PROBABILITY

David Kauchak
CS451 – Fall 2013

Admin

Midterm

Grading

Assignment 6

No office hours tomorrow from 10-11am (though I'll be around most of the rest of the day)

Basic Probability Theory: terminology

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

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 ***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 a particular values
- An example, i.e. a particular setting of features 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 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 a 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	P(MLPass)
true	0.89
false	0.11

EngPass	P(EngPass)
true	0.92
false	0.08

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

Conditional probability

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

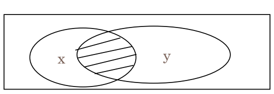
P(X|Y) models this (read "probability of X *given* Y")

- 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

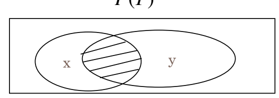
$$p(X \mid Y) = \ ?$$



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

Conditional probability

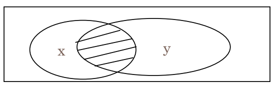
$$p(X \mid Y) = \frac{P(X,Y)}{P(Y)}$$



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 assignment, you should technically write p(X=x), etc.

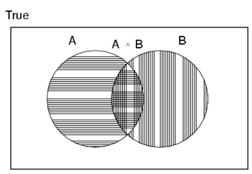
However, when it's clear , we'll often shorten it

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

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

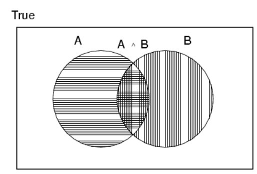
Properties of probabilities

P(*A or B*) = ?



Properties of probabilities

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



Properties of probabilities

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

More generally:

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

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

$P(E1, E2) \leq P(E1)$

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:

Maggie Louise Gal (aka "ML" Gal) 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

Maggie Louise Gal (aka "ML" Gal) 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

Maggie Louise Gal (aka "ML" Gal) 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

Maggie Louise Gal (aka "ML" Gal) 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

Maggie Louise Gal (aka "ML" Gal) 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

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

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