CS 383: Machine Learning

Prof Adam Poliak

Fall 2024

09/12/2024

Lecture 04

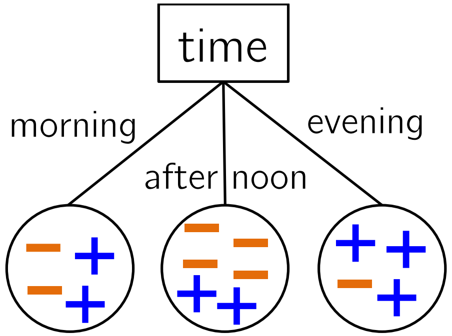
Quiz 2

1)
- internal nodes
- branches
- leaves

class labels
feature names
feature values

2) (a) +

(b) 5/14



3) high

4) zero/one loss

5) Case when no training examples have the feature value

Top-Down Decision Tree Algorithm

Dataset (X,y)

Features

```
MakeSubtree(D, F)
  if stopping criteria met
      make a leaf node N
      determine class label/probabilities for N
  else
      make an internal node N
      S = FindBestFeature(D, F)
      for each outcome k of S
          Dk = subset of instances that have outcome k
          N.child[k] = MakeSubtree(Dk,F-S)
  return subtree rooted at N
```

Why don't we want to use this feature again?

Design choice: stopping criteria

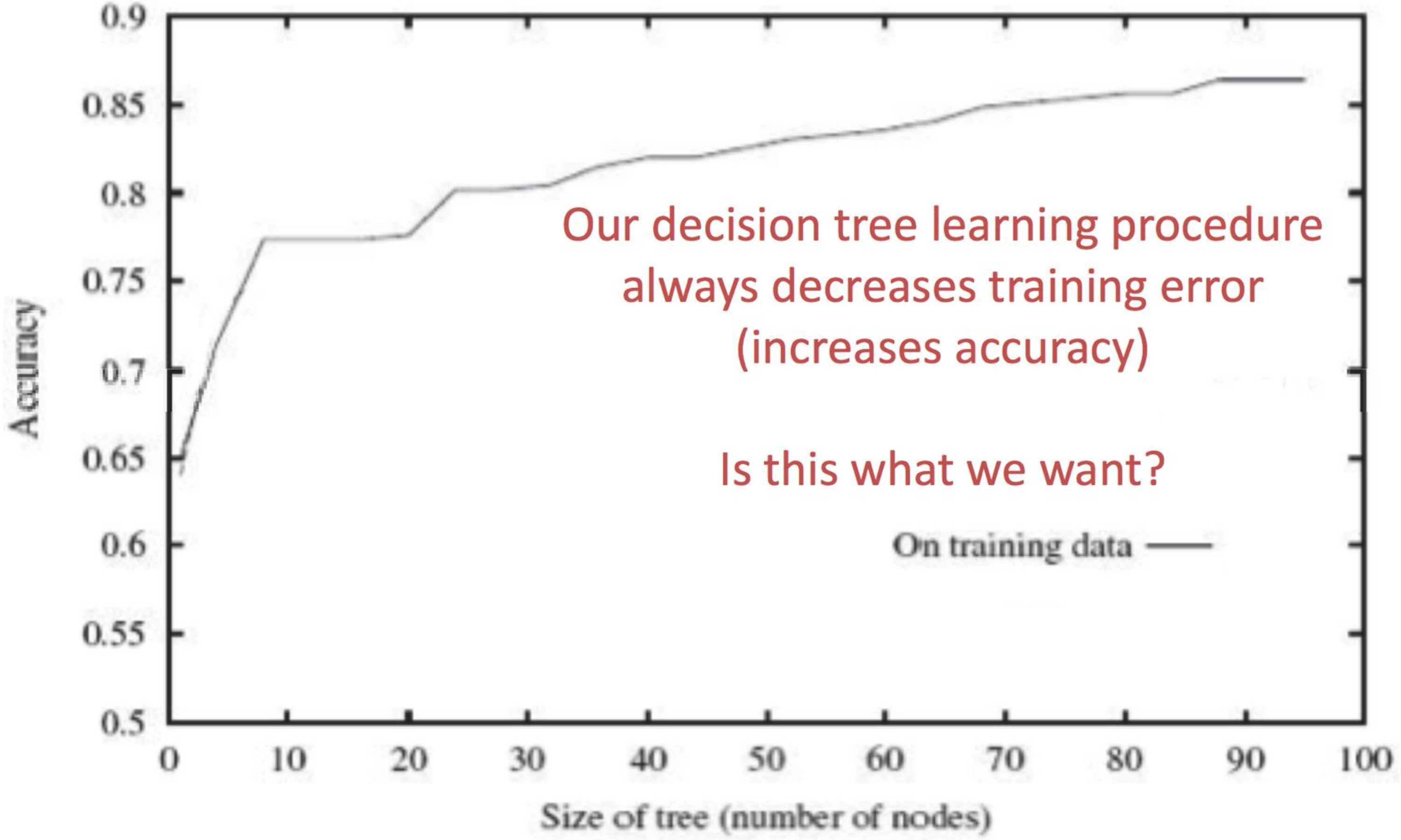
1. All the data points in our partition have the same label

2. No more features remain to split on

3. No features are informative about the label

4. Reached (user specified) max depth in the tree

Overfitting

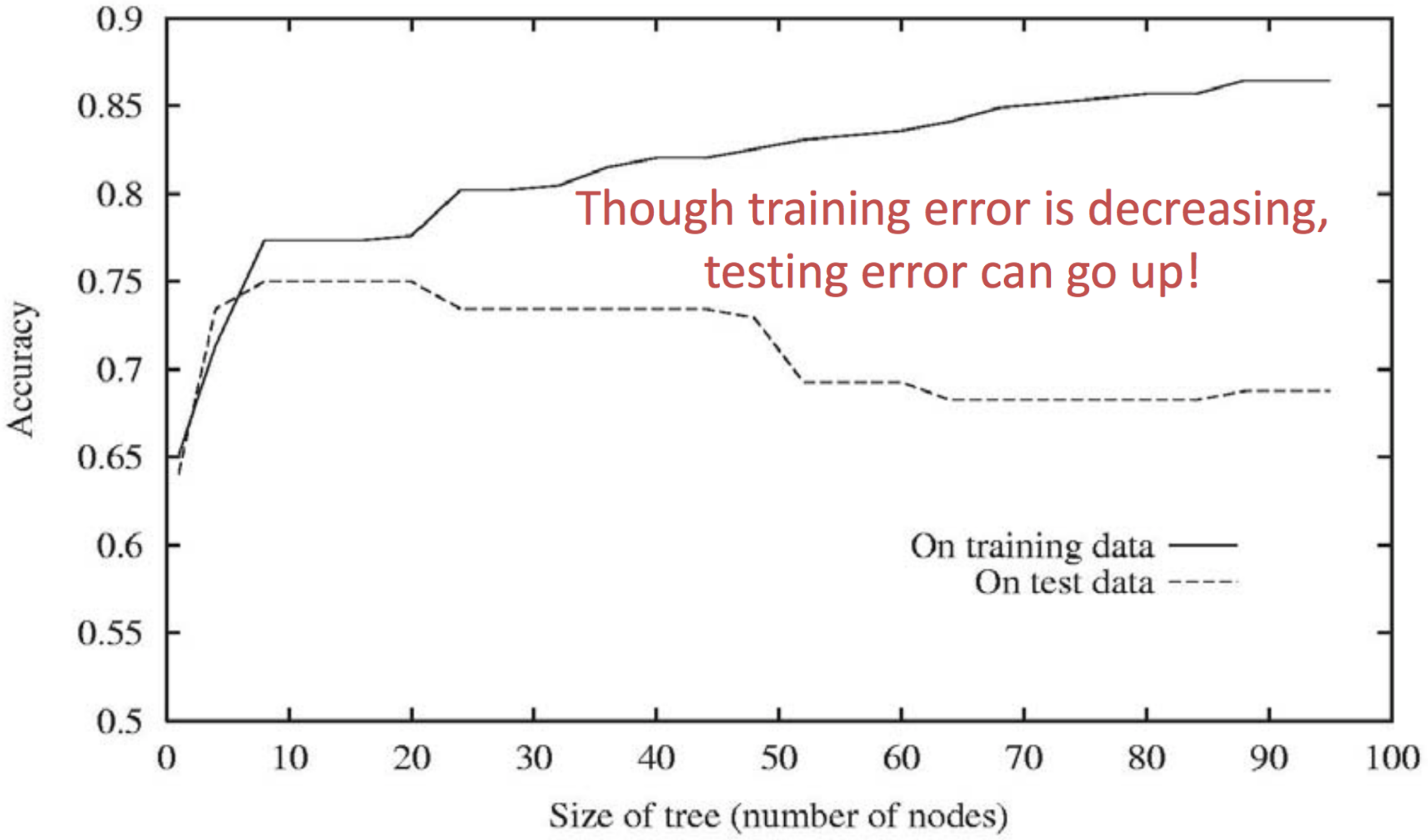


Our decision tree learning procedure always decreases training error (increases accuracy)

Is this what we want?

Slide: modified from Jessica Wu
Based on slide by David Kauchak
(originally by Pedro Domingos)

Overfitting



Slide: modified from Jessica Wu
Based on slide by David Kauchak
(originally by Pedro Domingos)

Overfitting definition

Consider a hypothesis (tree): h

- Training error: $error_{train}(h)$
- Error over all possible data: $error_D(h)$

Slide: modified from Ameet Soni

Overfitting definition

Consider a hypothesis (tree): h

- Training error: $error_{train}(h)$
- Error over all possible data: $error_D(h)$

A hypothesis h **overfits** training data if there exists another hypothesis h' s.t.

$$error_{train}(h) < error_{train}(h') \text{ AND } error_D(h) > error_D(h')$$

Slide: modified from Ameet Soni

Avoiding overfitting in decision trees

- Stop when leaf label reaches a certain fraction (i.e. 95% “yes”, 5% “no”)

HW2 implementation

- Set a maximum depth for the tree

- Set a minimum number of examples in leaf (i.e. if we have a 2-1 split, stop)

How to select “best” features?

X			Y
Color	**Shape**	**Size**	**Likes toy?**
red	square	big	+
blue	square	big	+
red	circle	small	-
blue	square	small	-
red	circle	big	+

Example by: Ameet Soni

How to select “best” features?

X			Y
Color	**Shape**	**Size**	**Likes toy?**
red	square	big	+
blue	square	big	+
red	circle	**big**	-
blue	square	**big**	-
red	circle	big	+

Example by: Ameet Soni

Most Informative Feature

Goal: Which feature will allow us to learn the most information

Information Gain – metric from information theory

“*the amount of information gained about a random variable or signal from observing another random variable*”

Wikipedia

Entropy

measure of uncertainty in a group of observations

Which feature has the highest IG

X			Y
Color	Shape	Size	Likes toy?
red	square	big	+
blue	square	big	+
red	circle	small	-
blue	square	small	-
red	circle	big	+

Example by: Ameet Soni

Size and entropy

X

Color	Shape	Size
red	square	big
blue	square	big
red	circle	small
blue	square	small
red	circle	big

Y

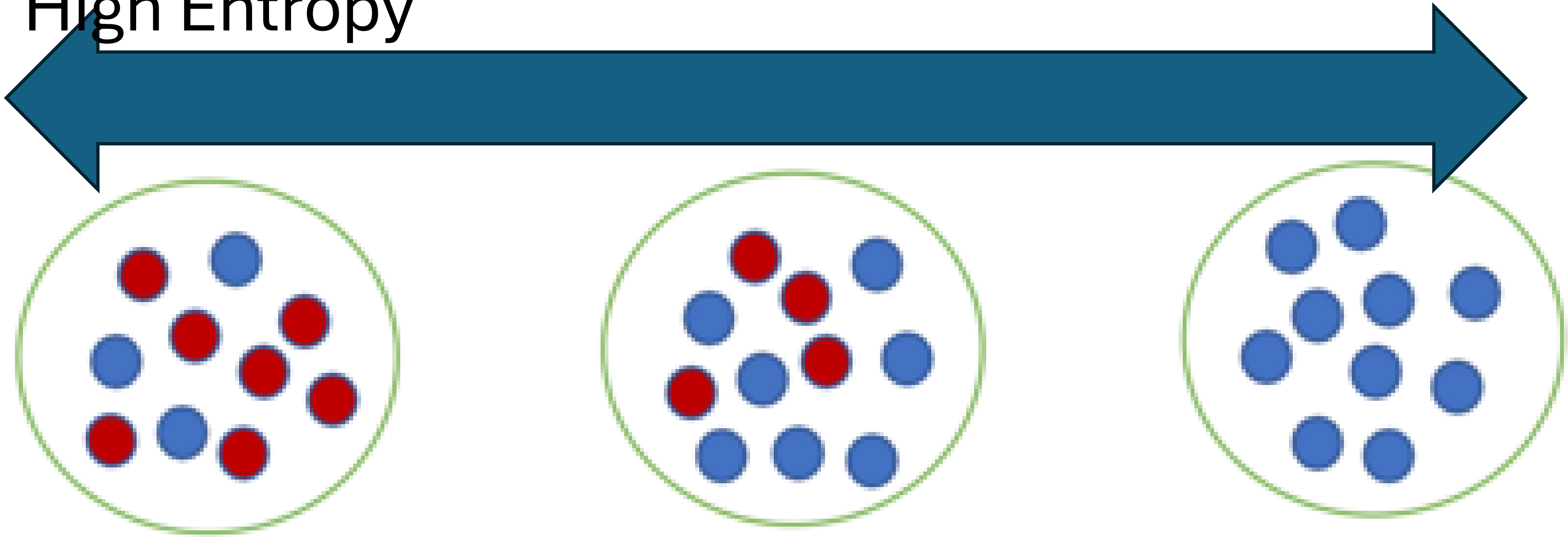
Likes toy?
+
+
-
-
+

Example by: Ameet Soni

Entropy & Information Gain

Low IG
High Entropy

High IG
Low Entropy



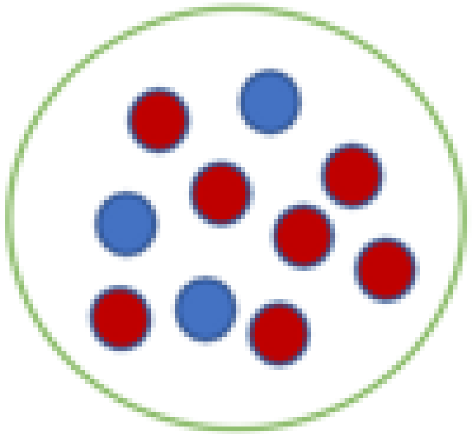
$p(red) = .7$
$p(blue) = .3$

$p(red) = .4$
$p(blue) = .6$

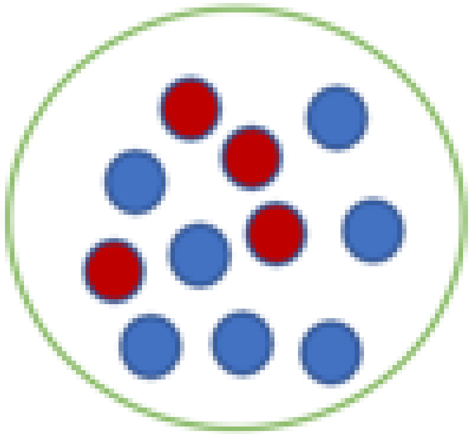
$p(red) = 0$
$p(blue) = 1$

Quantify Entropy

$$Entropy: \sum_{i}^{C} p_i \log(p_i)$$



$p(red) = .7$
$p(blue) = .3$

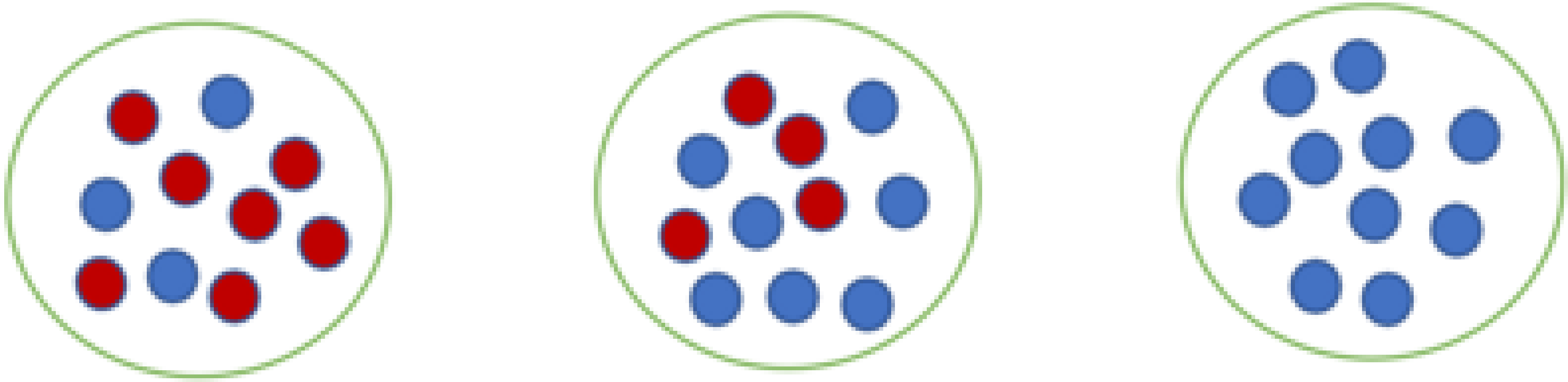


$p(red) = .4$
$p(blue) = .6$

$p(red) = 0$
$p(blue) = 1$

Quantify Entropy

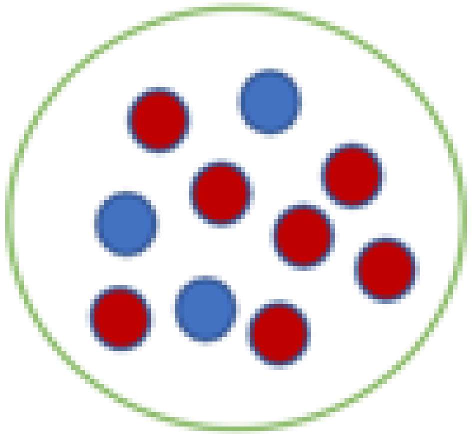
$$Entropy: \sum_{i}^{C} p_i \log(p_i)$$



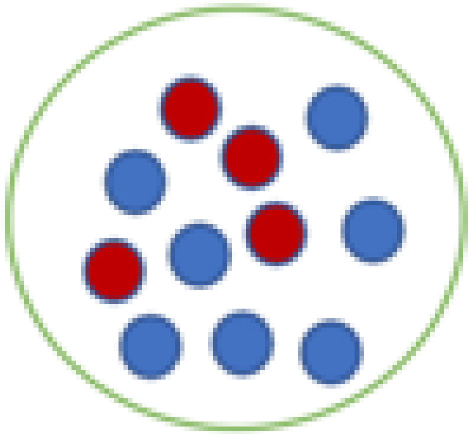
-0.265 $\quad$ -0.292 $\quad$ 0

Quantify Entropy

$$Entropy: -\sum_{i}^{C} p_i \log(p_i)$$



0.265



0.292

0

Conditional Entropy

X

Color	Shape	Size
red	square	big
blue	square	big
red	circle	small
blue	square	small
red	circle	big

Y

Likes toy?
+
+
-
-
+

Example by: Ameet Soni

Conditional Entropy

X

Color	Shape	Size
red	square	big
blue	square	big
red	circle	small
blue	square	small
red	circle	big

Y

Likes toy?
+
+
-
-
+

$$H\ (Y \mid X = v\ ): \sum_{i}^{C} p(Y_i \mid X = v)\ \log(p(Y_i \mid X = v)$$

Example by: Ameet Soni

Conditional Entropy

$$H\ (Y \mid X = v\ ):\ \sum_{i}^{C} p(y_i \mid X = v)\ \log(p(y_i \mid X = v)$$

$$H\ (Y \mid X):\ \sum_{v \in dom(X)}^{|X|} p(X = v) H\ (Y|X = v)$$

X

Y

Color	Shape	Size	Likes toy?
red	square	big	+
blue	square	big	+
red	circle	small	-
blue	square	small	-
red	circle	big	+

Information Gain

$$H\ (Y \mid X = v): \sum_{i}^{C} p(y_i \mid X = v)\ \log(p(y_i \mid X = v)$$

$$H\ (Y \mid X): \sum_{v \in dom(X)}^{|X|} p(X = v) H\ (Y | X = v)$$

$$IG\ (Y, X): H(Y) - H(Y|X)$$

X				Y
Color	**Shape**	**Size**		**Likes toy?**
red	square	big		+
blue	square	big		+
red	circle	small		-
blue	square	small		-
red	circle	big		+