DECISION TREES

David Kauchak
CS 451 – Fall 2013

Admin

Assignment 1 available and is due on Friday (printed out at the beginning of class)

Door code for MBH632

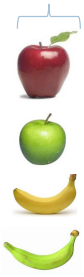
Keep up with the reading

Videos

Quick refresher from last time…

Representing examples

examples



What is an example?
How is it represented?

Features

examples

features

$f_1, f_2, f_3, \ldots, f_n$

$f_1, f_2, f_3, \ldots, f_n$

$f_1, f_2, f_3, \ldots, f_n$

$f_1, f_2, f_3, \ldots, f_n$

How our algorithms actually "view" the data

Features are the questions we can ask about the examples

Features

examples

features

red, round, leaf, 3oz, …

green, round, no leaf, 4oz, …

yellow, curved, no leaf, 4oz, …

green, curved, no leaf, 5oz, …

How our algorithms actually "view" the data

Features are the questions we can ask about the examples

Classification revisited

examples	label
red, round, leaf, 3oz, …	apple
green, round, no leaf, 4oz, …	apple
yellow, curved, no leaf, 4oz, …	banana
green, curved, no leaf, 5oz, …	banana

learn → model/ classifier

During learning/training/induction, learn a model of what distinguishes apples and bananas *based on the features*

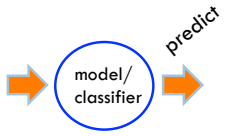
Classification revisited

red, round, no leaf, 4oz, … → model/ classifier → predict → Apple or banana?

The model can then classify a new example *based on the features*

Classification revisited

red, round, no leaf, 4oz, …



Apple

Why?

The model can then classify a new example *based on the features*

Classification revisited

Training data		Test set
examples	label	
red, round, leaf, 3oz, …	apple	
green, round, no leaf, 4oz, …	apple	red, round, no leaf, 4oz, … ?
yellow, curved, no leaf, 4oz, …	banana	
green, curved, no leaf, 5oz, …	banana	

Classification revisited

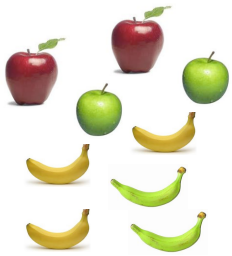
Training data		Test set
examples	label	
red, round, leaf, 3oz, …	apple	
green, round, no leaf, 4oz, …	apple	red, round, no leaf, 4oz, … ?
yellow, curved, no leaf, 4oz, …	banana	
green, curved, no leaf, 5oz, …	banana	

Learning is about ***generalizing*** from the training data

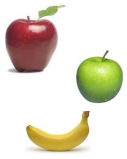
What does this assume about the training and test set?

Past predicts future

Training data

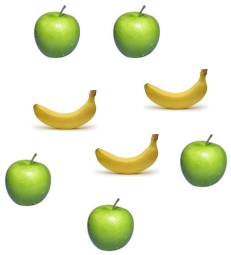


Test set

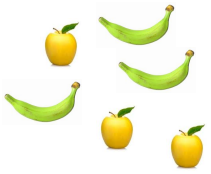


Past predicts future

Training data



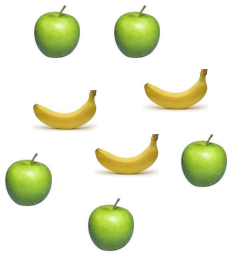
Test set



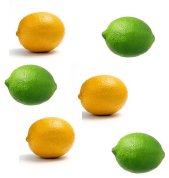
Not always the case, but we'll often assume it is!

Past predicts future

Training data



Test set



Not always the case, but we'll often assume it is!

More technically…

We are going to use the *probabilistic model* of learning

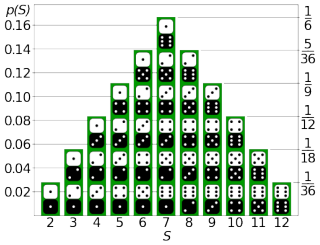
There is some probability distribution over example/label pairs called the *data generating distribution*

Both the training data **and** the test set are generated based on this distribution

What is a probability distribution?

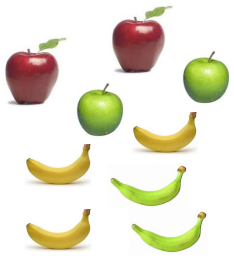
Probability distribution

Describes how likely (i.e. probable) certain events are



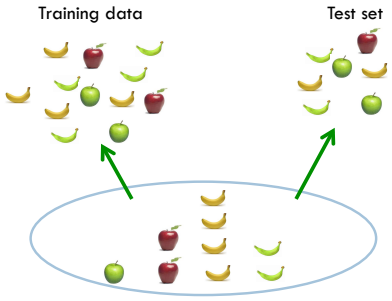
Probability distribution

Training data

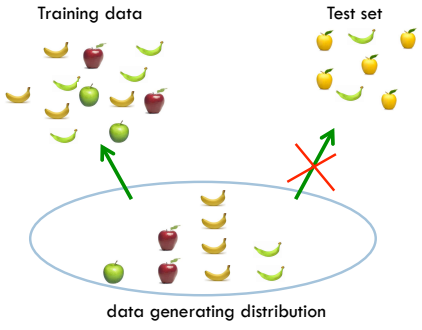


High probability	Low probability
round apples	curved apples
curved bananas	red bananas
apples with leaves	yellow apples
…	…

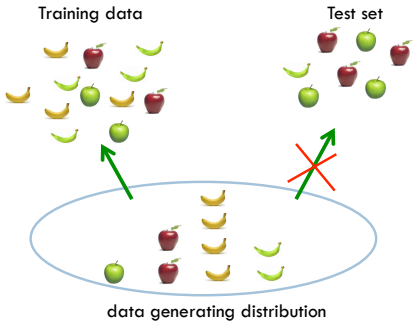
data generating distribution



data generating distribution



data generating distribution



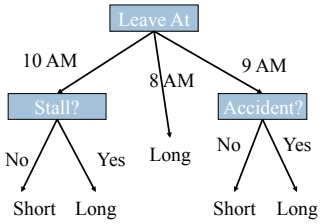
A sample data set

Features				Label
Hour	Weather	Accident	Stall	Commute
8 AM	Sunny	No	No	Long
8 AM	Cloudy	No	Yes	Long
10 AM	Sunny	No	No	Short
9 AM	Rainy	Yes	No	Long
9 AM	Sunny	Yes	Yes	Long
10 AM	Sunny	No	No	Short
10 AM	Cloudy	No	No	Short
9 AM	Sunny	Yes	No	Long
10 AM	Cloudy	Yes	Yes	Long
10 AM	Rainy	No	No	Short
8 AM	Cloudy	Yes	No	Long
9 AM	Rainy	No	No	Short

8 AM, Rainy, Yes, No?
10 AM, Rainy, No, No?

Can you describe a "model" that could be used to make decisions in general?

Decision trees

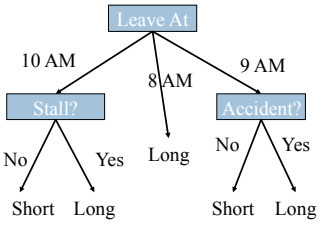


Tree with internal nodes labeled by features

Branches are labeled by tests on that feature

Leaves labeled with classes

Decision trees



Tree with internal nodes labeled by features

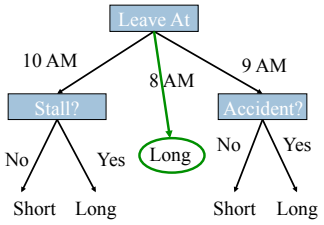
Branches are labeled by tests on that feature

Leaves labeled with classes

Leave = 8 AM
Weather = Rainy

Accident = Yes
Stall = No

Decision trees



Tree with internal nodes labeled by features

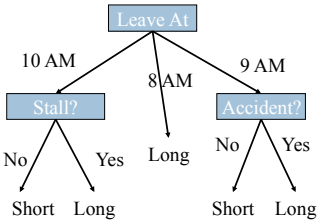
Branches are labeled by tests on that feature

Leaves labeled with classes

Leave = 8 AM
Weather = Rainy

Accident = Yes
Stall = No

Decision trees



Tree with internal nodes labeled by features

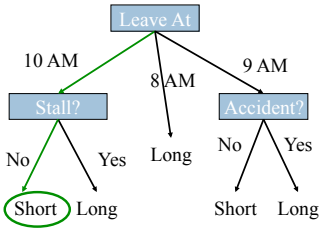
Branches are labeled by tests on that feature

Leaves labeled with classes

Leave = 10 AM
Weather = Rainy

Accident = No
Stall = No

Decision trees



Tree with internal nodes labeled by features

Branches are labeled by tests on that feature

Leaves labeled with classes

Leave = 10 AM
Weather = Rainy

Accident = No
Stall = No

To ride or not to ride, that is the question…

Terrain	Unicycle-type	Weather	Go-For-Ride?
Trail	Normal	Rainy	NO
Road	Normal	Sunny	YES
Trail	Mountain	Sunny	YES
Road	Mountain	Rainy	YES
Trail	Normal	Snowy	NO
Road	Normal	Rainy	YES
Road	Mountain	Snowy	YES
Trail	Normal	Sunny	NO
Road	Normal	Snowy	NO
Trail	Mountain	Snowy	YES

Build a decision tree

Recursive approach

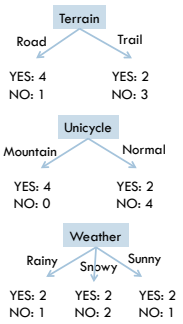
Base case: If all data belong to the same class, create a leaf node with that label

Otherwise:

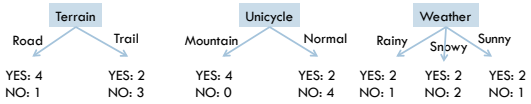
- calculate the "score" for each feature if we used it to split the data
- pick the feature with the highest score, partition the data based on that data value and call recursively

Partitioning the data

Terrain	Unicycle-type	Weather	Go-For-Ride?
Trail	Normal	Rainy	NO
Road	Normal	Sunny	YES
Trail	Mountain	Sunny	YES
Road	Mountain	Rainy	YES
Trail	Normal	Snowy	NO
Road	Normal	Rainy	YES
Road	Mountain	Snowy	YES
Trail	Normal	Sunny	NO
Road	Normal	Snowy	NO
Trail	Mountain	Snowy	YES



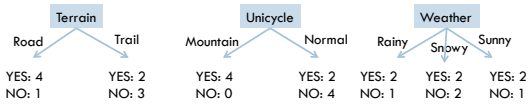
Partitioning the data



calculate the "score" for each feature
if we used it to split the data

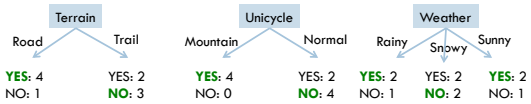
What score should we use?
If we just stopped here, which tree would be best?
How could we make these into decision trees?

Decision trees

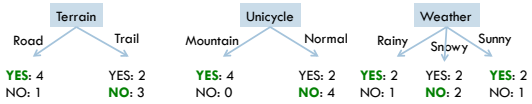


How could we make these into decision trees?

Decision trees



Decision trees

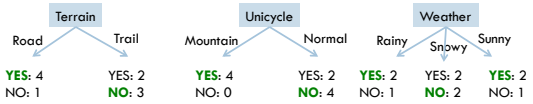


Training error: the average error over the training set

For classification, the most common "error" is the number of mistakes

Training error for each of these?

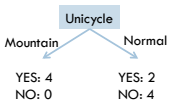
Decision trees



3/10 2/10 4/10

Training error: the average error over the training set

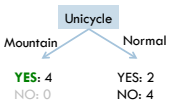
Recurse



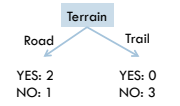
Terrain	Unicycle-type	Weather	Go-For-Ride?
Trail	Mountain	Sunny	YES
Road	Mountain	Rainy	YES
Road	Mountain	Snowy	YES
Trail	Mountain	Snowy	YES

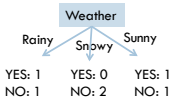
Terrain	Unicycle-type	Weather	Go-For-Ride?
Trail	Normal	Rainy	NO
Road	Normal	Sunny	YES
Trail	Normal	Snowy	NO
Road	Normal	Rainy	YES
Trail	Normal	Sunny	NO
Road	Normal	Snowy	NO

Recurse

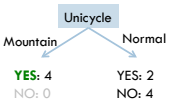


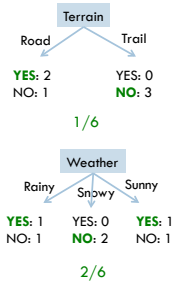
Terrain	Unicycle-type	Weather	Go-For-Ride?
Trail	Normal	Rainy	NO
Road	Normal	Sunny	YES
Trail	Normal	Snowy	NO
Road	Normal	Rainy	YES
Trail	Normal	Sunny	NO
Road	Normal	Snowy	NO





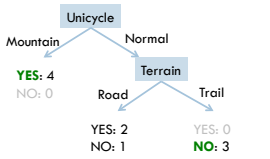
Recurse





Terrain	Unicycle-type	Weather	Go-For-Ride?
Trail	Normal	Rainy	NO
Road	Normal	Sunny	YES
Trail	Normal	Snowy	NO
Road	Normal	Rainy	YES
Trail	Normal	Sunny	NO
Road	Normal	Snowy	NO

Recurse



Terrain	Unicycle-type	Weather	Go-For-Ride?
Road	Normal	Sunny	YES
Road	Normal	Rainy	YES
Road	Normal	Snowy	NO

Recurse