

PROBABILITY

David Keuchak
CS158 – Fall 2023

1

Admin

Midterm

Assignment 6

2

Basic probability theory: terminology

An **experiment** has a set of potential outcomes, e.g., throw a die, "look at" another example

The **sample space** of an experiment is the set of all possible outcomes, e.g., $\{1, 2, 3, 4, 5, 6\}$

For machine learning the sample spaces can be very large

3

Basic probability theory: terminology

An **event** is a subset of the sample space

Dice rolls

- $\{2\}$
- $\{3, 6\}$
- $\text{even} = \{2, 4, 6\}$
- $\text{odd} = \{1, 3, 5\}$

Machine learning

- A particular feature has particular values
- An example, i.e. a particular setting of feature values
- label = Chardonnay

4

Events

We're interested in probabilities of events

- $p(\{2\})$
- $p(\text{label}=\text{survived})$
- $p(\text{label}=\text{Chardonnay})$
- $p(\text{"Pinot" occurred})$

5

Random variables

A random variable is a mapping from the sample space to a number (think events)

It represents all the possible values of something we want to measure in an experiment

For example, random variable, X , could be the number of heads for a coin

space	HHH	HHT	HTH	HTT	THH	THT	TTH	TTT
X	3	2	2	1	2	1	1	0

Really for notational convenience, since the event space can sometimes be irregular

6

Random variables

We're interested in the probability of the different values of a random variable

The definition of probabilities over *all* of the possible values of a random variable defines a **probability distribution**

space	HHH	HHT	HTH	HTT	THH	THT	TTH	TTT
X	3	2	2	1	2	1	1	0

X	$P(X)$
3	$P(X=3) = 1/8$
2	$P(X=2) = 3/8$
1	$P(X=1) = 3/8$
0	$P(X=0) = 1/8$

7

Probability distribution

To be explicit

- A probability distribution assigns probability values to *all possible values* of a random variable
- These values must be ≥ 0 and ≤ 1
- These values must sum to 1 for all possible values of the random variable

X	$P(X)$
3	$P(X=3) = 1/2$
2	$P(X=2) = 1/2$
1	$P(X=1) = 1/2$
0	$P(X=0) = 1/2$

X	$P(X)$
3	$P(X=3) = -1$
2	$P(X=2) = 2$
1	$P(X=1) = 0$
0	$P(X=0) = 0$

8

Unconditional/prior probability

Simplest form of probability is

- $P(X)$

Prior probability: without any additional information, what is the probability

- What is the probability of heads?
- What is the probability of surviving the titanic?
- What is the probability of a wine review containing the word "banana"?
- What is the probability of a passenger on the titanic being under 21 years old?
- ...

9

Joint distribution

We can also talk about probability distributions over multiple variables

$P(X,Y)$

- probability of X and Y
- a distribution over the cross product of possible values

MLPass AND EngPass	$P(\text{MLPass}, \text{EngPass})$
true, true	.88
true, false	.01
false, true	.04
false, false	.07

10

Joint distribution

Still a probability distribution

- all values between 0 and 1, inclusive
- all values sum to 1

All questions/probabilities of the two variables can be calculate from the joint distribution

MLPass AND EngPass	$P(\text{MLPass}, \text{EngPass})$
true, true	.88
true, false	.01
false, true	.04
false, false	.07

What is $P(\text{ENGPass})$?

11

Joint distribution

Still a probability distribution

- all values between 0 and 1, inclusive
- all values sum to 1

All questions/probabilities of the two variables can be calculate from the joint distribution

MLPass AND EngPass	$P(\text{MLPass}, \text{EngPass})$
true, true	.88
true, false	.01
false, true	.04
false, false	.07

0.92

How did you figure that out?

12

Joint distribution

$P(x) = \sum_{y \in Y} p(x,y)$ This is called "summing over" or "marginalizing out" a variable

MLPass AND EngPass	P(MLPass, EngPass)
true, true	.88
true, false	.01
false, true	.04
false, false	.07

MLPass	P(MLPass)
true	0.89
false	0.11

EngPass	P(EngPass)
true	0.92
false	0.08

13

Conditional probability

As we learn more information, we can update our probability distribution

$P(X|Y)$ models this (read "probability of X given Y")

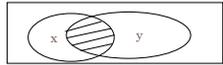
- What is the probability of a heads given that both sides of the coin are heads?
- What is the probability the document is about Chardonnay, given that it contains the word "pinot"?
- What is the probability of the word "noir" given that the sentence also contains the word "pinot"?

Notice that it is still a distribution over the values of X

14

Conditional probability

$p(X|Y) = ?$

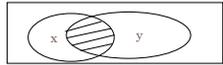


In terms of prior and joint distributions, what is the conditional probability distribution?

15

Conditional probability

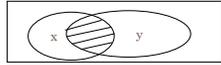
$p(X|Y) = \frac{P(X,Y)}{P(Y)}$



Given that y has happened, in what proportion of those events does x also happen

16

Conditional probability

$$p(X|Y) = \frac{P(X,Y)}{P(Y)}$$


Given that y has happened, what proportion of those events does x also happen

MLPass AND EngPass	P(MLPass, EngPass)
true, true	.88
true, false	.01
false, true	.04
false, false	.07

What is: $p(\text{MLPass}=\text{true} \mid \text{EngPass}=\text{false})$?

17

Conditional probability

$$p(X|Y) = \frac{P(X,Y)}{P(Y)}$$

MLPass AND EngPass	P(MLPass, EngPass)
true, true	.88
true, false	.01
false, true	.04
false, false	.07

What is: $p(\text{MLPass}=\text{true} \mid \text{EngPass}=\text{false})$?

$$\frac{P(\text{true}, \text{false}) = 0.01}{P(\text{EngPass} = \text{false}) = 0.01 + 0.07 = 0.08} = 0.125$$

Notice this is very different than $p(\text{MLPass}=\text{true}) = 0.89$

18

Both are distributions over X

Unconditional/prior probability		Conditional probability	
$p(X)$		$p(X Y)$	
MLPass	P(MLPass)	MLPass	P(MLPass EngPass=false)
true	0.89	true	0.125
false	0.11	false	0.875

19

A note about notation

When talking about a particular random variable value, you should technically write $p(X=x)$, etc.

However, when it's clear, we'll often shorten it

Also, we may also say $P(X)$ or $p(x)$ to generically mean any particular value, i.e. $P(X=x)$

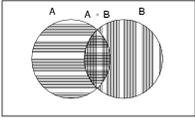
$$\frac{P(\text{true}, \text{false}) = 0.01}{P(\text{EngPass} = \text{false}) = 0.01 + 0.07 = 0.08} = 0.125$$

20

Properties of probabilities

$P(A \text{ or } B) = ?$

True

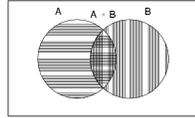


21

Properties of probabilities

$P(A \text{ or } B) = P(A) + P(B) - P(A, B)$

True



22

Properties of probabilities

$P(\neg E) = 1 - P(E)$

More generally:

- Given events $E = e_1, e_2, \dots, e_n$

$$P(e_i) = 1 - \sum_{j=1, j \neq i}^n P(e_j)$$

$P(E_1, E_2) \leq P(E_1)$

23

Chain rule (aka product rule)

$$p(X|Y) = \frac{P(X, Y)}{P(Y)} \implies p(X, Y) = P(X|Y)P(Y)$$

We can view calculating the probability of X AND Y occurring as two steps:

1. Y occurs with some probability $P(Y)$
2. Then, X occurs, given that Y has occurred

or you can just trust the math... 😊

24

Chain rule

$$p(X,Y,Z) = P(X|Y,Z)P(Y,Z)$$

$$p(X,Y,Z) = P(X,Y|Z)P(Z)$$

$$p(X,Y,Z) = P(X|Y,Z)P(Y|Z)P(Z)$$

$$p(X,Y,Z) = P(Y,Z|X)P(X)$$

$$p(X_1, X_2, \dots, X_n) = ?$$

25

Applications of the chain rule

We saw that we could calculate the individual prior probabilities using the joint distribution

$$p(x) = \sum_{y \in Y} p(x,y)$$

What if we don't have the joint distribution, but do have conditional probability information:

- $P(Y)$
- $P(X|Y)$

$$p(x) = \sum_{y \in Y} p(y)p(x|y)$$

26

Bayes' rule (theorem)

$$p(X|Y) = \frac{P(X,Y)}{P(Y)} \quad \Rightarrow \quad p(X,Y) = P(X|Y)P(Y)$$

$$p(Y|X) = \frac{P(X,Y)}{P(X)} \quad \Rightarrow \quad p(X,Y) = P(Y|X)P(X)$$

$$p(X|Y) = \frac{P(Y|X)P(X)}{P(Y)}$$

27

Bayes' rule

Allows us to talk about $P(Y|X)$ rather than $P(X|Y)$

Sometimes this can be more intuitive

Why?

$$p(X|Y) = \frac{P(Y|X)P(X)}{P(Y)}$$

28

Bayes' rule

$p(\text{disease} \mid \text{symptoms})$

- For everyone who had those symptoms, how likely is the disease?
- $p(\text{symptoms} \mid \text{disease})$
 - For everyone that had the disease, how likely is the symptom?

$p(\text{label} \mid \text{features})$

- For all examples that had those features, how likely is that label?
- $p(\text{features} \mid \text{label})$
 - For all the examples with that label, how likely is this feature?

$p(\text{cause} \mid \text{effect})$ vs. $p(\text{effect} \mid \text{cause})$

29

Gaps

I just won't put these away.

These, I just won't put away.

I just won't put _____ away.

30

Gaps

What did you put _____ away?

The socks that I put _____ away.

31

Gaps

Whose socks did you fold _____ and put _____ away?

Whose socks did you fold _____ ?

Whose socks did you put _____ away?

32

Parasitic gaps

These I'll put gap away without folding gap .

↓

These I'll put gap away.

These without folding gap .

33

Parasitic gaps

These I'll put gap away without folding gap .

1. Cannot exist by themselves (parasitic)

These I'll put my pants away without folding gap .

2. They're optional

These I'll put gap away without folding them.

34

Parasitic gaps

<http://literal-minded.wordpress.com/2009/02/10/dogs-parasitic-gap/>

35

Frequency of parasitic gaps

Parasitic gaps occur on average in 1/100,000 sentences

Problem:

Your friend has developed a machine learning approach to identify parasitic gaps. If a sentence has a parasitic gap, it correctly identifies it 95% of the time. If it doesn't, it will incorrectly say it does with probability 0.005. **Suppose we run it on a sentence and the algorithm says it is a parasitic gap, what is the probability it actually is?**

36

Prob of parasitic gaps

Your friend has developed a machine learning approach to identify parasitic gaps. If a sentence has a parasitic gap, it correctly identifies it 95% of the time. If it doesn't, it will incorrectly say it does with probability 0.005. Suppose we run it on a sentence and the algorithm says it is a parasitic gap, what is the probability it actually is?

G = gap
T = test positive

What question do we want to ask?

37

Prob of parasitic gaps

Your friend has developed a machine learning approach to identify parasitic gaps. If a sentence has a parasitic gap, it correctly identifies it 95% of the time. If it doesn't, it will incorrectly say it does with probability 0.005. Suppose we run it on a sentence and the algorithm says it is a parasitic gap, what is the probability it actually is?

G = gap
T = test positive

$$p(g | t) = ?$$

38

Prob of parasitic gaps

Your friend has developed a machine learning approach to identify parasitic gaps. If a sentence has a parasitic gap, it correctly identifies it 95% of the time. If it doesn't, it will incorrectly say it does with probability 0.005. Suppose we run it on a sentence and the algorithm says it is a parasitic gap, what is the probability it actually is?

G = gap
T = test positive

$$\begin{aligned} p(g | t) &= \frac{p(t | g)p(g)}{p(t)} \\ &= \frac{p(t | g)p(g)}{\sum_{g \in G} p(g)p(t | g)} = \frac{p(t | g)p(g)}{p(g)p(t | g) + p(\bar{g})p(t | \bar{g})} \end{aligned}$$

39

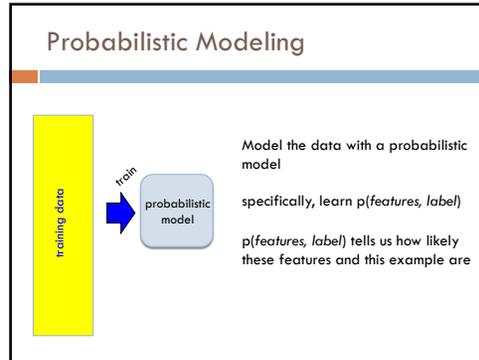
Prob of parasitic gaps

Your friend has developed a machine learning approach to identify parasitic gaps. If a sentence has a parasitic gap, it correctly identifies it 95% of the time. If it doesn't, it will incorrectly say it does with probability 0.005. Suppose we run it on a sentence and the algorithm says it is a parasitic gap, what is the probability it actually is?

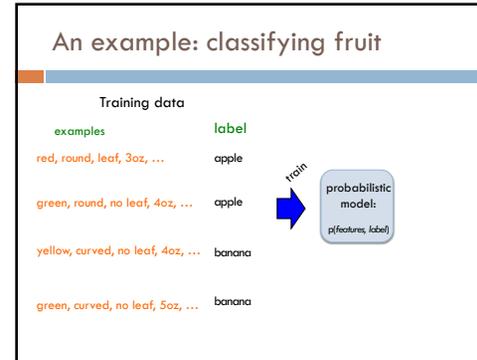
G = gap
T = test positive

$$\begin{aligned} p(g | t) &= \frac{p(t | g)p(g)}{p(g)p(t | g) + p(\bar{g})p(t | \bar{g})} \\ &= \frac{0.95 * 0.00001}{0.00001 * 0.95 + 0.99999 * 0.005} \approx 0.002 \end{aligned}$$

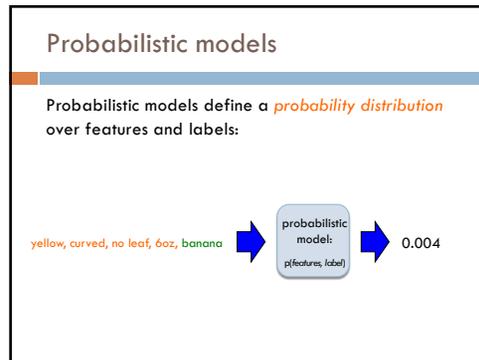
40



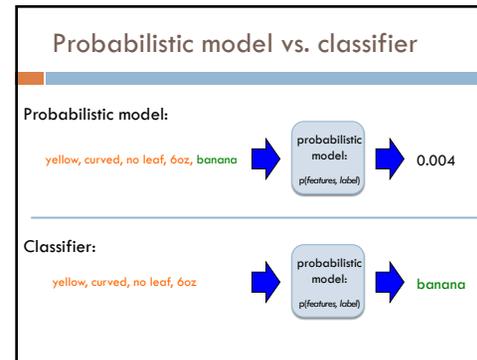
41



42



43



44

Probabilistic models: classification

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

yellow, curved, no leaf, 6oz, banana → probabilistic model: $p(\text{features}, \text{label})$ → 0.004

Given an unlabeled example: yellow, curved, no leaf, 6oz predict the label

How do we use a probabilistic model for classification/prediction?

45

Probabilistic models

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

yellow, curved, no leaf, 6oz, banana → probabilistic model: $p(\text{features}, \text{label})$ → 0.004

yellow, curved, no leaf, 6oz, apple → probabilistic model: $p(\text{features}, \text{label})$ → 0.00002

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

46

Probabilistic model vs. classifier

Probabilistic model:

yellow, curved, no leaf, 6oz, banana → probabilistic model: $p(\text{features}, \text{label})$ → 0.004

Classifier:

yellow, curved, no leaf, 6oz → probabilistic model: $p(\text{features}, \text{label})$ → banana

Why probabilistic models?

47

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
- an aside: to get the benefit of probabilistic output you can sometimes *calibrate* the confidence output of a non-probabilistic classifier

Provide a strong, well-founded groundwork

- Allow us to make clear decisions about things like regularization
- Tend to be much less "heuristic" than the models we've seen
- Different models have very clear meanings

48

Probabilistic models: big questions

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

49

Same problems we've been dealing with so far

Probabilistic models	ML in general
Which model do we use, i.e. how do we calculate $p(\text{feature}, \text{label})$?	Which model do we use (decision tree, linear model, non-parametric)
How do train the model, i.e. how do we estimate the probabilities for the model?	How do train the model?
How do we deal with overfitting?	How do we deal with overfitting?

50

Basic steps for probabilistic modeling

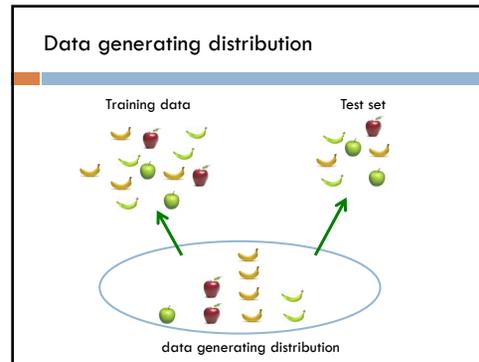
	Probabilistic models
Step 1: pick a model	Which model do we use, i.e. how do we calculate $p(\text{feature}, \text{label})$?
Step 2: figure out how to estimate the probabilities for the model	How do train the model, i.e. how do we estimate the probabilities for the model?
Step 3 (optional): deal with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

51

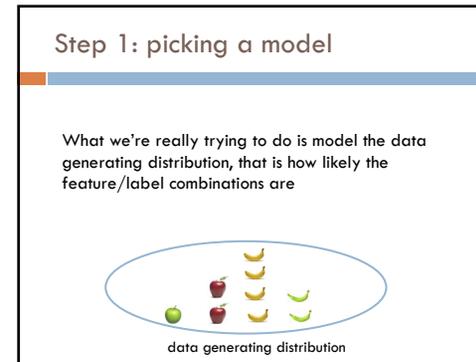
Basic steps for probabilistic modeling

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

52



53



54

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?

55

Some math

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

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

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

$$= p(y)p(x_1 | y)p(x_2 | y, x_1)p(x_3, \dots, x_m | y, x_1, x_2)$$

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

56

Step 1: pick a model

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

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 (e.g. for the wine data set)?

57

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	*
...

Wine problem:

- all possible combination of features
- ~7000 binary features
- Sample space size: $2^{7000} = ?$

58

27000

```

16216967556220202646666508547837709519111243036374325623598206415152702316270232987080237879
4460046519960190993209845386525789254651320410702211023356465864743158522707599373340842842
7234200120818785002793108501704119448426395077841200999986016936006600112008175292960787
8196252377065529257256678055895293846271864021610886260081609713287474920432087401101862
89084232701744002211593952230595484241434472094906307889478094839793934112304473438
61912152968484743440641204474020887540271869421701550220735398381224299258743537536161041593
43945376656170179090417297022365266628021808493892812699709528570890963757541434487608
834836941993024151975145101251270432087280919538476208781185402409958895964193277401355
3604811562453499947146160905708424923179621199536793730129447956002483337078998392029910222
34598039330494290714009017252108913079712420169639720218330073897845195284833710885
8195631737000743805167411891346175014845217679842967828428737312742212202517597533994839257
0298779077053534790244935435386460512591079567291431216297788784818532928196541764009803789
0799148140474738421274251580260811510482846078797048292203404277565073765075455022714
46526348768570962126107476270520304889072089785936804706342854853168865657327174660658185
6090648490801276175461407216176765575199211750751400775104467283906225854771447242334900
784023271760892113525612411943870260279904001183885057193694897593661121688883848023840
93256738077501891470300496215099698383397520015493963323720287592041517294937079077833625108
320092839604807379488768954642168844652115493767900919007174255509135117441523737479300
8955830218884135347984641136800049940373745600354288112326282186611310645507738922996946
9156018580839820747046085212438815202609938469688161375826292102947434888882163627122302
9312297938464854535357106024077817417026363656202795545217780741313451018100094688094
0781120037803337112463293891623708938047622459509182530163690923640671411644331656159828058
37207834298856239089208440902553829376
    
```

Any problems with this?

59

Midterm

- Quartile 1: 28.5 (84%)
- Median: 31 (91%)
- Quartile 3: 32 (94%)
- Mean: 30.3 (89%)

60