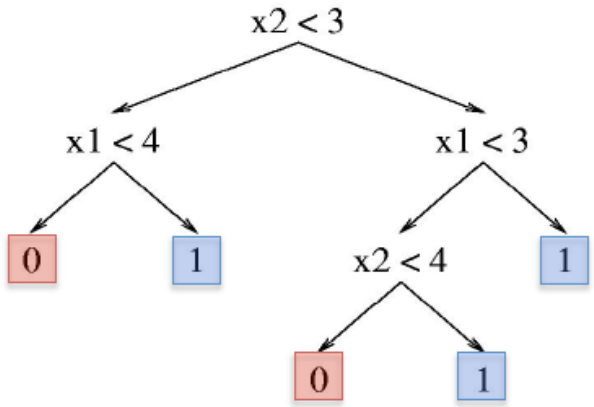
candidate splits

  - Choose among splits by InfoGain()

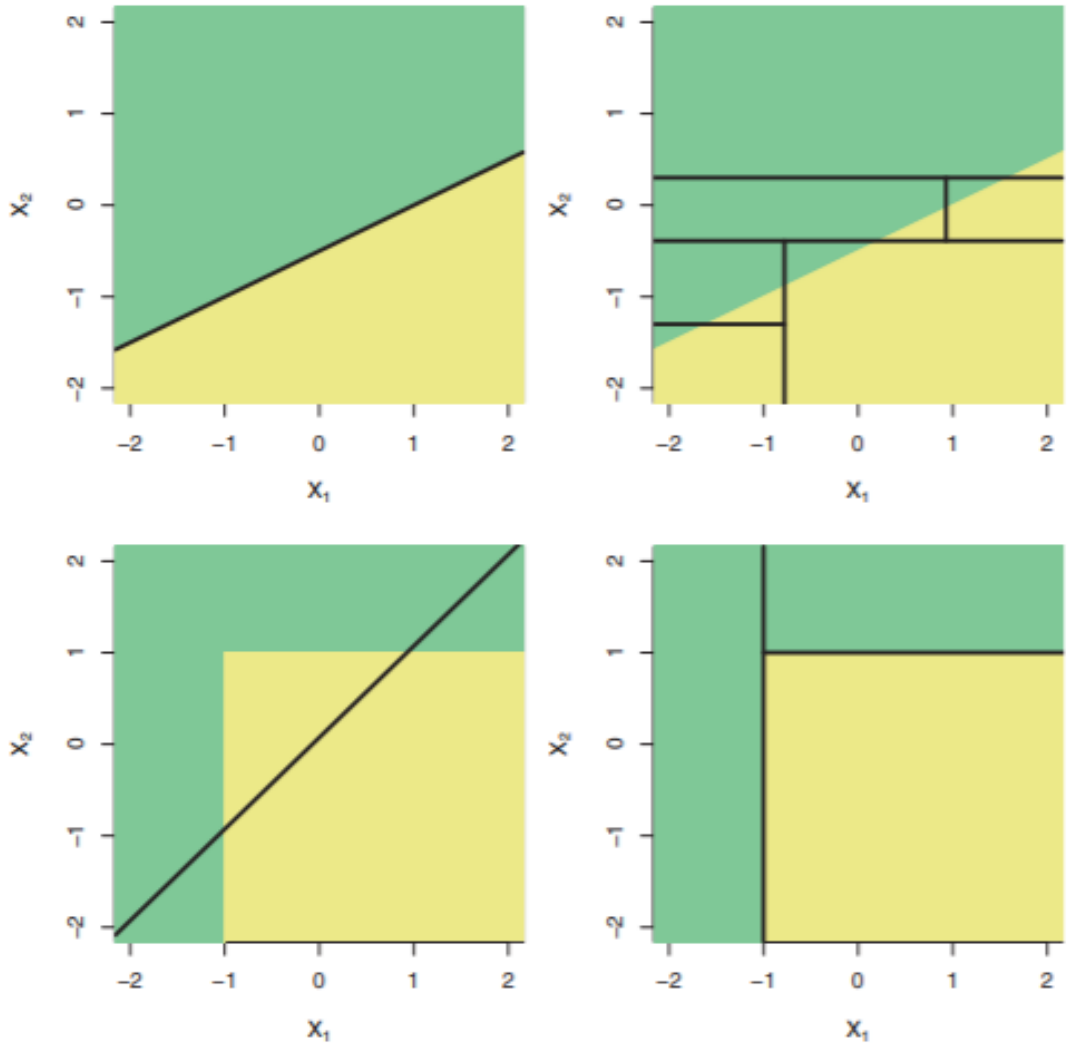
Decision Boundary

- Decision trees divide the feature space into axis-parallel (hyper-)rectangles
- Each rectangular region is labeled with one label





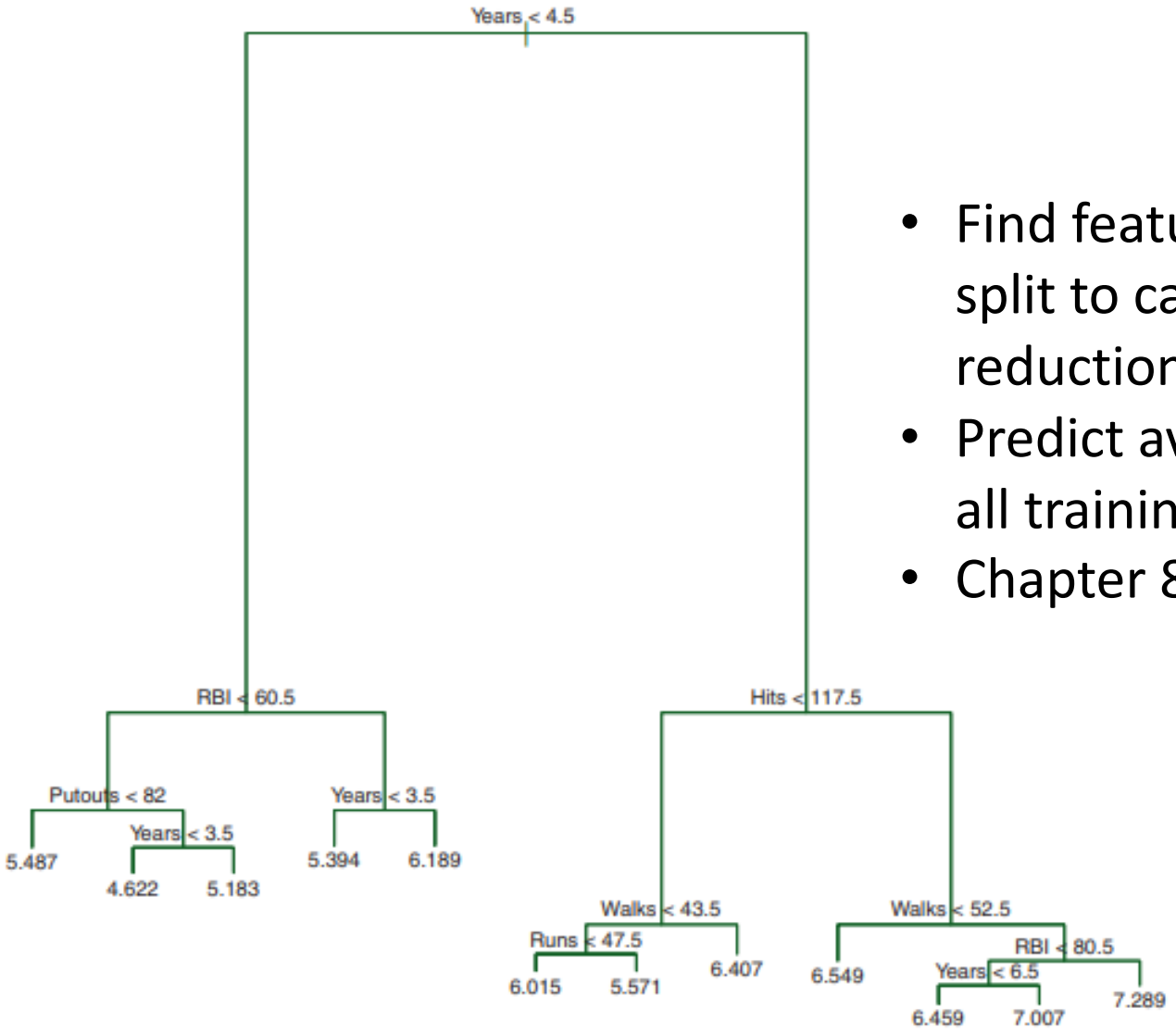
Decision Trees vs Linear Models



Linear model

Decision tree

Regression Trees

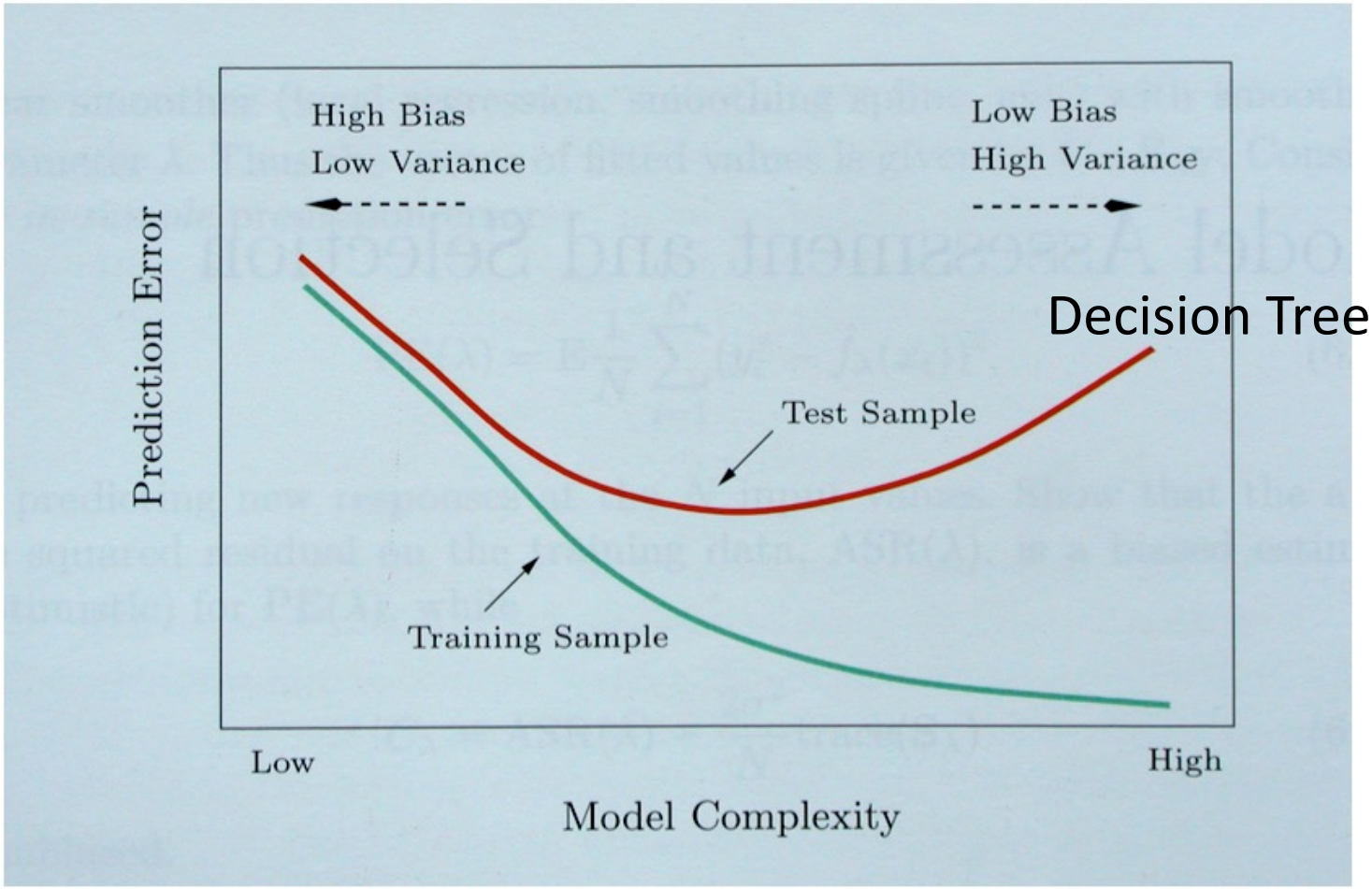


- Find feature and value to split to cause the maximum reduction in MSE
- Predict average response of all training data at each leaf
- Chapter 8.1 from textbook

Summary Decision Trees

- Greedy method for training
  - Not based on optimization or probabilities
- Uses impurity metric (e.g., information gain or Gini index) for splitting
- Advantages
  - Interpretability of decisions
- Limitations
  - Decision trees are prone to overfitting
  - Can be addressed by pruning or using ensembles of decision trees

Bias/Variance Tradeoff



Decision Trees

Hastie, Tibshirani, Friedman "Elements of Statistical Learning" 2001

How to reduce variance of single decision tree?