PROBABILITY

David Kauchak
CS158 – Fall 2025

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}
- even = {2, 4, 6}
- 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(label=survived)
- p(label=Chardonnay)
- p("Pinot" occurred)

5

Random variables

A random variable is a mapping from the sample space to a set of possible outcomes, often numbers (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(MLPass, 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 calculated from the joint distribution

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

What is P(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(MLPass, 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")

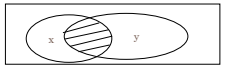
- What is the probability of a heads *given* that both sides of the coin are heads?
- What is the probability the document is about Chardonnay, given that it contains the word "Pinot"?
- What is the probability of the word "noir" given that the sentence also contains the word "pinot"?

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

14

Conditional probability

$$p(X \mid Y) = \ ?$$

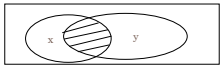


In terms of the pior (p(x) or p(y)) and joint distributions (p(x,y)), what is the conditional probability distribution?

15

Conditional probability

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

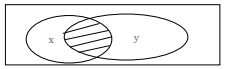


Given that y has happened, in what proportion of those events does x also happen

16

Conditional probability

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



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(MLPass=true | EngPass=false)?

17

Conditional probability

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

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

What is:
p(MLPass=true | EngPass=false)?

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

Notice this is very different than p(MLPass=true) = 0.89

18

Both are distributions over X

Unconditional/prior probability

$$p(X)$$

MLPass	P(MLPass)
true	0.89
false	0.11

Conditional probability

$$p(X \mid Y)$$

MLPass	P(MLPass | EngPass=false)
true	0.125
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

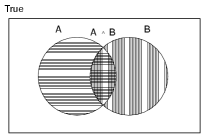
Also, we may also say P(X) or p(x) to generically mean any particular value, i.e. P(X=x)

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

20

Properties of probabilities

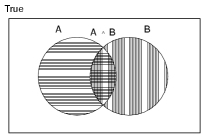
P(*A* or *B*) = ?



21

Properties of probabilities

P(*A* or *B*) = P(*A*) + P(*B*) - P(*A*,*B*)



22

Properties of probabilities

P(¬E) = 1– P(E)

More generally:

- Given events E = e_1, e_2, …, e_n

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

P(E1, E2) ≤ P(E1)

23

Chain rule (aka product rule)

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

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

$p(X,Y,Z) = P(X \mid Y,Z)P(Y,Z)$

$p(X,Y,Z) = P(X,Y \mid Z)P(Z)$

$p(X,Y,Z) = P(X \mid Y,Z)P(Y \mid Z)P(Z)$

$p(X,Y,Z) = P(Y,Z \mid X)P(X)$

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

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)

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

26

Bayes' rule (theorem)

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

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

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

27

Bayes' rule

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

Sometimes this can be more intuitive

Why?

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

28

Bayes' rule

p(disease | symptoms)
- For everyone who had those symptoms, how likely is the disease?
- p(symptoms|disease)
  - For everyone that had the disease, how likely is the symptom?

p(label| features)
- For all examples that had those features, how likely is that label?
- p(features | label)
  - For all the examples with that label, how likely is this feature

- p(cause | effect) vs. p(effect | cause)

29

Gaps

These, I just won't put away.

30

Gaps

What did you put ___ away?
gap

The socks that I put ___ away.
gap

31

Gaps

Whose socks did you fold ___ and put ___ away?
gap gap

Whose socks did you fold ___ ?
gap

Whose socks did you put ___ away?
gap

32

Parasitic gaps

These I'll put ___ away without folding ___.
gap gap

These I'll put ___ away.
gap

These without folding ___ .
gap

33

Parasitic gaps

These I'll put ___ away without folding ___.
gap 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 ___ away without folding them.
gap

34

Parasitic gaps

http://literalminded.wordpress.com/2009/02/10/dougs-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

$$p(g \mid t) = ?$$

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

$$p(g \mid t) = \frac{p(t \mid g)p(g)}{p(t)}$$

$$= \frac{p(t \mid g)p(g)}{\sum_{g \in G} p(g)p(t \mid g)} = \frac{p(t \mid g)p(g)}{p(g)p(t \mid g) + p(\overline{g})p(t \mid \overline{g})}$$

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?

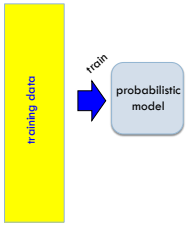
G = gap
T = test positive

$$p(g \mid t) = \frac{p(t \mid g)p(g)}{p(g)p(t \mid g) + p(\overline{g})p(t \mid \overline{g})}$$

$$= \frac{0.95 * 0.00001}{0.00001 * 0.95 + 0.99999 * 0.005} \approx 0.002$$

40

Probabilistic Modeling



Model the data with a probabilistic model

specifically, learn p(*features*, *label*)

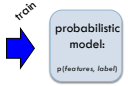
p(*features*, *label*) tells us how likely these features and this example are

41

An example: classifying fruit

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



42

Probabilistic models

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

43

Probabilistic model vs. classifier

Probabilistic model:

Classifier:

44

Probabilistic models: classification

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

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:

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

46

Probabilistic model vs. classifier

Probabilistic model:

Classifier:

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(*feature, label*)?

How do train the model, i.e. how do we 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(*feature, label*)?	Which model do we use (decision tree, linear model, non-parametric)
How do train the model, i.e. how to we 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

Step 1: pick a model

Step 2: figure out how to estimate the probabilities for the model

Step 3 (optional): deal with overfitting

Probabilistic models

Which model do we use, i.e. how do we calculate p(*feature, label*)?

How do train the model, i.e. how to we we estimate the probabilities for the model?

How do we deal with overfitting?

51

Basic steps for probabilistic modeling

Step 1: pick a model

Step 2: figure out how to estimate the probabilities for the model

Step 3 (optional): deal with overfitting

Probabilistic models

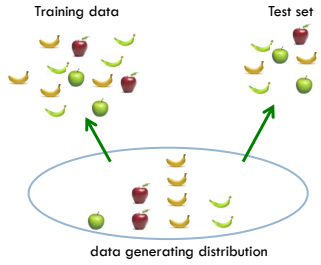
Which model do we use, i.e. how do we calculate p(*feature, label*)?

How do train the model, i.e. how to we we estimate the probabilities for the model?

How do we deal with overfitting?

52

Data generating distribution



53

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

54

Some math

$$p(features, label) = p(x_1, x_2, ..., x_m, y)$$

$$= p(y)p(x_1, x_2, ..., x_m \mid y)$$

What rule?

55

Some math

$$p(features, label) = p(x_1, x_2, ..., x_m, y)$$

$$= p(y)p(x_1, x_2, ..., x_m \mid y)$$

$$= p(y)p(x_1 \mid y)p(x_2, ..., x_m \mid y, x_1)$$

$$= p(y)p(x_1 \mid y)p(x_2 \mid y, x_1)p(x_3, ..., x_m \mid y, x_1, x_2)$$

$$= p(y)\prod_{j=1}^{m} p(x_i \mid y, x_1, ..., x_{i-1})$$

56

Step 1: pick a model

$$p(features, label) = p(y)\prod_{j=1}^{m} p(x_i \mid y, x_1, ..., x_{i-1})$$

So, far we have made NO assumptions about the data

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

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( )
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

2^{7000}

16216967556622020264666650854783770951911124303637432562359820841515270231627023529870802378794460004651996019099530984538652557892546513204107022110233664588474315822270763599373340842842722420012281878260072931082617043194484266392077784125099996860169436006660011209817579296678781962552377006552947572566780558092938446272186402161088626008160971328747492043520874011018626980432275017246032311293955235505905454421454772509509096507889478094683592939574112569473438619121529684847434446074120417402088754037186942170155022073539838122429925874353753616104159343594557666561701790904172597025336526666268202180849389281269970952857089069637557541434487608824836994199380241519751451012512704382908728091953847630285781185402409995889596419227760125536049115624034999471441609053708422923139621199536795730129447956002483335707389983920299102223465980389530690429801740098017325210691307971242016963397230218335007589784519525848553710885819563173700074380516741118913461750148452176798429678282422873731274221220225175975359948392579298779077063553347902449354353866601259107956729143121629778878481855229281965417660098039899799168140474938421574351580260381151068286406789730483829220346042775765507377656754750702714466226348768570962126107476270502030494889072089785936890470634285485316686665657327174660658185609066484950801276175461457216176955575199211730731406777510449672859082255854777144724233490076402632176089211355256124119453870268029904400183858505767193696897593661213568888386800238409325673807775018914703049621509969838539752071549396339237202875920415172949370790977853625108320092839604807237954887069546621488044632112493076299091990717742355039135117441332973747930089955830518884135334798464113680004999403737245600354288112326328218661131064550772899229969469156018588039820741704606832124388152026099584696588161375826382931029547343888832163627122302921229795384868355483535710603407789177417026365620272695543751778074131345510181000946880940781112057380335371124632958916237089580476224959091825301636909236240671411644331656159828058372078343988856239089202844090255382937 6

Any problems with this?

59