MACHINE LEARNING BASICS

David Kauchak
CS159 Fall 2020

1

Admin

Assignment 6a

2

Machine Learning is...

Machine learning, a branch of artificial intelligence, concerns the construction and study of systems that can learn from data.

WIKIPEDIA
The Free Encyclopedia

3

Machine Learning is...

Machine learning is programming computers to optimize a performance criterion using example data or past experience.

-- Ethem Alpaydin

The goal of machine learning is to develop methods that can automatically detect patterns in data, and then to use the uncovered patterns to predict future data or other outcomes of interest.

-- Kevin P. Murphy

The field of pattern recognition is concerned with the automatic discovery of regularities in data through the use of computer algorithms and with the use of these regularities to take actions.

-- Christopher M. Bishop
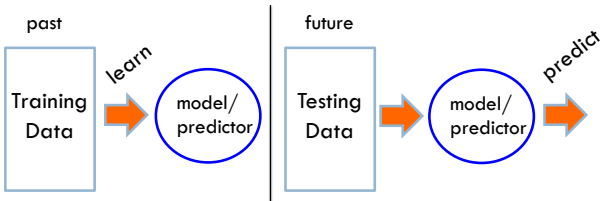
4

Machine Learning is…

Machine learning is about predicting the future based on the past.
-- Hal Daume III

5

Machine Learning is…

Machine learning is about predicting the future based on the past.
-- Hal Daume III

6

Why machine learning?

Lot's of data

Hand-written rules just don't do it

Performance can be much better than what people can do

Why not just study machine learning?

- Domain knowledge/expertise is still very important
- What types of features to use
- What models are important

7

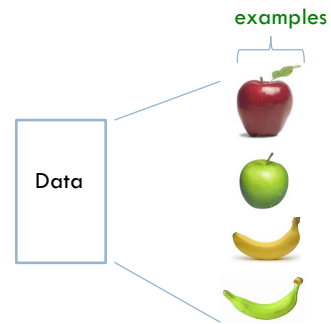## Why machine learning?

Be able to laugh at these signs

8

Machine learning problems

What high-level machine learning problems have you seen or heard of before?
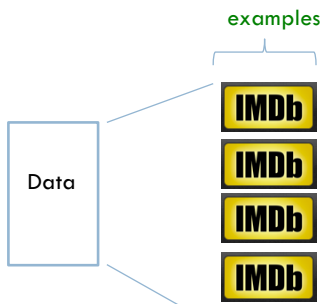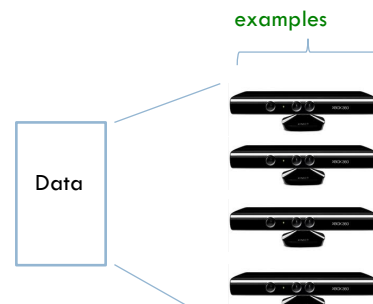
9

Data

examples

Data

10

Data
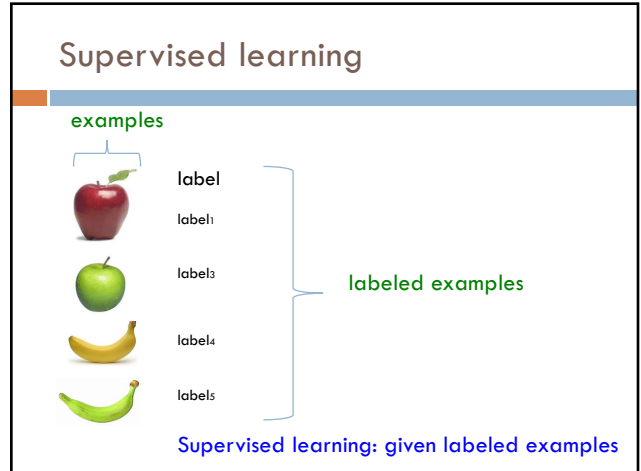
examples

Data

IMDb
IMDb
IMDb
IMDb

11

Data

examples

Data

12

Data

examples

Data

13

Supervised learning

examples

label
label1
label3
label4
label5

labeled examples

Supervised learning: given labeled examples

14

Supervised learning

label
label1
label3
label4
label5

model/
predictor

Supervised learning: given labeled examples

15

Supervised learning

model/
predictor

predicted label

Supervised learning: learn to predict new example
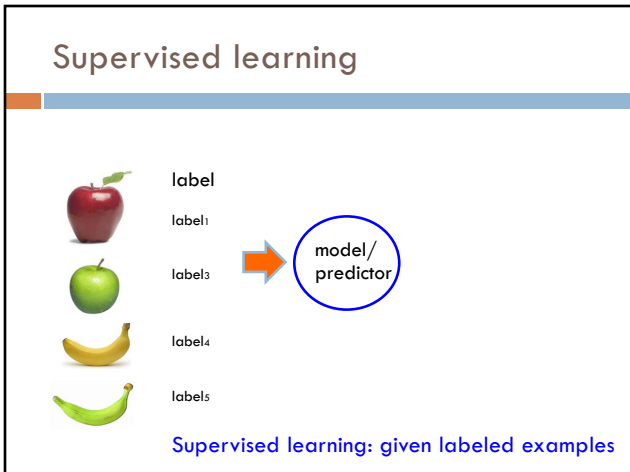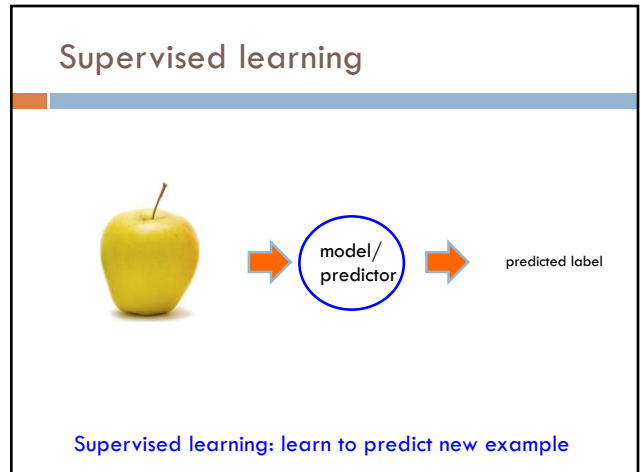
16

Supervised learning: classification

label

apple

apple

banana

banana

Classification: a finite set of labels

Supervised learning: given labeled examples

17

NLP classification applications

Document classification
- spam
- sentiment analysis
- topic classification

Turn SafeSearch on or off
https://support.**google**.com/web**search**/answer/510
1. Visit the Search Settings page.
2. In the "SafeSearch filters" section, select or unselect **Filter explicit results**.
3. Click **Save** at the bottom of the page.

Does linguistics phenomena X occur in text Y?

Digit recognition

Grammatically correct or not?

Word sense disambiguation

Any question you can pose as to have a discrete set of labels/answers!

18

Supervised learning: regression

label

-4.5

10.1

3.2

4.3

Regression: label is real-valued
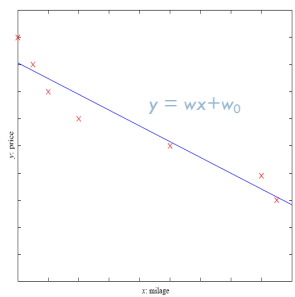
Supervised learning: given labeled examples

19

Regression Example

Price of a used car

x : car attributes
(e.g. mileage)

y : price

$y = wx + w_0$

20

Regression applications

How many clicks will a particular website, ad, etc. get?

Predict the readability level of a document
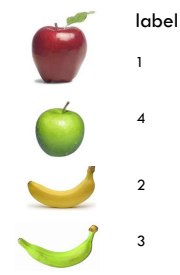
Predict pause between spoken sentences?

Economics/Finance: predict the value of a stock

Car/plane navigation: angle of the steering wheel, acceleration, …

…

21

Supervised learning: ranking

label

1

4

2

3

Ranking: label is a ranking

Supervised learning: given labeled examples

22

NLP Ranking Applications

reranking N-best output lists (e.g. parsing, machine translation, …)

Rank possible simplification options
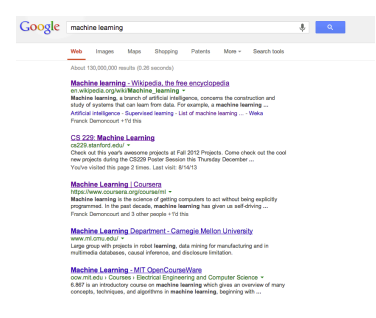
flight search (search in general)

…

23

Ranking example

Given a query and a set of web pages, rank them according to relevance

24

Unsupervised learning

Unsupervised learning: given data, i.e. examples, but no labels

25

Unsupervised learning applications

learn clusters/groups without any label

- cluster documents
- cluster words (synonyms, parts of speech, …)

compression

bioinformatics: learn motifs

…

26

Reinforcement learning

left, right, straight, left, left, left, straight	GOOD
left, straight, straight, left, right, straight, straight	BAD
left, right, straight, left, left, left, straight	18.5
left, straight, straight, left, right, straight, straight	-3

Given a *sequence* of examples/states and a *reward* after completing that sequence, learn to predict the action to take in for an individual example/state

27

Reinforcement learning example

Backgammon

Given sequences of moves and whether or not the player won at the end, learn to make good moves

28

Reinforcement learning example

https://www.youtube.com/watch?v=tXlM99xPQC8

29

Other learning variations

What data is available:
- Supervised, unsupervised, reinforcement learning
- semi-supervised, active learning, …

How are we getting the data:
- online vs. offline learning

Type of model:
- generative vs. discriminative
- parametric vs. non-parametric

30

Text classification

label

spam

For this class, I'm mostly going to focus on classification

I'll use text classification as a running example

not spam

not spam

31

Representing examples

examples

What is an example?
How is it represented?

32

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

33

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

34

Text: raw data

Raw data   Features?

35

Feature examples

Raw data   Features

Clinton said banana repeatedly last week on tv, "banana, banana, banana"

(1, 1, 1, 0, 0, 1, 0, 0, …)

banana clinton said california across tv wrong capital

Occurrence of words (unigrams)

36

Feature examples

Raw data

Features

Clinton said banana repeatedly last week on tv, "banana, banana, banana"

(4, 1, 1, 0, 0, 1, 0, 0, ...)

banana clinton said california across tv wrong capital

Frequency of word occurrence (unigram frequency)

37

Feature examples

Raw data

Features

Clinton said banana repeatedly last week on tv, "banana, banana, banana"

(1, 1, 1, 0, 0, 1, 0, 0, ...)

banana repeatedly clinton said said banana california schools across the tv banana wrong way capital city

Occurrence of bigrams

38

Feature examples

Raw data

Features

Clinton said banana repeatedly last week on tv, "banana, banana, banana"

(1, 1, 1, 0, 0, 1, 0, 0, ...)

banana repeatedly clinton said said banana california schools across the tv banana wrong way capital city

Other features?

39

Lots of other features

POS: occurrence, counts, sequence

Constituents

Whether 'V1agra' occurred 15 times

Whether 'banana' occurred more times than 'apple'

If the document has a number in it

…

Features are very important, but we're going to focus on the model

40

10

Slide 41 — 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*

41

Slide 42 — 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*

42

Slide 43 — Classification revisited

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

Why?

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

43

Slide 44 — Classification revisited

Training data

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

Test set

red, round, no leaf, 4oz, … ?

44

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	Learning is about *generalizing* from the training data
green, curved, no leaf, 5oz, …	banana	What does this assume about the training and test set?

45

Past predicts future

Training data   Test set

46

Past predicts future

Training data   Test set

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

47

Past predicts future

Training data   Test set

Not always the case, but we'll often assume it is!
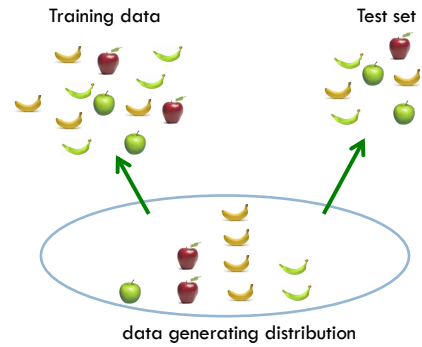
48

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*

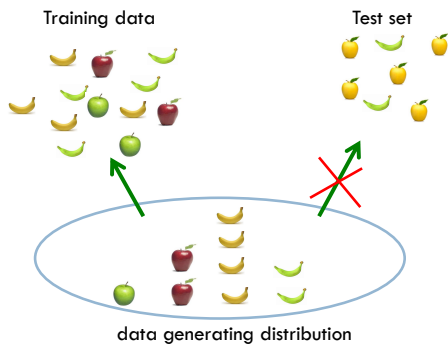**Both** the training data **and** the test set are generated based on this distribution

49

data generating distribution

Training data

Test set

data generating distribution

50

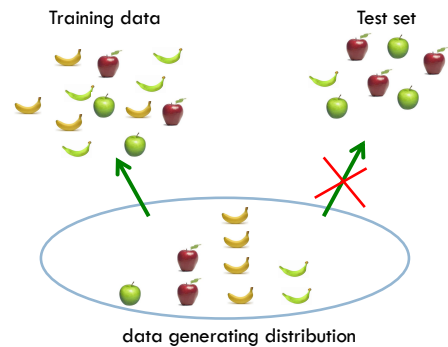## data generating distribution

Training data

Test set

data generating distribution

51

data generating distribution

Training data

Test set

data generating distribution

52