CSE419 – Artificial Intelligence and Machine Learning 2021

PhD Furkan Gözükara, Toros University

https://github.com/FurkanGozukara/CSE419-ArtificiaHntelligence-and-Machine-Learning-2021

Lecture 13 Part 1 Probability

Based on Asst. Prof. Dr. David Kauchak (Pomona College) Lecture Slides

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

- **1** {2}
- **3**, 6
- \blacksquare 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 = chicken

Events

We're interested in probabilities of events

- **□** p({2})
- p(label=survived)
- p(label=chicken)
- p(parasitic gap)
- p("meat" 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	ннн	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	ннн	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</p>
- 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 meal review containing the word "banana"?
- What is the probability of a passenger on the titanic being under 21 years old?

We can also talk about probability distributions over multiple **depended** variables

P(X,Y)

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

MLPas s	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	

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)?

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?

$$P(x) = \mathop{\mathrm{a}}_{y \, \widehat{I} \, Y} p(x,y)$$

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

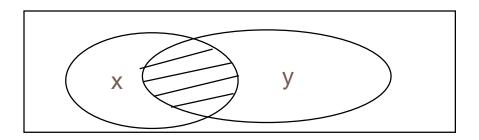
As we learn more information, we can update our probability distribution

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 chicken, given that it contains the word "meat"?
- What is the probability of the word "fish" given that the sentence also contains the word "meat"?

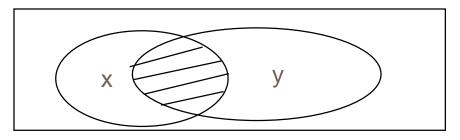
Notice that it is still a distribution over the values of X

$$p(X|Y) = ?$$



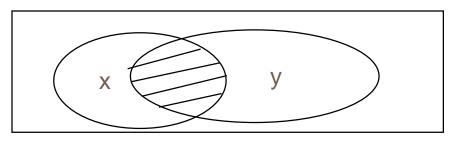
In terms of prior and joint distributions, what is the conditional probability distribution?

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



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

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



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)?

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

MLPas s	P(MLPass)
true	0.89
false	0.11

Conditional probability

MLPass	P(MLPass Eng Pass=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

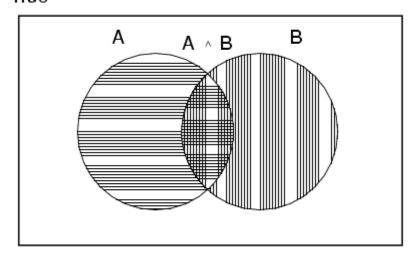
Also, we may also say P(X) or p(x) to generically mean any particular value, i.e. P(X=x)

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

Properties of probabilities

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

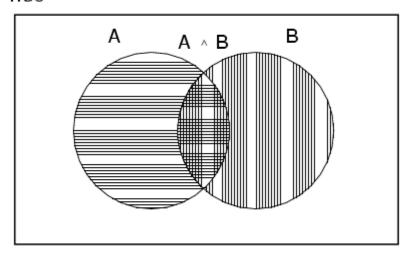




Properties of probabilities

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





Properties of probabilities

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

More generally:

□ Given events $E = e_1, e_2, ..., e_n$

$$p(e_i) = 1 - \mathop{a}_{j=1:n,j^{\perp}i} p(e_j)$$

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

Chain rule (aka product rule)

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

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... ©

Chain rule

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

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

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

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

$$p(X_1, X_2, ..., X_n) = ?$$

Applications of the chain rule

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

$$p(x) = \mathop{\mathrm{a}}_{y \, \widehat{}} p(x,y)$$

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

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

$$p(x) = \mathop{\mathrm{a}}_{y \, \hat{l} \, Y} p(y) p(x \mid y)$$

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

Bayes' rule (theorem)

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

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

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

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

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?

The socks that I put away.

Gaps

Whose socks did you fold away?

and put gap



Whose socks did you fold?

Whose socks did you put gap

away?

Parasitic gaps

These I'll put away without folding gap



These I'll put away.

These without folding

Parasitic gaps

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.

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?

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

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

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?

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

$$= \frac{p(t \mid g)p(g)}{\mathop{\mathring{o}}_{g \mid G}} = \frac{p(t \mid g)p(g)}{p(g)p(t \mid g)} = \frac{p(t \mid g)p(g)}{p(g)p(t \mid g) + p(\overline{g})p(t \mid \overline{g})}$$

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?

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

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

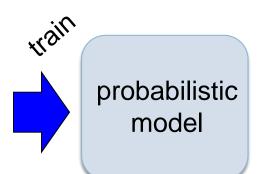
CSE419 – Artificial Intelligence and Machine Learning 2021

PhD Furkan Gözükara, Toros University

https://github.com/FurkanGozukara/CSE419 2021

Lecture 12.1 Probabilistic Models

Based on Asst. Prof. Dr. David Kauchak (Pomona College) Lecture Slides



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

trair.



probabilistic model:

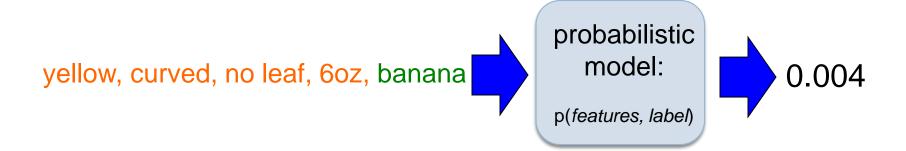
p(features, label)

yellow, curved, no leaf, 4oz, ...banana

green, curved, no leaf, 5oz, ...banana

Probabilistic models

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



Probabilistic model vs. classifier

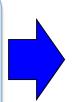
Probabilistic model:

yellow, curved, no leaf, 6oz, banana



probabilistic model:

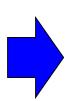
p(features, label)



0.004

Classifier:

yellow, curved, no leaf, 6oz



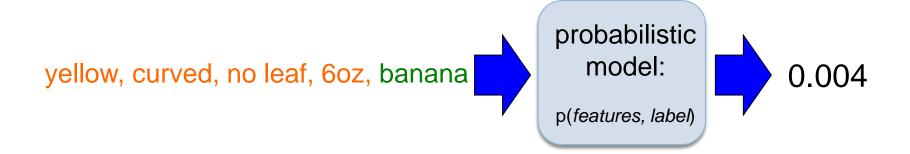
probabilistic model:

p(features, label)



Probabilistic models: classification

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



Given an unlabeled example:

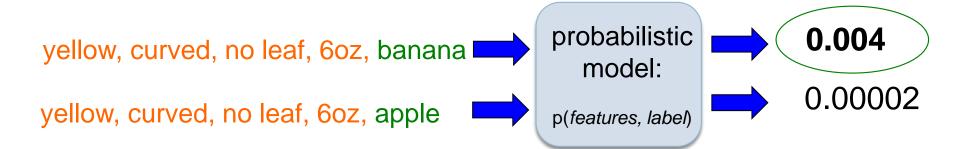
predict the label

yellow, curved, no leaf, 6oz

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

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

Probabilistic model vs. classifier

Probabilistic model:

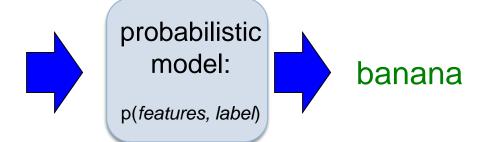
yellow, curved, no leaf, 6oz, banana probabilistic model:

p(features, label)

probabilistic model:

Classifier:

yellow, curved, no leaf, 6oz



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 nonprobabilistic 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?

Same problems we've been dealing with so far

Probabilistic models

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

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

How do we deal with overfitting?

ML in general

Which model do we use (decision tree, linear model, non-parametric)

How we train the model?

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?

Basic steps for probabilistic modeling

Step 1: pick a model

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

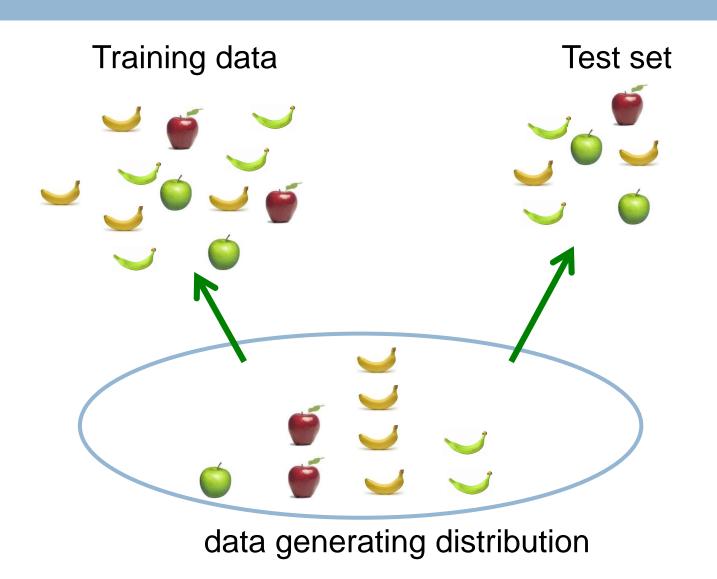
Step 3 (optional): deal with overfitting

Probabilistic models

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

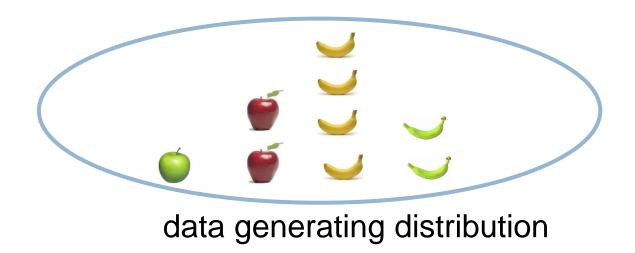
How do train the model, i.e. how to we we estimate the probabilities for the model?

What was the data generating distribution?



Step 1: picking a model

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



Some maths

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

= $p(y)p(x_1, x_2, ..., x_m | y)$

What rule?

Some maths

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

$$= p(y)p(x_1, x_2, ..., x_m | y)$$

$$= p(y)p(x_1 | y)p(x_2, ..., x_m | y, x_1)$$

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

$$= p(y) \bigcap_{i=1}^{m} p(x_i | y, x_1, ..., x_{i-1})$$

Step 1: pick a model

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

So, far we have made NO assumptions about the data

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

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

Full distribution tables

X ₁	X ₂	X_3	 у	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	*

chicken problem:

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

Any problems with this?

Full distribution tables

X ₁	X ₂	X ₃	 у	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	*

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

Step 1: pick a model

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

So, far we have made NO assumptions about the data

Model selection involves making assumptions about the data

We did this before, e.g. assume the data is linearly separable

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

An aside: independence

Two variables are independent if one has nothing whatever to do with the other

For two independent variables, knowing the value of one does not change the probability distribution of the other variable (or the probability of any individual event)

- the result of the toss of a coin is independent of a roll of a dice
- price of tea in England is independent of the whether or not you pass AI

independent or dependent?

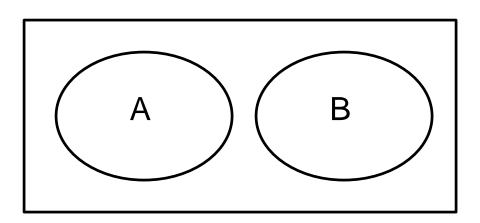
Catching a cold and having cat-allergy

Miles per gallon and driving habits

Height and longevity of life

Independent variables

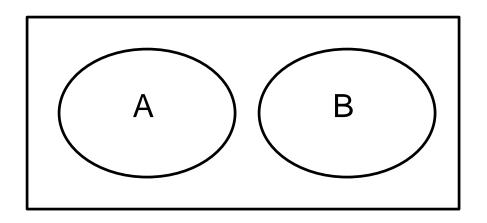
How does independence affect our probability equations/properties?



If A and B are independent (written ...)

- P(A,B) = ?
- P(A|B) = ?
- P(B|A) = ?

Independent variables



If A and B are independent (written ...)

- P(A,B) = P(A)P(B)
- \square P(A|B) = P(A)

How does independence help us?

 \square P(B|A) = P(B)

Independent variables

If A and B are independent

- P(A,B) = P(A)P(B)
- \square P(A|B) = P(A)
- \square P(B|A) = P(B)

Reduces the storage requirement for the distributions

Reduces the complexity of the distribution

Reduces the number of probabilities we need to estimate

Conditional Independence

Dependent events can become independent given certain other events

Examples,

- height and length of life
- "correlation" studies
 - size of your lawn and length of life

If A, B are conditionally independent of C

- P(A,B|C) = P(A|C)P(B|C)
- P(A|B,C) = P(A|C)
- P(B|A,C) = P(B|C)
- but $P(A,B) \neq P(A)P(B)$

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

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

What does this assume?

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

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

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

For the chicken problem?

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

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

Assumes the probability of a word occurring in a review is independent of the other words *given the label*

For example, the probability of "fish" occurring is independent of whether or not "meat" occurs given that the review is about "chicken"

Is this assumption true?

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

For most applications, this is not true!

For example, the fact that "chicken" occurs will probably make it *more likely* that "meat" occurs

However, this is often a reasonable approximation:

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

Naïve Bayes model

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

$$= p(y) \bigodot_{i=1}^{\underline{m}} p(x_i \mid y)$$
 naïve bayes assumption

 $p(x_i|y)$ is the probability of a particular feature value given the label

How do we model this?

- for binary features
- for discrete features, i.e. counts
- for real valued features

p(x|y)

Binary features:

$$p(x_i | y) = \begin{cases} q_i & \text{if } x_i = 1 \\ 1 - q_i & \text{otherwise} \end{cases}$$

biased coin toss!

Other features:

Could use lookup table for each value, but doesn't generalize well

Better, model as a distribution:

- gaussian (i.e. normal) distribution
- poisson distribution
- multinomial distribution (more on this later)
- ...

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?

Obtaining probabilities

















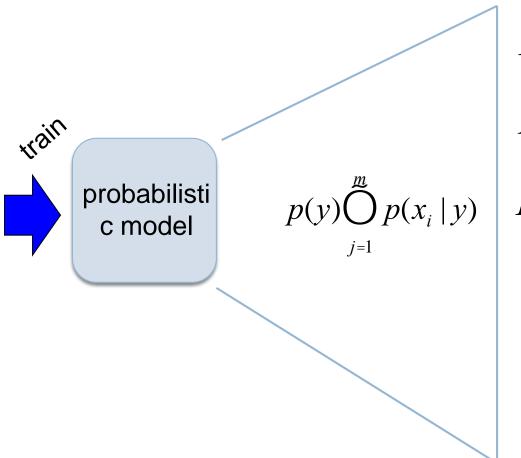




We've talked a lot about probabilities, but not where they come from

- How do we calculate $p(x_i|y)$ from training data?
- What is the probability of surviving the titanic?
- What is that any review is about Pinot Noir?
- What is the probability that a particular review is about Pinot Noir?

training data



p(y)

 $p(x_1|y)$

 $p(x_2 | y)$

:

 $p(x_m | y)$

Estimating probabilities

What is the probability of a chicken meat review?

We don't know!

We can **estimate** that based on data, though:

number of review labeled chicken meat

total number of reviews

This is called the maximum likelihood estimation. Why?

Maximum Likelihood Estimation (MLE)

Maximum likelihood estimation picks the values for the model parameters that maximize the likelihood of the training data

You flip a coin 100 times. 60 times you get heads.

What is the MLE for heads?

p(head) = 0.60

Watch Videos to Understand Better

- StatQuest: The Normal Distribution, Clearly
 Explained!!! > https://youtu.be/rzFX5NWojp0
- StatQuest: Maximum Likelihood, clearly explained!!! > https://youtu.be/XepXtl9YKwc
- StatQuest: Probability vs Likelihood > https://youtu.be/pYxNSUDSFH4
- Maximum Likelihood For the Normal Distribution, step-by-step! > https://youtu.be/Dn6b9fCIUpM

Maximum Likelihood Estimation (MLE)

Maximum likelihood estimation picks the values for the model parameters that maximize the likelihood of the training data

You flip a coin 100 times. 60 times you get heads.

What is the likelihood of the data under this model (each coin flip is a data point)?

MLE example

You flip a coin 100 times. 60 times you get heads.

MLE for heads: p(head) = 0.60

What is the likelihood of the data under this model (each coin flip is a data point)?

$$likelihood(data) = \tilde{O}_{i}p(x_{i})$$

$$\log(0.60^{60} * 0.40^{40}) = -67.3$$

MLE example

Can we do any better?

$$likelihood(data) = \widetilde{O}_{i} p(x_{i})$$

p(heads) =
$$0.5$$

log($0.50^{60} * 0.50^{40}$) =-69.3

p(heads) =
$$0.7$$

$$\log(0.70^{60} * 0.30^{40}) = -69.5$$

Useful Articles

- Probability concepts explained: Maximum likelihood estimation > https://towardsdatascience.com/probability-concepts-explained-maximum-likelihood-estimation-c7b4342fdbb1
- A Gentle Introduction to Maximum Likelihood Estimation > https://towardsdatascience.com/a-gentle-introduction-to-maximum-likelihood-estimation-9fbff27ea12f

Very Useful Video

StatQuest: Probability vs Likelihood >
 https://www.youtube.com/watch?v=pYxNS
 UDSFH4&feature=youtu.be

CSE419 – Artificial Intelligence and Machine Learning 2021

PhD Furkan Gözükara, Toros University

https://github.com/FurkanGozukara/CSE419 2021

Lecture 13 Part 2 Probabilistic Models

Based on Asst. Prof. Dr. David Kauchak (Pomona College) Lecture Slides

Maximum Likelihood Estimation (MLE)

You flip a coin 100 times. 60 times you get heads and 40 times you get tails.

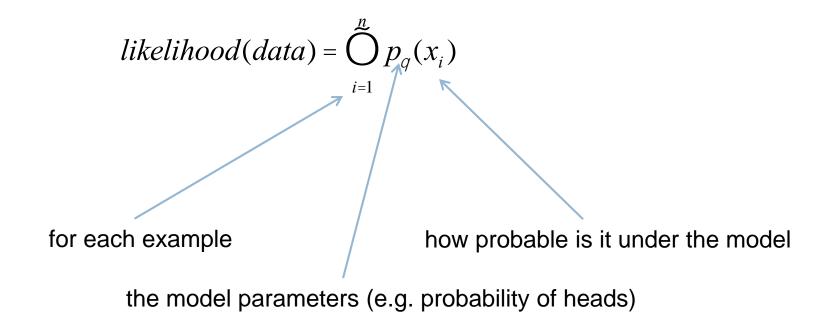
What is the probability for heads?

p(head) = 0.60

Why?

Likelihood

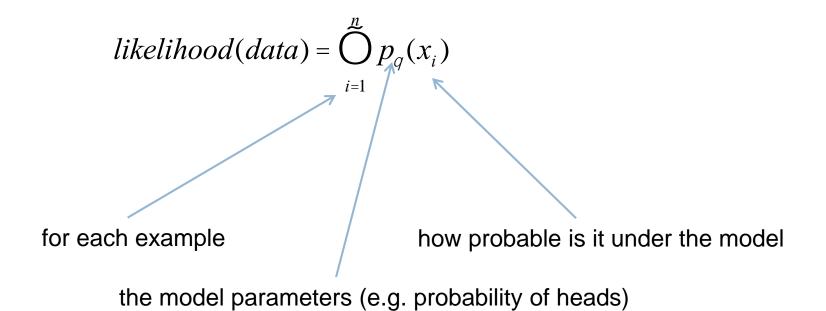
The *likelihood* of a data set is the probability that a particular model (i.e. a model and estimated probabilities) assigns to the data



Likelihood

You flip a coin 100 times. 60 times you get heads and 40 times you get tails.

What is the likelihood of this data with Θ =p(head) = 0.6 ?



Likelihood

You flip a coin 100 times. 60 times you get heads and 40 times you get tails.

What is the likelihood of this data with Θ =p(head) = 0.6 ?

$$likelihood(data) = \bigcap_{i=1}^{n} p_q(x_i)$$

 $0.60^{60} * 0.40^{40} = 5.908465121038621e-30$

60 heads with p(head) = 0.6

40 tails with p(tail) = 0.4

Maximum Likelihood Estimation (MLE)

The *maximum likelihood* estimate for a model parameter is the one that maximize the likelihood of the training data

$$MLE = \arg\max_{q} \bigcap_{i=1}^{n} p_{q}(x_{i})$$

Often easier to work with log-likelihood:

$$MLE = \underset{i=1}{\operatorname{argmax}} \log(\bigodot_{p_q}^{n}(x_i))$$

$$= \underset{i=1}{\operatorname{argmax}} \mathop{\Diamond}_{q}^{n} \log(p(x_i))$$

$$= \underset{i=1}{\operatorname{argmax}} \log(p(x_i))$$

The *maximum likelihood* estimate for a model parameter is the one that maximize the likelihood of the training data

$$MLE = \operatorname{argmax}_{q} \stackrel{n}{\overset{n}{\circlearrowleft}} \log(p(x_{i}))$$

Given some training data, how do we calculate the MLE?

You flip a coin 100 times. 60 times you get heads and 40 times you get tails.

You flip a coin 100 times. 60 times you get heads and 40 times you get tails.

$$\log - likelihood = \mathop{\Diamond}_{i=1}^{n} \log(p(x_i))$$

$$= 60\log(p(heads)) + 40\log(p(tails))$$

$$= 60\log(q) + 40\log(1-q)$$

$$MLE = \arg\max_{q} 60\log(q) + 40\log(1 - q)$$

How do we find the max?

You flip a coin 100 times. 60 times you get heads and 40 times you get tails.

$$\frac{d}{dq}60\log(q) + 40\log(1 - q) = 0$$

$$\frac{60}{q} - \frac{40}{1 - q} = 0$$

$$\frac{40}{1 - q} = \frac{60}{q}$$

$$40q = 60 - 60q$$

$$100q = 60$$

$$q = \frac{60}{100}$$
Yay!

You flip a coin n times. a times you get heads and b times you get tails.

$$\frac{d}{dq}a\log(q) + b\log(1-q) = 0$$

. . .

$$Q = \frac{a}{a+b}$$

MLE: sanity check

You flip a coin 100 times. 60 times you get heads and 40 times you get tails.

Can we do any better?

$$\log - likelihood = \mathop{a}_{i=1}^{n} \log(p(x_i))$$

$$p(heads) = 0.6$$

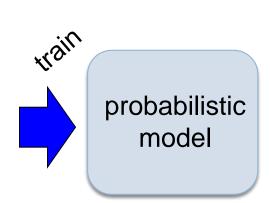
$$\log(0.60^{60} * 0.40^{40}) = -67.3$$

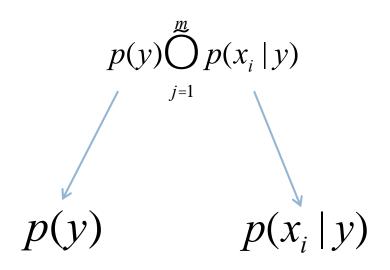
$$p(heads) = 0.5$$

 $log(0.50^{60} * 0.50^{40}) = -69.3$

p(heads) =
$$0.7$$

$$\log(0.70^{60} * 0.30^{40}) = -69.5$$





What are the MLE estimates for these?

Maximum likelihood estimates

$$p(y) = \frac{count(y)}{n}$$

number of examples with label

total number of examples

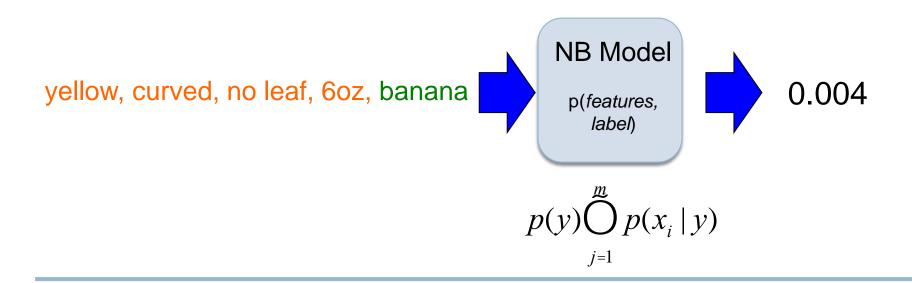
$$p(x_i \mid y) = \frac{count(x_i, y)}{count(y)}$$

number of examples with the label with feature

number of examples with label

What does training a NB model then involve? How difficult is this to calculate?

Naïve Bayes classification



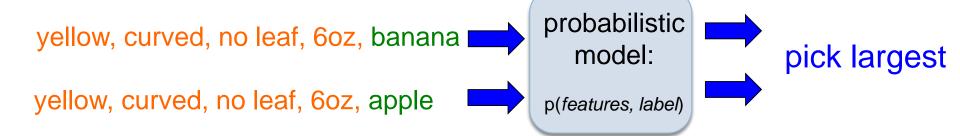
Given an unlabeled example:

predict the label

yellow, curved, no leaf, 6oz

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

Probabilistic models



$$p(y) \bigcap_{j=1}^{m} p(x_i \mid y)$$

label =
$$\underset{y \in labels}{\operatorname{arg max}} p(y) \bigodot_{j=1}^{\underline{m}} p(x_i \mid y)$$

Generative Story



To classify with a model, we're given an example and we obtain the probability

We can also ask how a given model would *generate* a document

This is the "generative story" for a model

Looking at the generative story can help understand the model

We also can use generative stories to help develop a model

NB generative story



$$p(y) \bigcap_{j=1}^{m} p(x_i \mid y)$$

What is the generative story for the NB model?

NB generative story



$$p(y) \bigcap_{j=1}^{m} p(x_i \mid y)$$

- Pick a label according to p(y)
 - roll a biased, num_labels-sided die
- 2. For each feature:
 - Flip a biased coin:
 - if heads, include the feature
 - if tails, don't include the feature

What about for modeling chicken reviews?

NB decision boundary

label =
$$\operatorname{argmax}_{y \mid labels} p(y) \bigcap_{j=1}^{m} p(x_i \mid y)$$

What does the decision boundary for NB look like if the features are binary?

Some maths

$$label = \log(\operatorname{argmax}_{y\hat{1} | labels} p(y) \bigodot_{j=1}^{m} p(x_i | y))$$

$$= \operatorname{argmax}_{y\hat{1} | labels} \log(p(y)) + \bigotimes_{i=1}^{m} \log(p(x_i | y))$$

$$= \operatorname{argmax}_{y\hat{1} | labels} \log(p(y)) + \bigotimes_{i=1}^{m} x_i \log(p(x_i | y)) + \overline{x}_i \log(1 - p(x_i | y))$$

$$p(x_i | y) = \int_{\hat{T}}^{\hat{T}} q_i & \text{if } x_i = 1 \\ \hat{T} - q_i & \text{otherwise}$$

Some more maths

$$labels = \operatorname{argmax}_{y \hat{1} \ labels} \log(p(y)) + \overset{m}{\underset{i=1}{\overset{m}{\bigcirc}}} x_i \log(p(x_i \mid y)) + \overline{x}_i \log(1 - p(x_i \mid y))$$

$$= \operatorname{argmax}_{y \hat{1} \ labels} \log(p(y)) + \overset{m}{\overset{m}{\bigcirc}} x_i \log(p(x_i \mid y)) + (1 - x_i) \log(1 - p(x_i \mid y))$$

$$= \operatorname{argmax}_{y \hat{1} \ labels} \log(p(y)) + \overset{m}{\overset{m}{\bigcirc}} x_i \log(p(x_i \mid y)) + (1 - x_i) \log(1 - p(x_i \mid y))$$

$$= \operatorname{argmax}_{y \hat{1} \ labels} \log(p(y)) + \overset{m}{\overset{m}{\bigcirc}} x_i \log(p(x_i \mid y)) + (1 - x_i) \log(1 - p(x_i \mid y))$$

$$= \operatorname{argmax}_{y \hat{1} \ labels} \log(p(y)) + \overset{m}{\overset{m}{\bigcirc}} x_i \log(p(x_i \mid y)) + (1 - x_i) \log(1 - p(x_i \mid y))$$

$$= \operatorname{argmax}_{y \hat{1} \ labels} \log(p(x_i \mid y)) + (1 - x_i) \log(1 - p(x_i \mid y))$$

$$= \operatorname{argmax}_{y \hat{1} \ labels} \log(p(x_i \mid y)) + (1 - x_i) \log(1 - p(x_i \mid y))$$

$$= \operatorname{argmax}_{y \hat{1} \ labels} \log(p(x_i \mid y)) + (1 - x_i) \log(1 - p(x_i \mid y))$$

=
$$\operatorname{argmax}_{y \mid labels} \log(p(y)) + \bigotimes_{i=1}^{m} x_i \log(p(x_i \mid y)) - x_i \log(1 - p(x_i \mid y)) + \log(1 - p(x_i \mid y))$$

$$= \operatorname{argmax}_{y \hat{1} \ labels} \log(p(y)) + \mathop{\tilde{o}}_{i=1}^{m} x_i \frac{\log(p(x_i \mid y))}{\log(1 - p(x_i \mid y))} + \log(1 - p(x_i \mid y))$$

And...

$$labels = \operatorname{argmax}_{y \hat{1} \ labels} \log(p(y)) + \bigotimes_{i=1}^{m} x_{i} \frac{\log(p(x_{i} \mid y))}{\log(1 - p(x_{i} \mid y))} + \log(1 - p(x_{i} \mid y))$$

$$= \operatorname{argmax}_{y \hat{1} \ labels} \log(p(y)) + \mathring{a} \log(1 - p(x_i | y)) + \mathring{a} x_i \frac{\log(p(x_i | y))}{\log(1 - p(x_i | y))}$$

What does this look like?

And...

$$labels = \operatorname{argmax}_{y \hat{1} \ labels} \log(p(y)) + \bigotimes_{i=1}^{m} x_i \frac{\log(p(x_i \mid y))}{\log(1 - p(x_i \mid y))} + \log(1 - p(x_i \mid y))$$

$$wx + b$$

Linear model !!!

What are the weights?

NB as a linear model

$$w_i = \frac{\log(p(x_i \mid y))}{\log(1 - p(x_i \mid y))}$$

How likely this feature is to be 1 given the label

How likely this feature is to be 0 given the label

- low weights indicate there isn't much difference
- larger weights (positive or negative) indicate feature is important

Maximum likelihood estimation

Intuitive

Sets the probabilities so as to maximize the probability of the training data

Problems?

- Overfitting!
- Amount of data
 - particularly problematic for rare events
- Is our training data representative

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

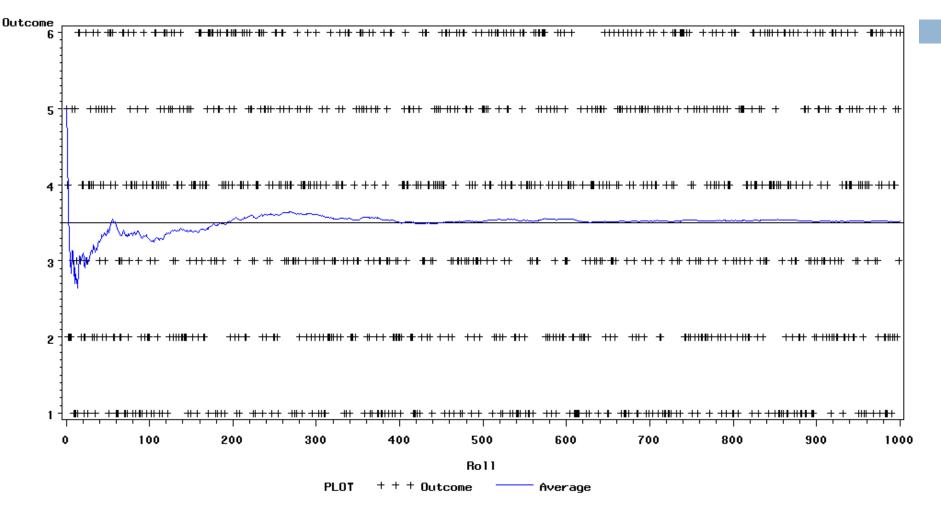
Which model do we use, i.e. how do we calculate p(feature, label)?

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

How do we deal with overfitting?

LAW OF LARGE NUMBERS IN AVERAGE OF DIE ROLLS

AVERAGE CONVERGES TO EXPECTED VALUE OF 3.5



Back to parasitic gaps

Say the actual probability is 1/100,000

We don't know this, though, so we're estimating it from a small data set of 10K sentences

What is the probability that we have a parasitic gap sentence in our sample?

Back to parasitic gaps

 $p(not_parasitic) = 0.99999$

p(not_parasitic)¹⁰⁰⁰⁰ ≈ 0.905 is the probability of us NOT finding one

So, probability of us finding one is ~10%, in which case we would incorrectly assume that the probability is 1/10,000 (10 times too large)

Solutions?

Priors

Coin1 data: 3 Heads and 1 Tail

Coin2 data: 30 Heads and 10 tails

Coin3 data: 2 Tails

Coin4 data: 497 Heads and 503 tails

If someone asked you what the probability of heads was for each of these coins, what would you say?

1: 3/4

2: 3/4

3: 0

4: 497/1000

 You can use Naive Bayes classifier for probabilistic classification task