PROBABILITY

David Kauchak
CS158 – Fall 2019

Admin

Midterm

Mean:	35.5 (89%)
Median:	35.5 (89%)

Assignment grading

Assignment 6

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

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

Events

We're interested in probabilities of events

- p({2})
- p(label=survived)
- p(label=Chardonnay)
- p(parasitic gap)
- p("Pinot" occurred)

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

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

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

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?
- …

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

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

What is P(ENGPass)?

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?

Joint distribution

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

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

Conditional probability

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

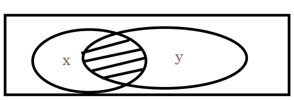
P(X|Y) models this (read "probability of X *given* Y")

- 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

Conditional probability

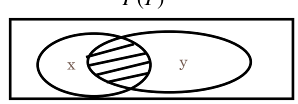
$$p(X \mid Y) = \ ?$$



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

Conditional probability

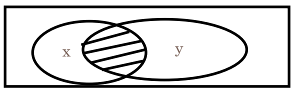
$$p(X \mid Y) = \frac{P(X,Y)}{P(Y)}$$



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

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)?

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

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

A note about notation

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

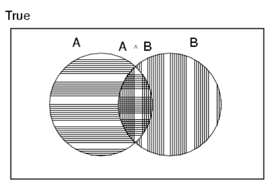
However, when it's clear , we'll often shorten it

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

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

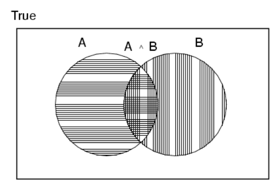
Properties of probabilities

P(*A or B*) = ?



Properties of probabilities

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



Properties of probabilities

P(¬E) = 1– P(E)

More generally:

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

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

P(E1, E2) ≤ P(E1)

Chain rule (aka product rule)

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

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… ☺

Chain rule

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

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

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

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

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

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)

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

This is called "summing over" or "marginalizing out" a variable

Bayes' rule (theorem)

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

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

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

Bayes' rule

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

Sometimes this can be more intuitive

Why?

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

Bayes' rule

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

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

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

Gaps

V

I just won't put these away.

direct object

These, I just won't put away.

filler

I just won't put ___ away.

gap

Gaps

What did you put ___ away?

gap

The socks that I put ___ away.

gap

Gaps

Whose socks did you fold ___ and put ___ away?

gap gap

Whose socks did you fold ___ ?

gap

Whose socks did you put ___ away?

gap

Parasitic gaps

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

These I'll put ____ away.
gap

These without folding ____ .
gap

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

Parasitic gaps

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

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?

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?

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

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

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

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

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

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

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

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

Probabilistic Modeling

training data

train

probabilistic model

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

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

train

probabilistic model: p(*features, label*)

Probabilistic models

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

yellow, curved, no leaf, 6oz, banana → probabilistic model: p(*features, label*) → 0.004

Probabilistic model vs. classifier

Probabilistic model:

yellow, curved, no leaf, 6oz, banana → probabilistic model: p(*features, label*) → 0.004

Classifier:

yellow, curved, no leaf, 6oz → probabilistic model: p(*features, label*) → banana

Probabilistic models: classification

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

yellow, curved, no leaf, 6oz, banana → probabilistic model: p(features, 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?

Probabilistic models

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

yellow, curved, no leaf, 6oz, banana → probabilistic model: p(features, label) → **0.004**

yellow, curved, no leaf, 6oz, apple → probabilistic model: p(features, label) → 0.00002

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

Probabilistic model vs. classifier

Probabilistic model:

yellow, curved, no leaf, 6oz, banana → probabilistic model: p(features, label) → 0.004

Classifier:

yellow, curved, no leaf, 6oz → probabilistic model: p(features, label) → banana

Why probabilistic models?

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

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 to we we estimate the probabilities for the model?

How do we deal with overfitting?

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?

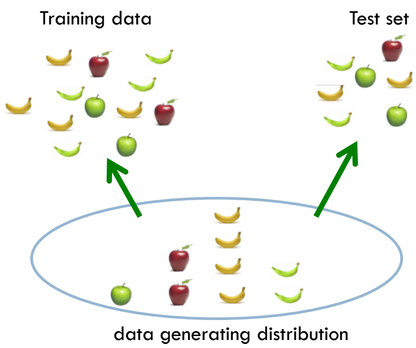
Basic steps for probabilistic modeling

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

Basic steps for probabilistic modeling

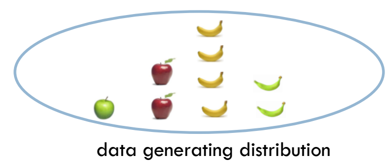
	Probabilistic models
Step 1: pick a model	Which model do we use, i.e. how do we calculate p(*feature, label*)?
Step 2: figure out how to estimate the probabilities for the model	How do train the model, i.e. how to we we estimate the probabilities for the model?
Step 3 (optional): deal with overfitting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

What was the data generating distribution?



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



Some maths

$$p(\mathit{features}, \mathit{label}) = p(x_1, x_2, \ldots, x_m, y)$$

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

What rule?

Some maths

$$p(\mathit{features}, \mathit{label}) = p(x_1, x_2, \ldots, x_m, y)$$

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

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

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

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

Step 1: pick a model

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

So, far we have made NO assumptions about the data

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

How many entries would the probability distribution table have if we tried to represent all possible values (e.g. for the wine data set)?

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} = ?

2^{7000}

1621696755662202026466665085478377095191112430363743256235982084151527023162702352987080237879
4460004651996019099530984538652557892546513204107022110253564658647431585227076599373340842842
7224200122818782600729310826170431944842663920777841250999968601694360066600112098175792966787
8196255237700655294757256678055809293844627218640216108862600816097132874749204352087401101862
6908423275017246052311293955235059054544214554772509509096507889478094683592939574112569473438
6191215296848474344406741204174020887540371869421701550220735398381224299258743537536161041593
4359455766656170179090417259702533652666268202180849389281269970952857089069637557541434487608
8248369941993802415197514510125127043829087280919538476302857811854024099958895964192277601255
3604911562403499947144160905730842429313962119953679373012944795600248333570738998392029910322
3465980389530690429801740098017325210691307971242016963397230218353007589784519525848553710885
8195631737000743805167411189134617501484521767984296782842287373127422122022517597535994839257
0298779077063553347902449354353866605125910795672914312162977887848185522928196541766009803989
9799168140474938421574351580260381151068286406789730483829220346042775765507377656754750702714
4662263487685709621261074762705203049488907208978593689047063428548531668665657327174660658185
6090664849508012761754614572161769555751992117507514067775104496728590822558547771447242334900
7640263217608921135525612411945387026802990440018385850576719369689759366121356888838680023840
9325673807775018914703049621509969838539752071549396339237202875920415172949370790977853625108
3200928396048072379548870695466216880446521124930762900919907177423550391351174415329737479300
8995583051888413533479846411368000499940373724560035428811232632821866113106455077289922996946
9156018580839820741704606832124388152026099584696588161375826382921029547343888832163627122302
9212297953848683554835357106034077891774170263636562027269554375177807413134551018100094688094
0781122057380335371124632958916237089580476224595091825301636909236240671411644331656159828058
3720783439888562390892028440902553829376

Any problems with this?