

MACHINE LEARNING
BASICS

David Kauchak
CS159 Fall 2024

1

Admin


Assignment 6: due Sunday

Quiz 3 next Thursday

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...

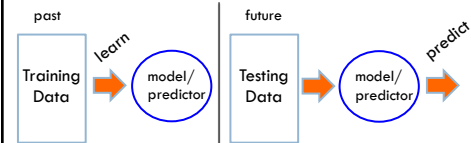
Machine learning is about predicting the future based on the past.
-- Hal Daume III



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?

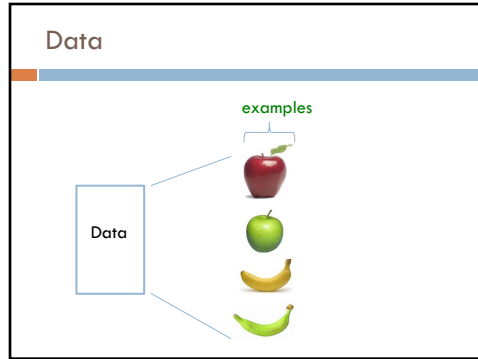
- Domain knowledge/expertise is still very important
- What types of features to use
- What models are important

7

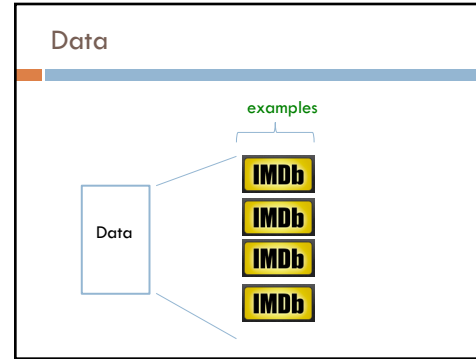
Machine learning problems

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

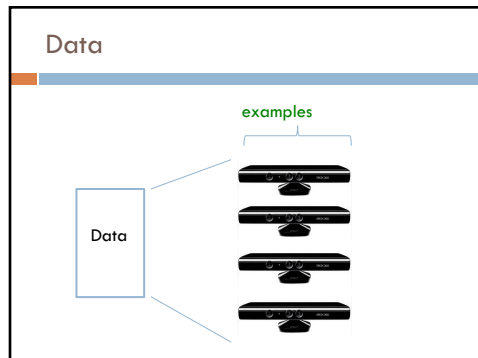
8



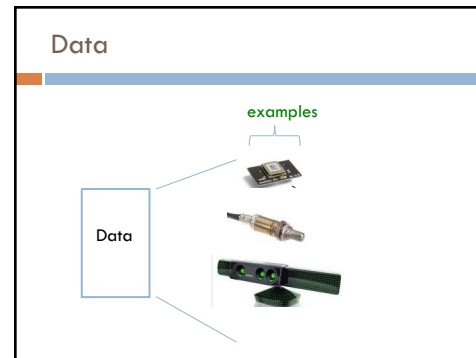
9



10



11



12

Supervised learning

examples

label
label1
label2
label4
label5

labeled examples

Supervised learning: given labeled examples

13

Supervised learning

label
label1
label2
label4
label5

model/
predictor

Supervised learning: given labeled examples

14

Supervised learning

model/
predictor

predicted label

Supervised learning: learn to predict new example

15

Supervised learning: classification

label
apple
apple
banana
banana

Classification: a finite set of labels

Supervised learning: given labeled examples

16





NLP classification applications

- Document classification
 - spam
 - sentiment analysis
 - topic classification
- Turn SafeSearch on or off
<https://support.google.com/websearch/answer/610>
 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!

17

Supervised learning: regression

Image	label
	-4.5
	10.1
	3.2
	4.3

Regression: label is real-valued

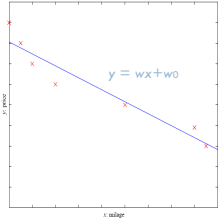
Supervised learning: given labeled examples

18

Regression Example

Price of a used car

x : car attributes (e.g. mileage)
y : price



19

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, ...
- ...

20

Supervised learning: ranking

label

1
4
2
3

Ranking: label is a ranking

Supervised learning: given labeled examples

21

NLP Ranking Applications

- reranking N-best output lists (e.g. parsing, machine translation, ...)
- Rank possible simplification options
- flight search (search in general)
- ...

22

Ranking example

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

Machine Learning - Wikipedia, the free encyclopedia

Machine Learning - Coursera

Machine Learning - MIT OpenCourseWare

23

Unsupervised learning

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

24

Unsupervised learning applications

- learn clusters/groups without any label
 - cluster documents
 - cluster words (synonyms, parts of speech, ...)
- compression
- bioinformatics: learn motifs
- ...

25

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

26

Reinforcement learning example

Backgammon

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

27

Reinforcement learning example

<https://www.youtube.com/watch?v=IXIM99xPQC8>

28

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

29

Text classification

label

spam

not spam

not spam


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

I'll use text classification as a running example

30

Representing examples

examples



What is an example?
How is it represented?

31

Features

examples

features

$f_1, f_2, f_3, \dots, f_n$

$f_1, f_2, f_3, \dots, f_n$

$f_1, f_2, f_3, \dots, f_n$





$f_1, f_2, f_3, \dots, f_n$

How our algorithms actually "view" the data

Features are the questions we can ask about the examples

32

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

33

Text: raw data

Raw data

Features?



34

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 last tv week on banana

california across in wrong capital

Occurrence of words (unigrams)

35

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 last tv week on banana

california across in wrong capital

Frequency of word occurrence (unigram frequency)

36

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

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

Other features?

38

Lots of other features

POS: occurrence, counts, sequence

Constituents

Whether 'Viagra' 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

39

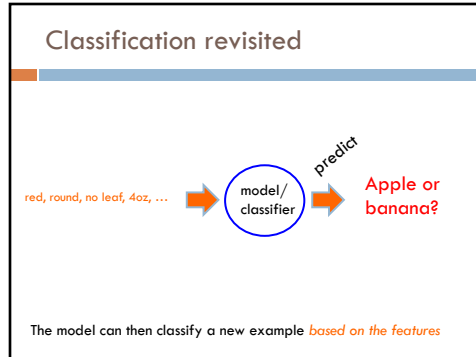
Classification revisited

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

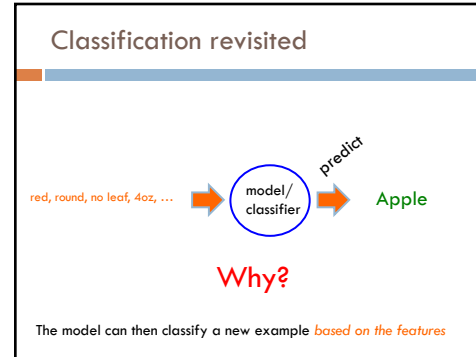
learn → model/classifier

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

40



41



42

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	

43

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?

44

Past predicts future

Training data

Test set

45

Past predicts future

Training data

Test set

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

46

Past predicts future

Training data

Test set

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

47

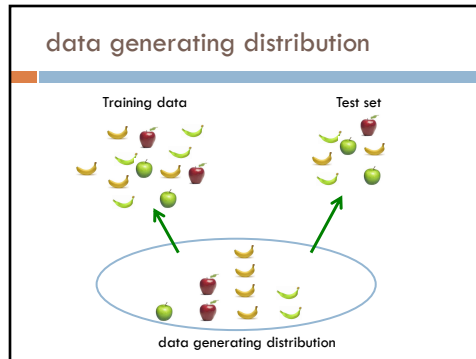
More technically...

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

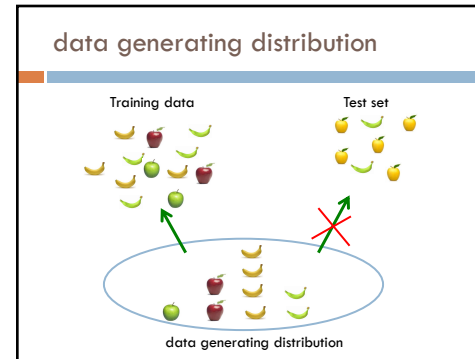
There is some probability distribution over example/label pairs called the *data generating distribution*

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

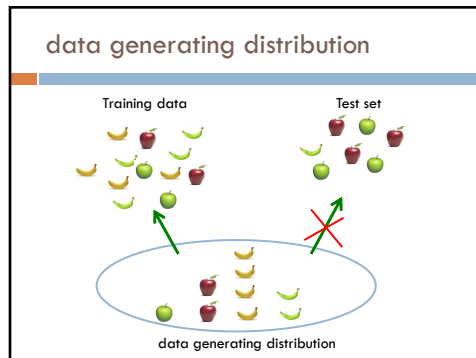
48



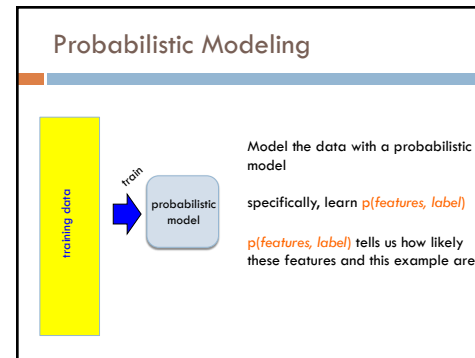
49



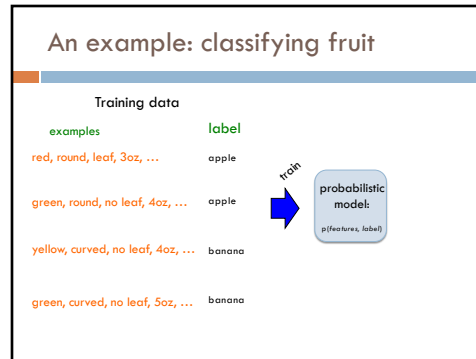
50



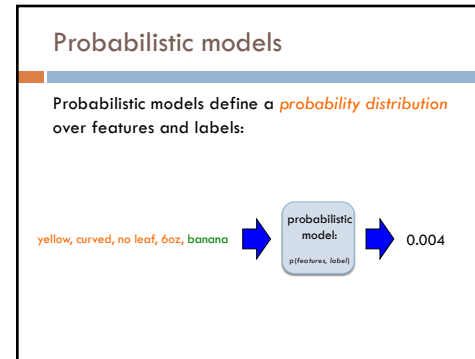
51



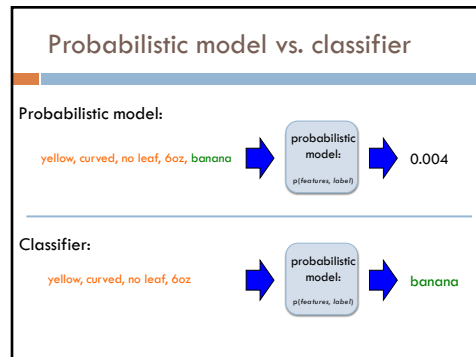
53



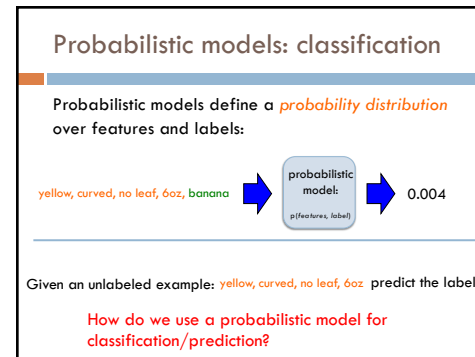
54



55



56



57

Probabilistic models

Probabilistic models define a *probability distribution* over features and labels:



For each label, ask for the probability under the model
Pick the label with the highest probability

58

Probabilistic model vs. classifier

Probabilistic model:



Classifier:



Why probabilistic models?

59

Probabilistic models

Probabilities are nice to work with

- range between 0 and 1
- can combine them in a well understood way
- lots of mathematical background/theory

Provide a strong, well-founded groundwork

- Allow us to make clear decisions about things like smoothing
- Tend to be much less "heuristic"
- Models have very clear meanings

60

Probabilistic models: big questions

1. Which model do we use, i.e. how do we calculate $p(\text{feature}, \text{label})$?
2. How do train the model, i.e. how to we we estimate the probabilities for the model?
3. How do we deal with overfitting (i.e. smoothing)?

61

Basic steps for probabilistic modeling

<p>Step 1: pick a model</p>	<p>Probabilistic models</p> <p>Which model do we use, i.e. how do we calculate $p(\text{feature}, \text{label})$?</p>
<p>Step 2: figure out how to estimate the probabilities for the model</p>	<p>How do we train the model, i.e. how do we estimate the probabilities for the model?</p>
<p>Step 3 (optional): deal with overfitting</p>	<p>How do we deal with overfitting?</p>

62

What was the data generating distribution?

The diagram illustrates the process of sampling from a data generating distribution. At the bottom, a blue oval labeled 'data generating distribution' contains several fruit icons: two apples, two bananas, and two oranges. Two green arrows point upwards from this oval to two separate groups of fruit icons. The group on the left is labeled 'Training data' and contains a mix of fruits. The group on the right is labeled 'Test set' and also contains a mix of fruits, representing a random sample from the same underlying distribution.

63

Step 1: picking a model

What we're really trying to do is model the data generating distribution, that is how likely the feature/label combinations are

The diagram shows a blue oval labeled 'data generating distribution' containing a small collection of fruit icons: one apple, one banana, and one orange, representing the underlying source of the data.

64

Some math

$$p(\text{features}, \text{label}) = p(x_1, x_2, \dots, x_m, y)$$

$$= p(y)p(x_1, x_2, \dots, x_m | y)$$

What rule?

65

Some math

$$\begin{aligned}
 p(\text{features}, \text{label}) &= p(x_1, x_2, \dots, x_m, y) \\
 &= p(y) p(x_1, x_2, \dots, x_m | y) \\
 &= p(y) p(x_1 | y) p(x_2, \dots, x_m | y, x_1) \\
 &= p(y) p(x_1 | y) p(x_2 | y, x_1) p(x_3, \dots, x_m | y, x_1, x_2) \\
 &= p(y) \prod_{j=1}^m p(x_j | y, x_1, \dots, x_{j-1})
 \end{aligned}$$

66

Step 1: pick a model

$$p(\text{features}, \text{label}) = p(y) \prod_{j=1}^m p(x_j | y, x_1, \dots, x_{j-1})$$

So, far we have made NO assumptions about the data

$$p(x_m | y, x_1, x_2, \dots, x_{m-1})$$

How many entries would the probability distribution table have if we tried to represent all possible values and we had 7000 binary features?

67

Full distribution tables

x_1	x_2	x_3	...	y	$p(\cdot)$
0	0	0	...	0	*
0	0	0	...	1	*
1	0	0	...	0	*
1	0	0	...	1	*
0	1	0	...	0	*
0	1	0	...	1	*
...

All possible combination of features!

Table size: $2^{7000} = ?$

68

2^{7000}

```

1621696755662202046666508547837709519111243036374326235982084151527023162702352987080237879
4400046519960190953098453865257892546513204107022110235646586474315852707599373340842842
7244001238187856007931085017043194684260390077841250999986816963006600112098173792646787
81962523770065229475725667805880929384627218640216108862600816097132874749204352087401101862
890823275017246052311393952355050545421454472059099650788478948339292957411250947338
6191215296848474344067412041740208875403718642170155020735398381224299258743537536161041593
43945576656170179090417297023536526662820218084938958126997095285708096963755754143487608
82485994190380241519751451012512704382908782091384763028781186024099589196419227701255
360491156240349947144160905730824293138621199536793730129449756002483337073899839209910322
34659039532049290174009801732321069130797124016933972021833007897845192584653710885
8195631737000743805167411891346175014845217698429678284287373274221202251759753994839257
0287790770635347902449354538660512591079567291431216297789784818552292819646176600903999
9799168140474928421574513802603811510482846078973048382920346040772764500776627475070714
462263487685709621261074727052049488907288785936890470634285485316886562732717466058185
6970648495801376175464527181765555719021170571400775104496782899232585777144724234900
764026321760892113552561241194538702680299040018385850276719369879536612136888838680023840
92267380777501891470304962150949838239720071497963393372028759204151729493709707783525108
32000782364807379548870694462168804462115492762900019907174215029135117441329737492500
895583051888413533479846411368004994037374560035428811232628218661131064507789292994946
91560185083982074170406821243881520209594949581813785353921029474388883163627123202
921229795384863554833571060340789177417026363656202726954375177807413134551018100094688094
0781122057380333711246329289162570895804762494920918253016369092362406714164433165159828058
3700784378885639087008490925382976
    
```

Any problems with this?

69

Full distribution tables

x_1	x_2	x_3	...	y	$p(\cdot)$
0	0	0	...	0	*
0	0	0	...	1	*
1	0	0	...	0	*
1	0	0	...	1	*
0	1	0	...	0	*
0	1	0	...	1	*
			...		

- Storing a table of that size is impossible!
- How are we supposed to learn/estimate each entry in the table?

70

Step 1: pick a model

$$p(\text{features}, \text{label}) = p(y) \prod_{j=1}^m p(x_j | y, x_1, \dots, x_{j-1})$$

So, far we have made NO assumptions about the data

Model selection involves making assumptions about the data

We've done this before, n-gram language model, parsing, etc.

These assumptions allow us to represent the data more compactly and to estimate the parameters of the model

71

Naïve Bayes assumption

$$p(\text{features}, \text{label}) = p(y) \prod_{j=1}^m p(x_j | y, x_1, \dots, x_{j-1})$$

$$p(x_i | y, x_1, x_2, \dots, x_{i-1}) = p(x_i | y)$$

What does this assume?

72

Naïve Bayes assumption

$$p(\text{features}, \text{label}) = p(y) \prod_{j=1}^m p(x_j | y, x_1, \dots, x_{j-1})$$

$$p(x_i | y, x_1, x_2, \dots, x_{i-1}) = p(x_i | y)$$

Assumes feature i is independent of the other features given the label

Is this true for text, say, with unigram features?

73

Naïve Bayes assumption

$$p(x_i | y, x_1, x_2, \dots, x_{i-1}) = p(x_i | y)$$

For most applications, this is not true!

For example, the fact that "San" occurs will probably make it *more likely* that "Francisco" occurs

However, this is often a reasonable approximation:

$$p(x_i | y, x_1, x_2, \dots, x_{i-1}) \approx p(x_i | y)$$

74