CSE419 – Artificial Intelligence and Machine Learning 2018

PhD Furkan Gözükara, Toros University

https://github.com/FurkanGozukara/CSE419_2018

Lecture 12

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

- {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 = 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	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 **depended** 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

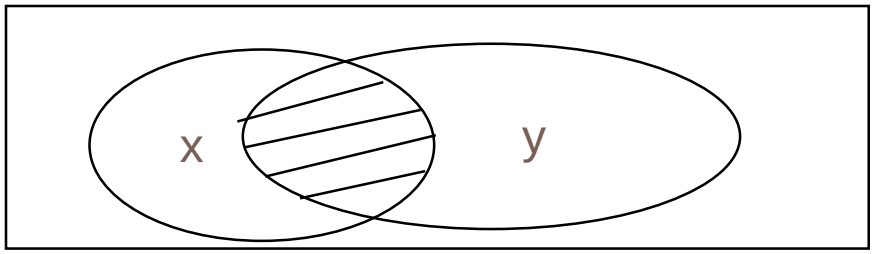
P(X|Y) models this (read "probability of X *given* Y")

- What is the probability of a heads *given* that both sides of the coin are heads?
- What is the probability the document is about 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

Conditional probability

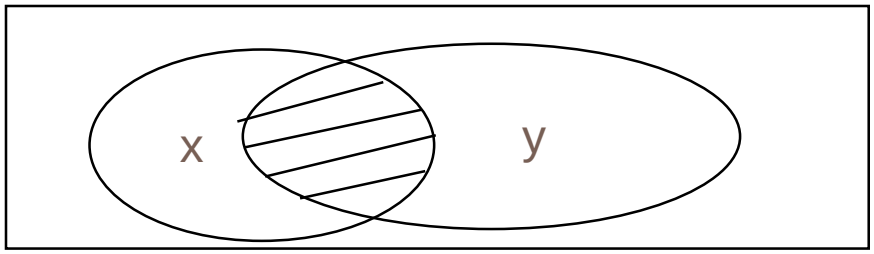
$$p(X \mid Y) = ?$$



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

Conditional probability

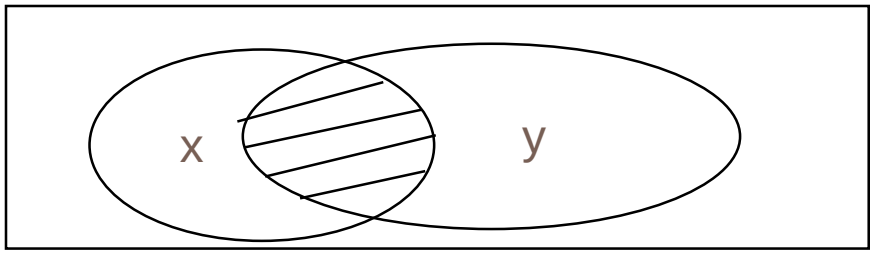
$$p(X \mid Y) = \frac{P(X,Y)}{P(Y)}$$



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.

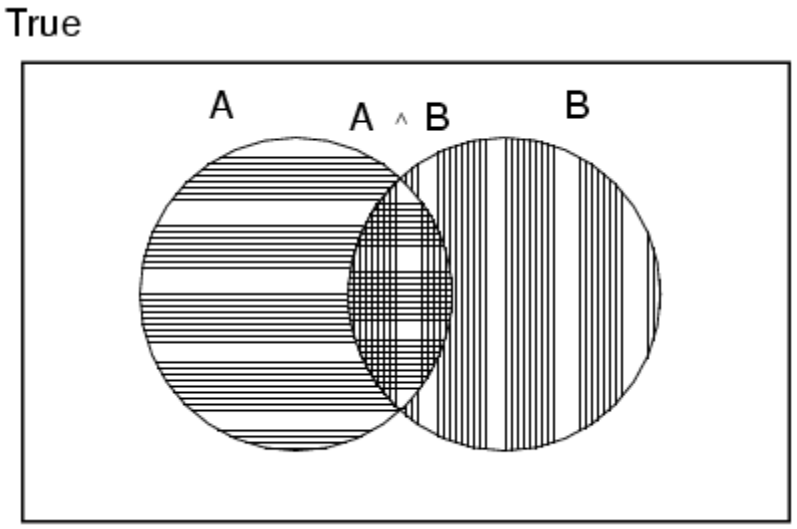
However, when it's clear , we'll often shorten it

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

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

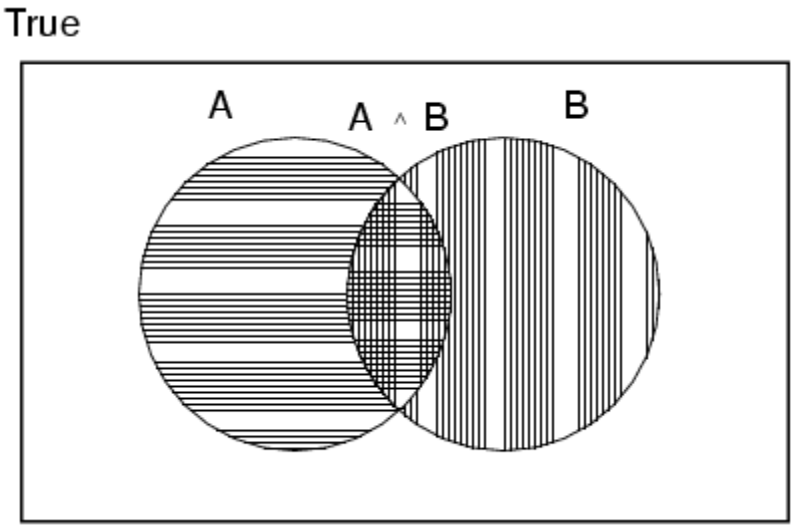
Properties of probabilities

P(*A or B*) = ?



Properties of probabilities

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



Properties of probabilities

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

More generally:

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

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

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

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?

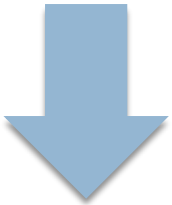
gap

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

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

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

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

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

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

CSE419 – Artificial Intelligence and Machine Learning 2018

PhD Furkan Gözükara, Toros University

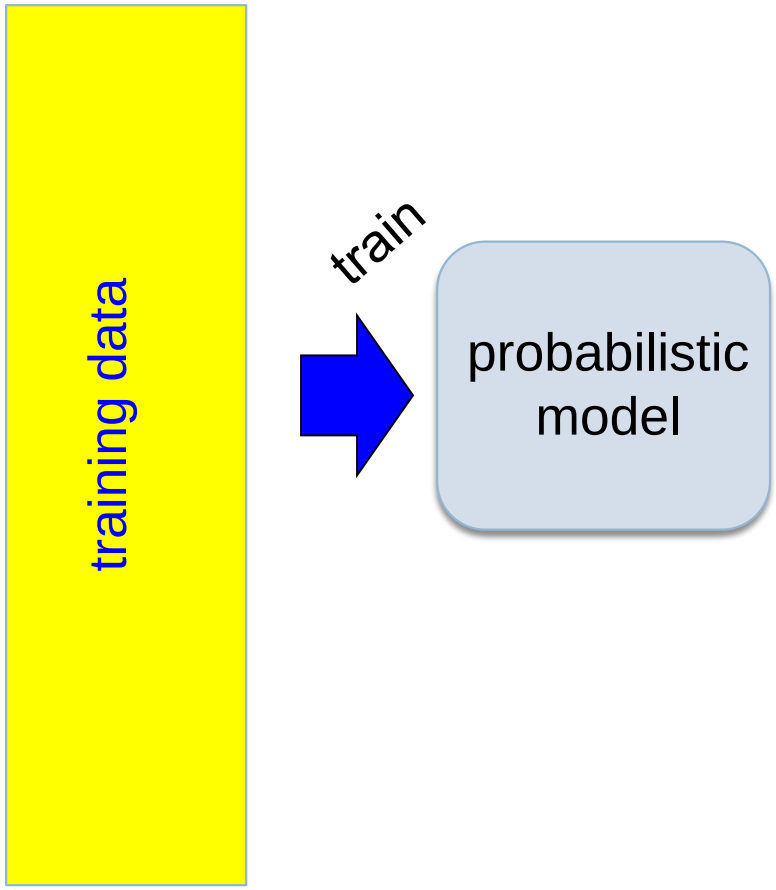
https://github.com/FurkanGozukara/CSE419_2018

Lecture 12.1

Probabilistic Models

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

Probabilistic Modeling



Model the data with a probabilistic model

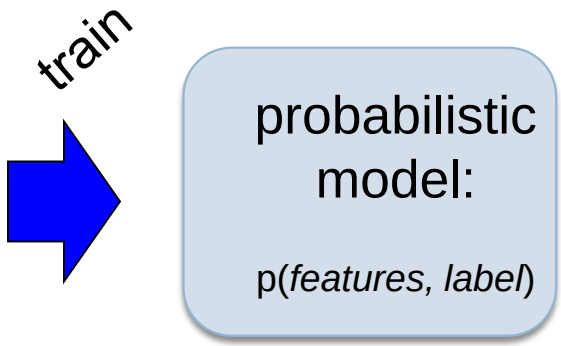
specifically, learn p(*features, label*)

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

An example: classifying fruit

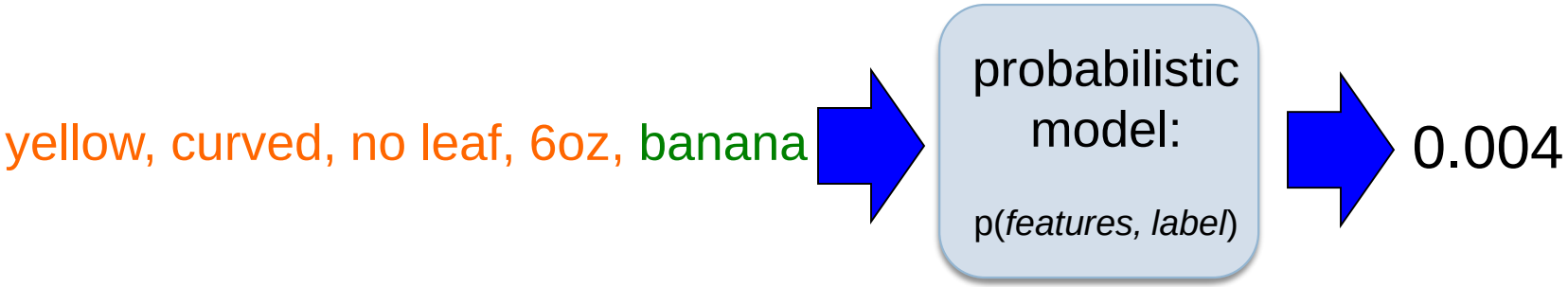
Training data

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



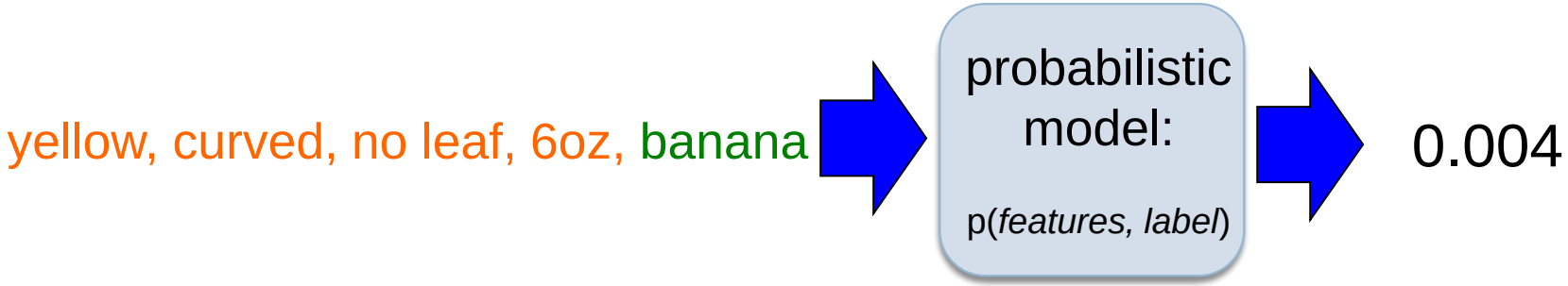
Probabilistic models

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

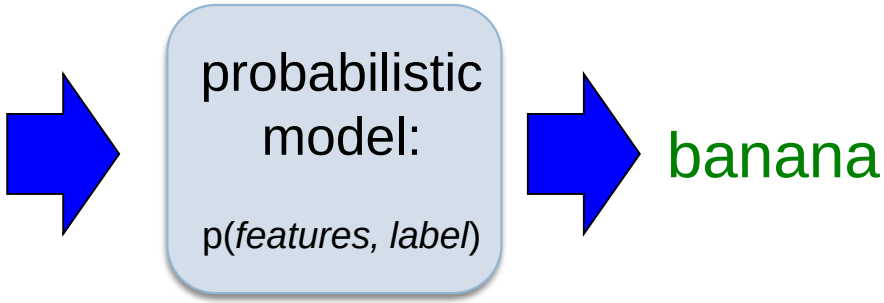


Probabilistic model vs. classifier

Probabilistic model:

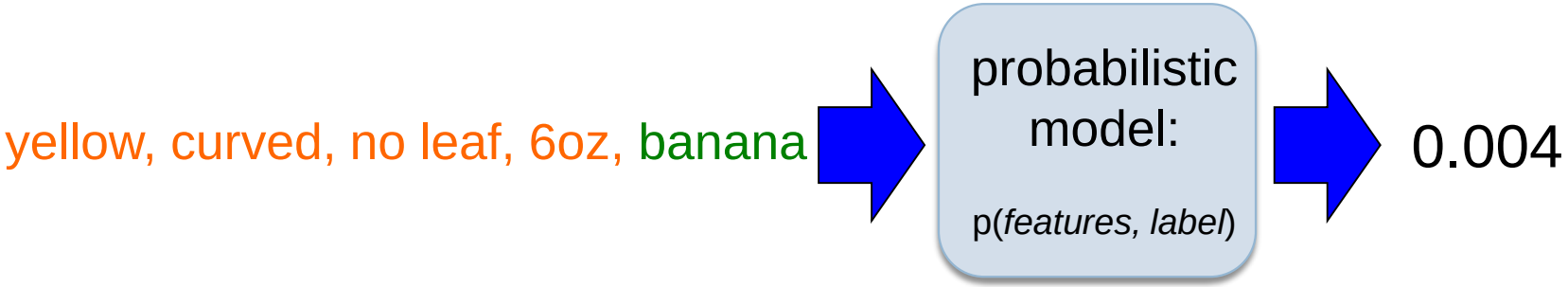


Classifier:



Probabilistic models: classification

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



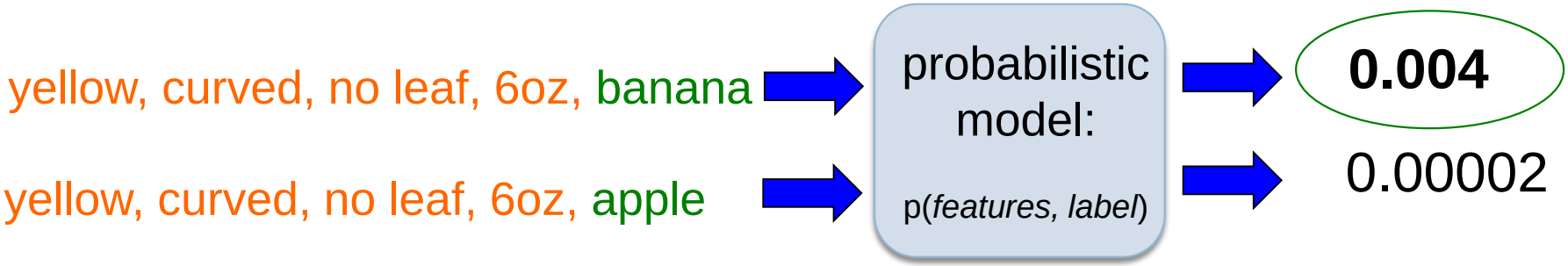
Given an unlabeled example: predict the label

yellow, curved, no leaf, 6oz

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

Probabilistic models

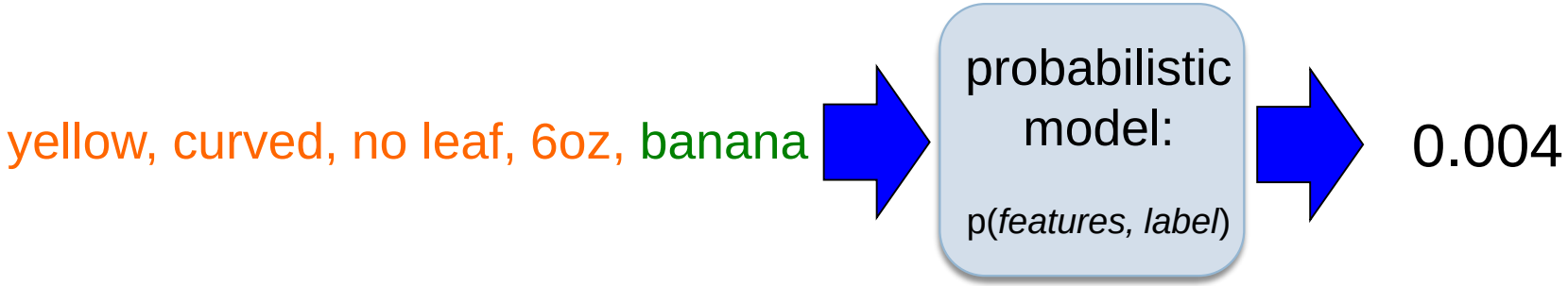
Probabilistic models define a *probability distribution* over features and labels:



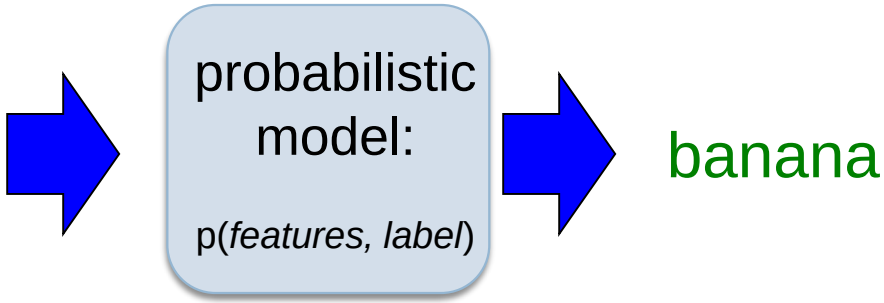
For each label, ask for the probability under the model
Pick the label with the highest probability

Probabilistic model vs. classifier

Probabilistic model:



Classifier:



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

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?

ML in general

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

How do train the model?

How do we deal with overfitting?

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?

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

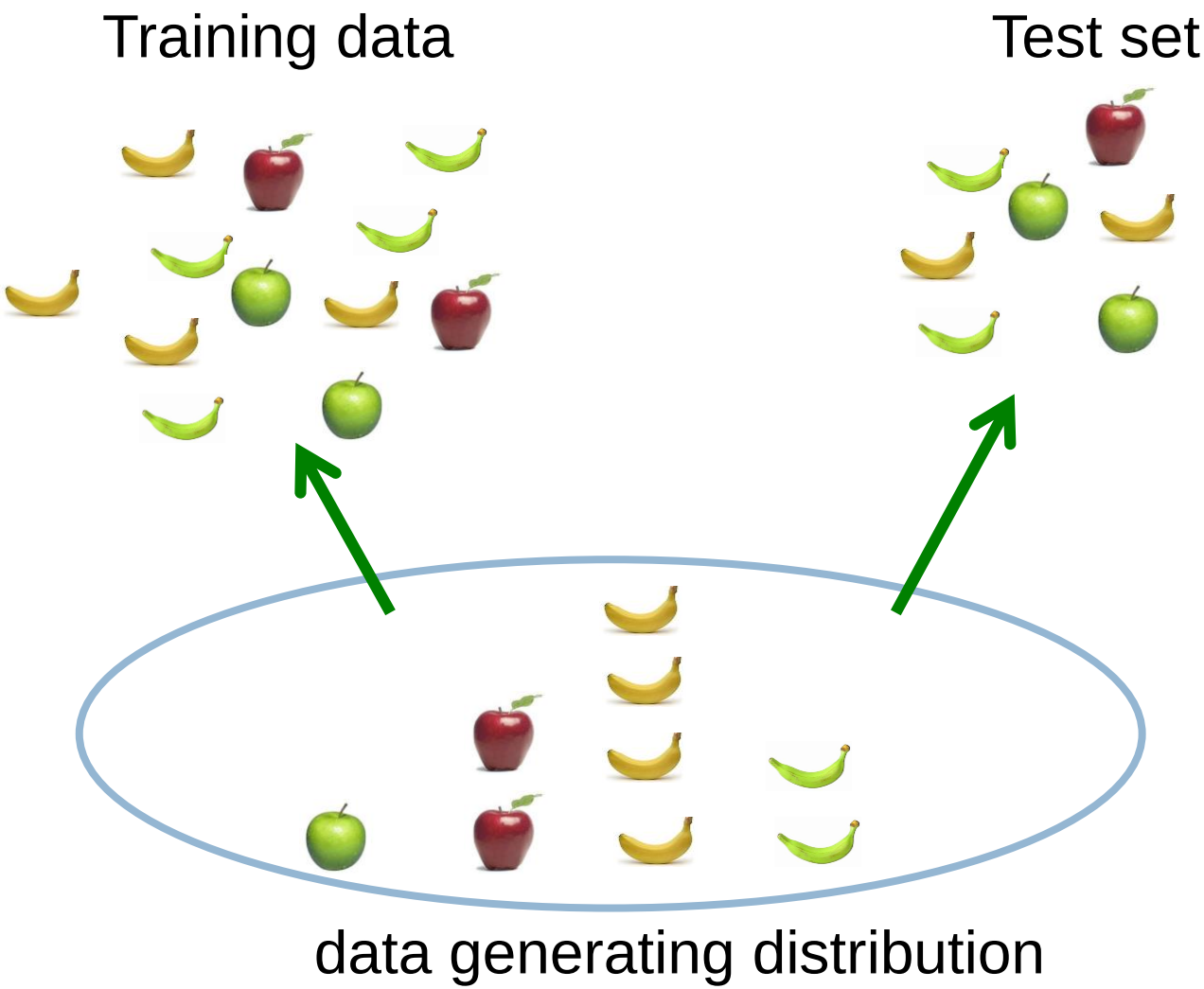
Probabilistic models

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?

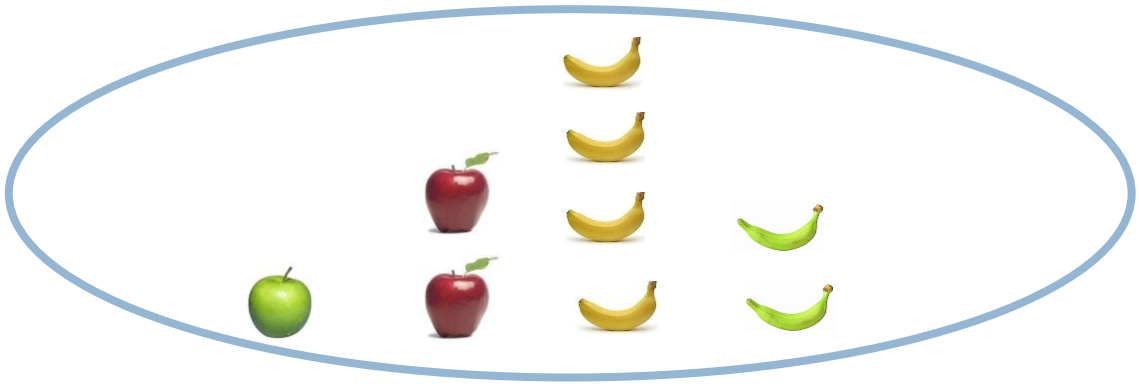
How do we deal with overfitting?

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



data generating distribution

Some maths

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

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

What rule?

Some maths

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

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

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

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

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

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	*
			…		

chicken problem:

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

2^{7000}

16216967556622020264666650854783770951911124303637432562359820841515270231627023529870802378794460004651996019099530984538652557892546513204107022110253564658647431585227076599373340842842722420012281878260072931082617043194484266392077784125099996860169436006660011209817579296678781962552377006552947572566780558092938446272186402161088626008160971328747492043520874011018626908423275017246052311293955235059054544214554772509509096507889478094683592939574112569473438619121529684847434440674120417402088754037186942170155022073539838122429925874353753616104159343594557666561701790904172597025336526662682021808493892812699709528570890696375575414344876088248369941993802415197514510125127043829087280919538476302857811854024099958895964192277601255360491156240349994714416090573084242931396211995367937301294479560024833357073899839202991032234659803895306904298017400980173252106913079712420169633972302183530075897845195258485537108858195631737000743805167411189134617501484521767984296782842287373127422122022517597535994839257029877907706355334790244935435386660512591079567291431216297788784818552292819654176600980398997991681404749384215743515802603811510682864067897304838292203460427757655073776567547507027144662263487685709621261074762705203049488907208978593689047063428548531668665657327174660658185609066484950801276175461457216176955575199211750751406777510449672859082255854777144724233490076402632176089211355256124119453870268029904400183858505767193696897593661213568888386800238409325673807775018914703049621509969838539752071549396339237202875920415172949370790977853625108320092839604807237954887069546621688044652112493076290091990717742355039135117441532973747930089955830518884135334798464113680004999403737245600354288112326328218661131064550772899229969469156018580839820741704606832124388152026099584696588161375826382921029547343888832163627122302921229795384868355483535710603407789177417026363656202726955437517780741313455101810009468809407811220573803353711246329589162370895804762245950918253016369092362406714116443316561598280583720783439888562390892028440902553829376

Any problems with this?

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	*
			…		

- 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

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

So, far we have made NO assumptions about the data

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?

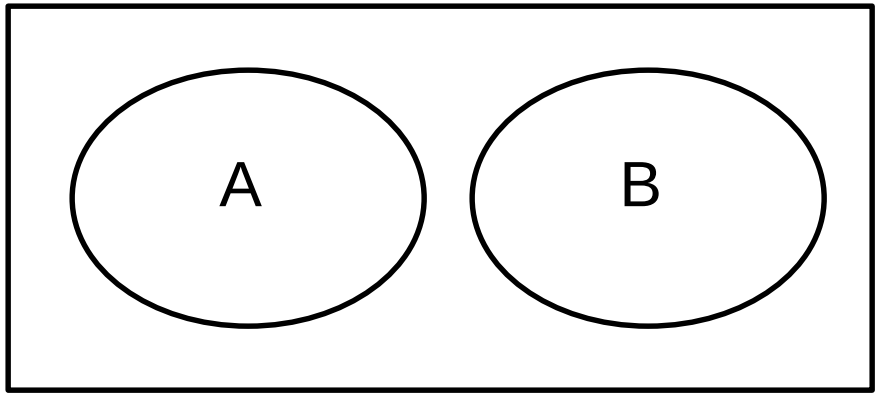
Catching a cold and having cat-allergy

Miles per gallon and driving habits

Height and longevity of life

Independent variables

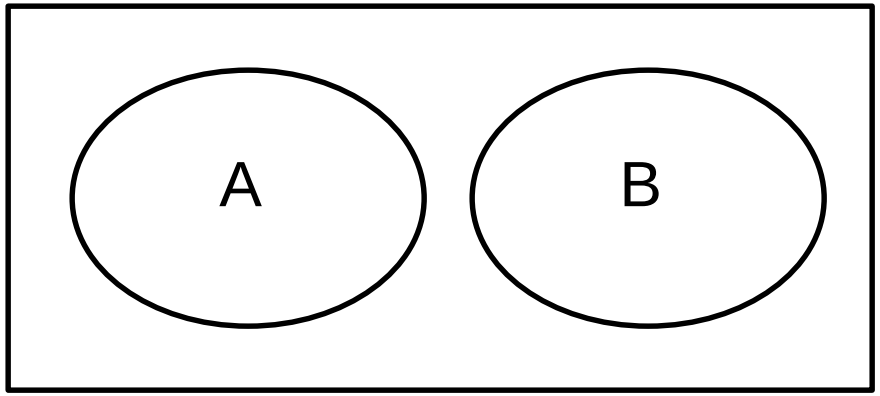
How does independence affect our probability equations/properties?



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

How does independence help us?

Independent variables

If A and B are independent

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

Naïve Bayes assumption

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

$$p(x_i \mid y, x_1, x_2, \ldots, x_{i-1}) = p(x_i \mid y)$$

What does this assume?

Naïve Bayes assumption

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

$$p(x_i \mid y, x_1, x_2, \ldots, x_{i-1}) = p(x_i \mid y)$$

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

For the chicken problem?

Naïve Bayes assumption

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

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?

Naïve Bayes assumption

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

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:

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

Naïve Bayes model

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

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

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:

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

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?

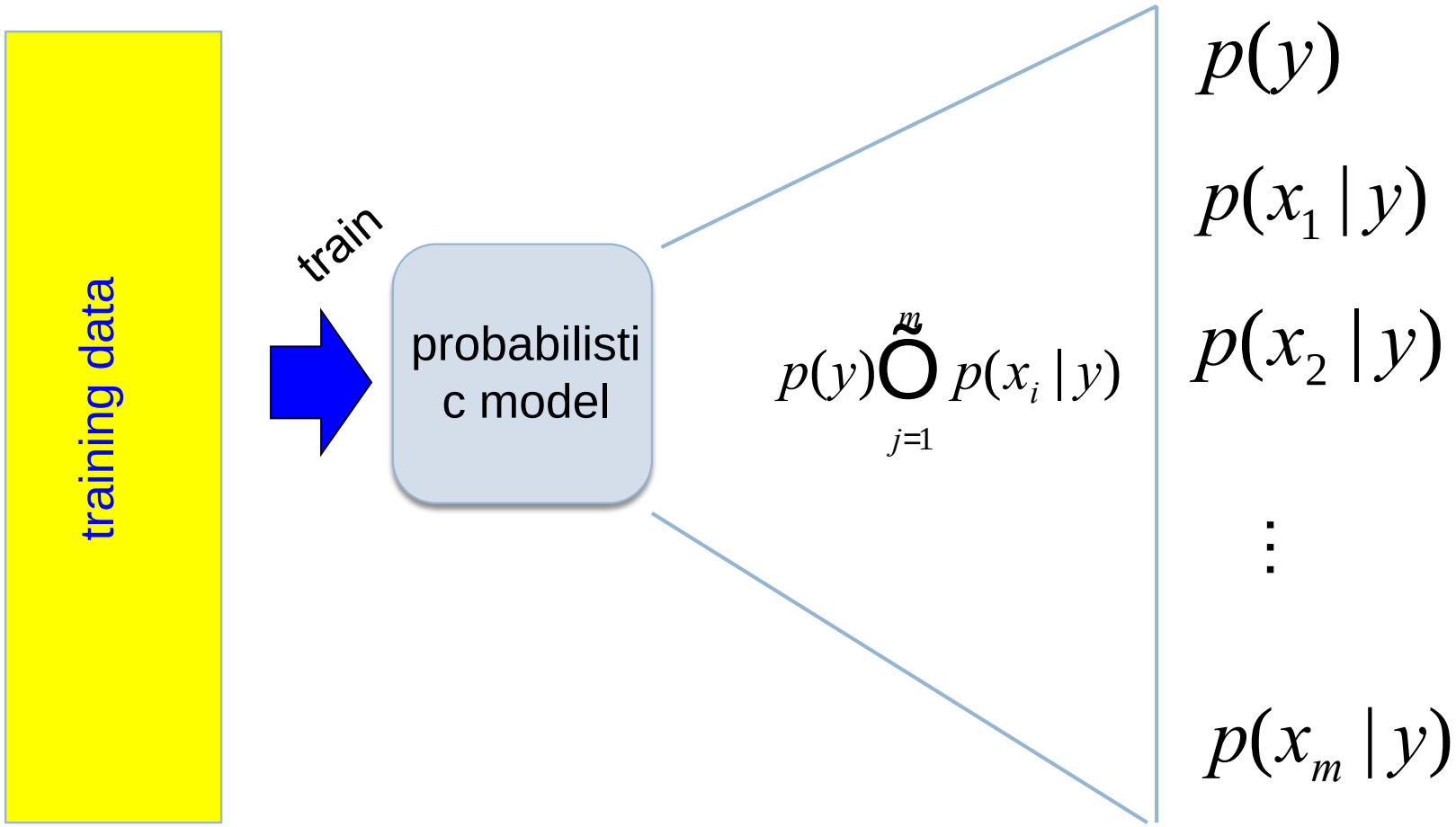
How do we deal with overfitting?

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?

Obtaining probabilities



Estimating probabilities

What is the probability of a chicken meat review?

We don't know!

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

$$\frac{\text{number of review labeled chicken meat}}{\text{total number of reviews}}$$

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

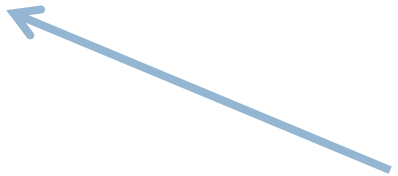
Watch Videos to Understand Better

- 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?

$$likelihood(data) = \tilde{O}_i p(x_i)$$

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=pYxNSUDSFH4&feature=youtu.be

CSE419 – Artificial Intelligence and Machine Learning 2018

PhD Furkan Gözükara, Toros University

https://github.com/FurkanGozukara/CSE419_2018

Lecture 12.2

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

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

for each example

how probable is it under the model

the model parameters (e.g. probability of heads)

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 ?

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

for each example

how probable is it under the model

the model parameters (e.g. probability of heads)

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 ?

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

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

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

$$MLE = \arg\max_q \prod_{i=1}^{n} p_q(x_i)$$

Often easier to work with log-likelihood:

$$MLE = \text{argmax}_q \log\left(\prod_{i=1}^{n} p_q(x_i)\right)$$

Why is this ok?

$$= \text{argmax}_q \sum_{i=1}^{n} \log(p(x_i))$$

Calculating MLE

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

$$MLE = \text{argmax}_q \sum_{i=1}^{n} \log(p(x_i))$$

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.

Calculating MLE

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

$$\text{log-}\ likelihood = \sum_{i=1}^{n} \log(p(x_i))$$

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

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

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

How do we find the max?

Calculating MLE

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!

Calculating MLE

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?

p(heads) = 0.6

$$\log\text{-}\ likelihood = \sum_{i=1}^{n} \log(p(x_i))$$

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

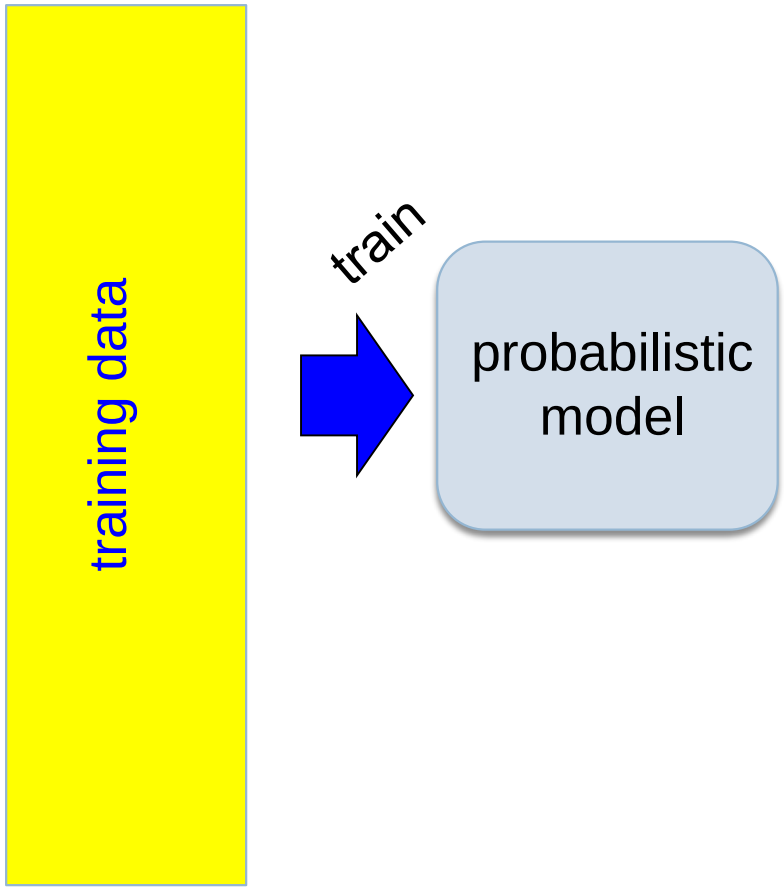
p(heads) = 0.5

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

p(heads) = 0.7

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

MLE estimation for NB



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

$$p(y) \qquad p(x_i \mid y)$$

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

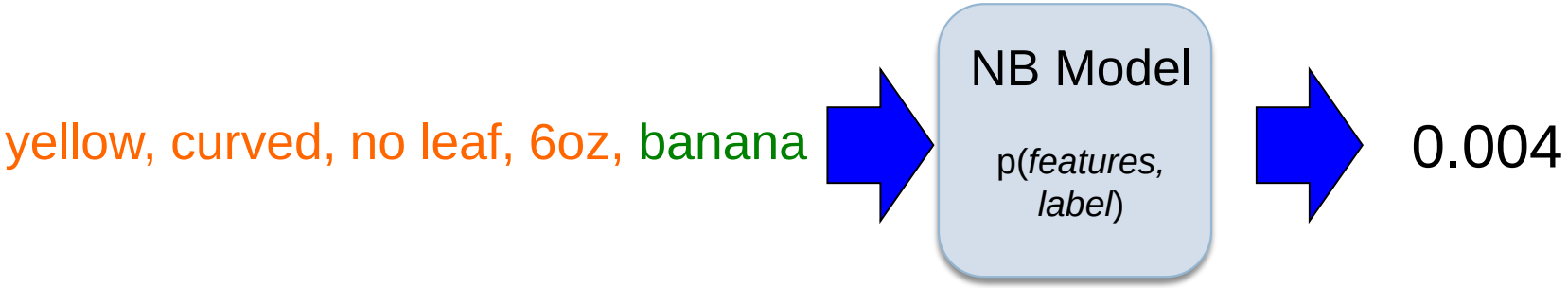
$$p(x_i \mid y) = \frac{count(x_i, y)}{count(y)}$$

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



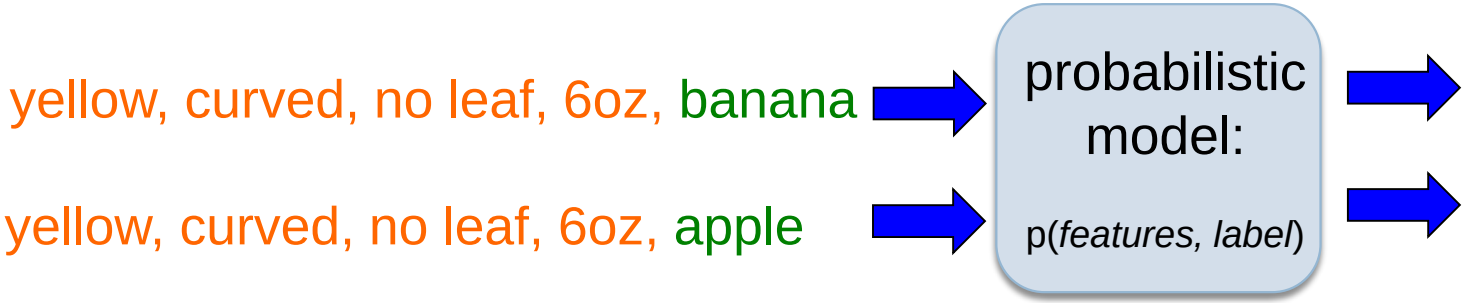
$$p(y)\prod_{j=1}^{m} p(x_i \mid y)$$

Given an unlabeled example: predict the label

yellow, curved, no leaf, 6oz

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

Probabilistic models



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

label = $$\arg\max_{y \in \text{labels}} p(y)\prod_{j=1}^{m} p(x_i \mid y)$$

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

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

What is the generative story for the NB model?

NB generative story

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

1. 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 wine reviews?

NB decision boundary

$$\text{label} = \text{argmax}_{y \in \text{labels}} \; p(y) \prod_{j=1}^{m} p(x_i \mid y)$$

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

Some maths

$$label = \log(\text{argmax}_{y \in labels}\, p(y) \prod_{j=1}^{m} p(x_i \mid y))$$

$$= \text{argmax}_{y \in labels} \log(p(y)) + \sum_{i=1}^{m} \log(p(x_i \mid y))$$

$$= \text{argmax}_{y \in labels} \log(p(y)) + \sum_{i=1}^{m} x_i \log(p(x_i \mid y)) + \overline{x}_i \log(1 - p(x_i \mid y))$$

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

Some more maths

$$labels = \text{argmax}_{y \in labels} \log(p(y)) + \sum_{i=1}^{m} x_i \log(p(x_i \mid y)) + \bar{x}_i \log(1 - p(x_i \mid y))$$

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

(because x_i are binary)

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

$$= \text{argmax}_{y \in labels} \log(p(y)) + \sum_{i=1}^{m} x_i \frac{\log(p(x_i \mid y))}{\log(1 - p(x_i \mid y))} + \log(1 - p(x_i \mid y)$$

And…

$$labels = \text{argmax}_{y \in labels} \log(p(y)) + \sum_{i=1}^{m} x_i \frac{\log(p(x_i \mid y))}{\log(1 - p(x_i \mid y))} + \log(1 - p(x_i \mid y)$$

$$= \text{argmax}_{y \in labels} \log(p(y)) + \sum_{i=1}^{m} \log(1 - p(x_i \mid y)) + \sum_{i=1}^{m} x_i \frac{\log(p(x_i \mid y))}{\log(1 - p(x_i \mid y))}$$

What does this look like?

And…

$$labels = \text{argmax}_{y \in labels} \log(p(y)) + \sum_{i=1}^{m} x_i \frac{\log(p(x_i \mid y))}{\log(1 - p(x_i \mid y))} + \log(1 - p(x_i \mid y)$$

$$= \text{argmax}_{y \in labels} \underbrace{\log(p(y)) + \sum_{i=1}^{m} \log(1 - p(x_i \mid y))}_{b} + \sum_{i=1}^{m} x_i \frac{\log(p(x_i \mid y))}{\log(1 - p(x_i \mid y))}$$

b + $x_i * w_i$

w x + b

What are the weights?

Linear model !!!

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

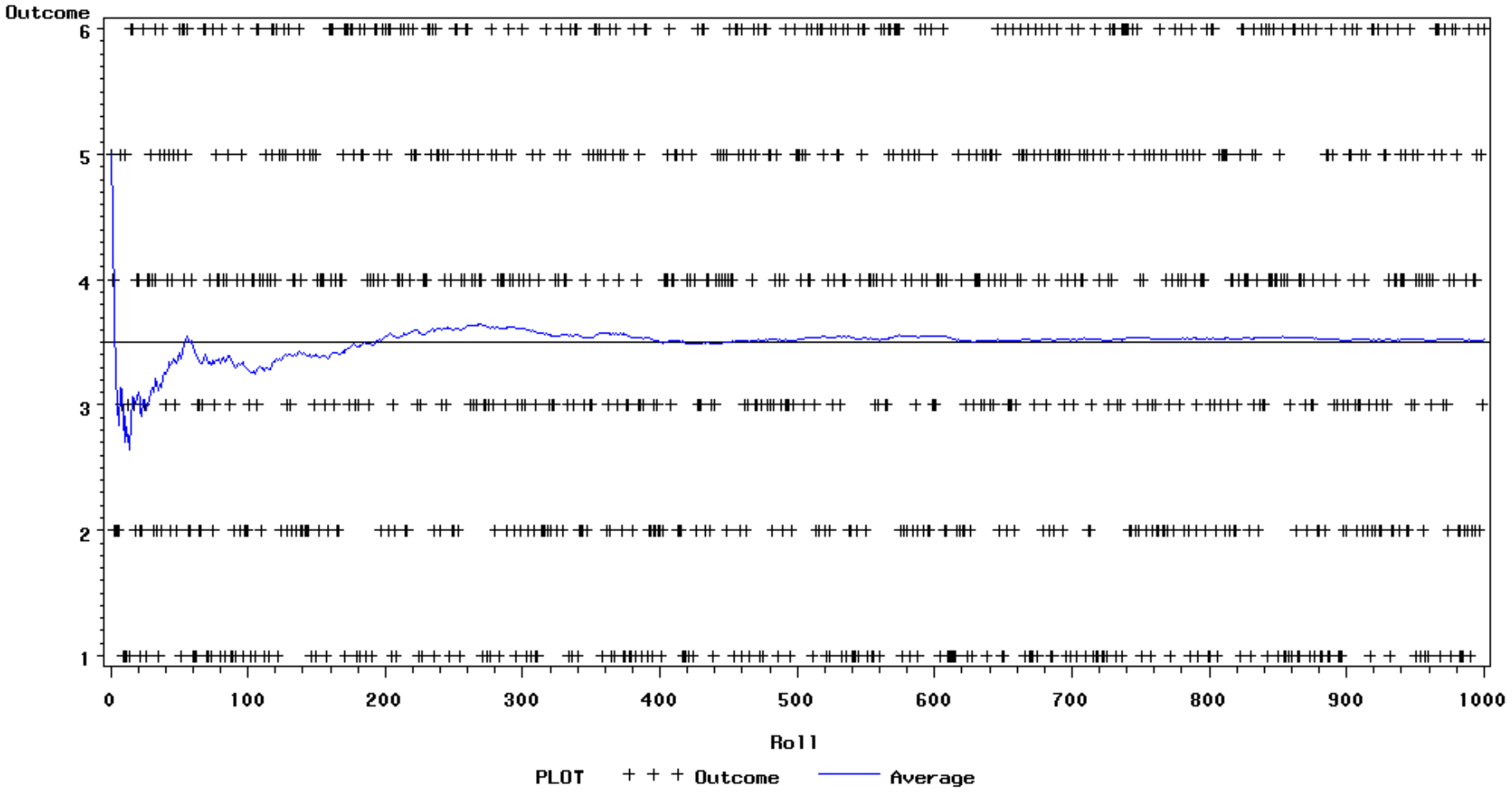
Which model do we use, i.e. how do we calculate p(*feature, label*)?

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

$p(\text{not\_parasitic})^{10000} \approx 0.905$ 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: ¾

2: ¾

3: 0

4: 497/1000

CSE419 – Artificial Intelligence and Machine Learning 2018

PhD Furkan Gözükara, Toros University

https://github.com/FurkanGozukara/CSE419_2018

Lecture 12.3

Logistic Regression

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

Good Regression Videos

- An Introduction to Linear Regression Analysis > https://youtu.be/zPG4NjIkCjc
- How to calculate linear regression using least square method > https://youtu.be/JvS2triCgOY
- How to Calculate R Squared Using Regression Analysis > https://youtu.be/w2FKXOa0HGA
- Standard Error of the Estimate used in Regression Analysis (Mean Square Error) > https://youtu.be/r-txC-dpI-E

Good Regression Videos

- Statistics 101: Logistic Regression, An Introduction > https://youtu.be/zAULhNrnuL4
- Statistics 101: Logistic Regression Probability, Odds, and Odds Ratio > https://youtu.be/ckkiG-SDuV8
- Linear Regression Playlist > https://www.youtube.com/playlist?list=PLIeGtxpvyG-LoKUpV0fSY8BGKIMIdmfCi
- Logistic Regression Playlist > https://www.youtube.com/playlist?list=PLIeGtxpvyG-JmBQ9XoFD4rs-b3hkcX7Uu

Good Regression Videos

- StatQuest: Linear Models Pt.1 - Linear Regression > https://youtu.be/nk2CQITm_eo
- StatQuest: Linear Models Pt.1.5 - Multiple Regression > https://youtu.be/zITIFTsivN8
- StatQuest: Logistic Regression > https://youtu.be/yIYKR4sgzI8

Training revisited

From a probability standpoint, what we're really doing when we're training the model is selecting the Θ that maximizes:

$$p(q \mid data)$$

i.e.

$$\text{argmax}_q \, p(q \mid data)$$

That we pick the most likely model parameters given the data

Estimating revisited

We can incorporate a prior belief in what the probabilities might be

To do this, we need to break down our probability

$$p(q \mid data) = ?$$

(Hint: Bayes rule)

Estimating revisited

What are each of these probabilities?

$$p(\theta \mid data) = \frac{p(data \mid \theta)p(\theta)}{p(data)}$$

Priors

likelihood of the data under the model

probability of different parameters, call the prior

$$p(\theta \mid data) = \frac{p(data \mid \theta)p(\theta)}{p(data)}$$

probability of seeing the data (regardless of model)

Priors

$$q = \text{argmax}_q \frac{p(data \mid q)p(q)}{p(data)}$$

Does p(data) matter for the argmax?

Priors

likelihood of the data under the model

probability of different parameters, call the prior

$$q = \text{argmax}_q \, p(data \mid q) p(q)$$

What does MLE assume for a prior on the model parameters?

Priors

likelihood of the data under the model

probability of different parameters, call the prior

$$q = \text{argmax}_q \, p(data \mid q) p(q)$$

- Assumes a uniform prior, i.e. all Θ are equally likely!
- Relies solely on the likelihood

A better approach

$$q = \text{argmax}_q \, p(data \mid q) p(q)$$

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

We can use any distribution we'd like

This allows us to impart addition bias into the model

Another view on the prior

Remember, the max is the same if we take the log:

$$\theta = \text{argmax}_\theta \log(p(data \mid \theta)) + \log(p(\theta))$$

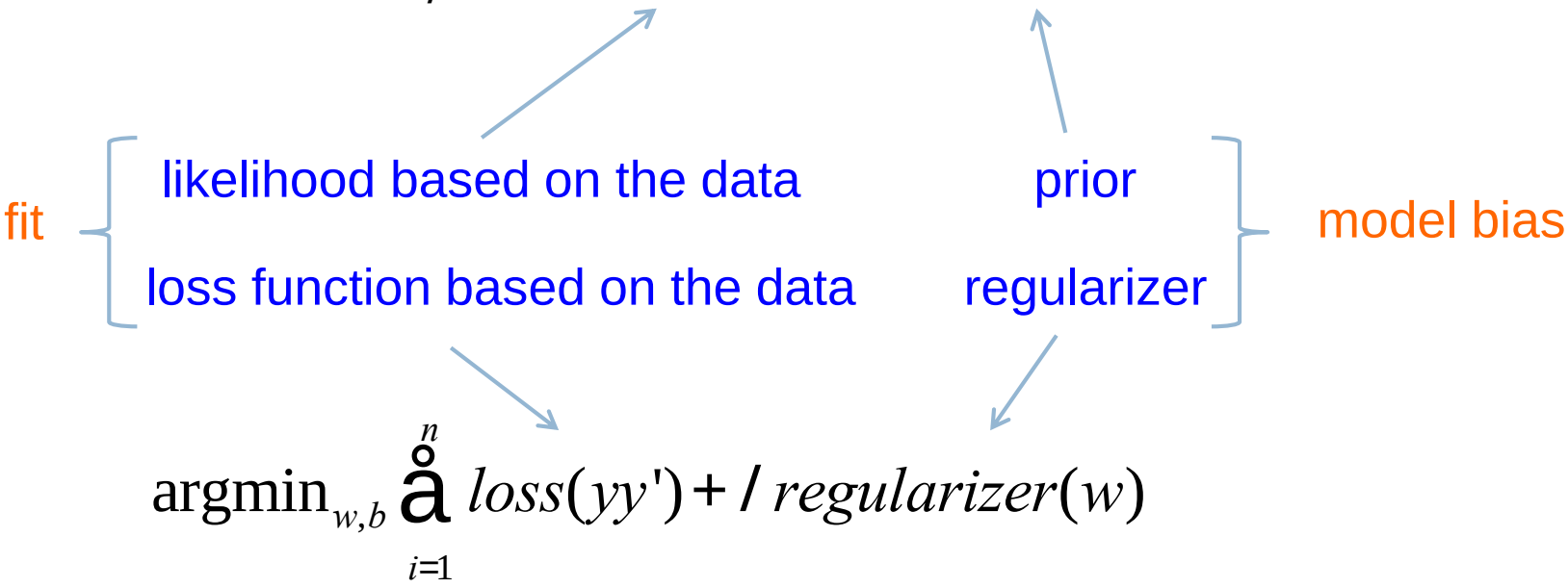
$$\log\text{-}\ likelihood = \sum_{i=1}^{n} \log(p(x_i))$$

We can use any distribution we'd like
This allows us to impart addition bias into the model

Does this look like something we've seen before?

Regularization vs prior

$$q = \text{argmax}_q \log(p(data \mid q)) + \log(p(q))$$

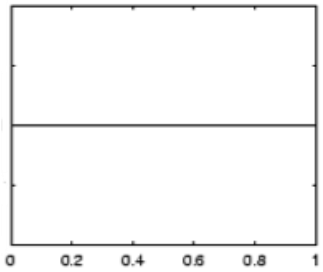


$$\text{argmin}_{w,b} \sum_{i=1}^{n} loss(yy') + \lambda\, regularizer(w)$$

Prior for NB

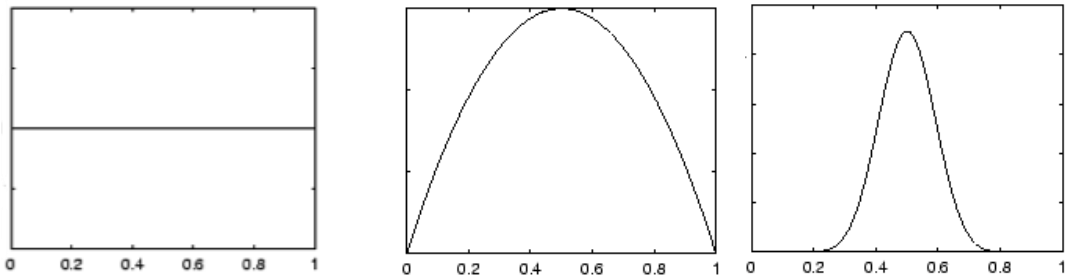
$$q = \text{argmax}_q \log(p(data \mid q)) + \log(p(q))$$

Uniform prior



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

Dirichlet prior



λ= 0 → increasing

$$p(x_i \mid y) = \frac{count(x_i, y) + l}{count(y) + possible\_values\_of\_x_i * l}$$

Prior: another view

$$p(x_1, x_2, \ldots, x_m, y) = p(y) \prod_{j=1}^{m} p(x_i \mid y)$$

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

What happens to our likelihood if, for one of the labels, we never saw a particular feature?

Goes to 0!

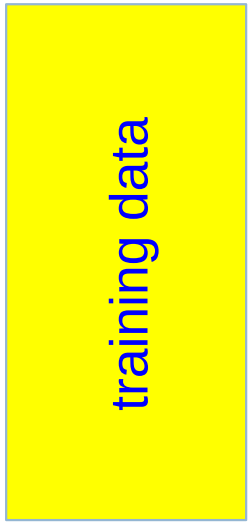
Prior: another view

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

$$p(x_i \mid y) = \frac{count(x_i, y) + l}{count(y) + possible\_values\_of\_x_i * l}$$

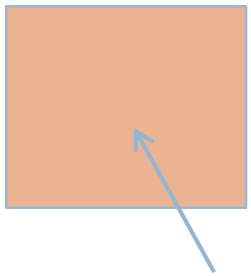
Adding a prior avoids this!

Smoothing



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

$$p(x_i \mid y) = \frac{count(x_i, y) + l}{count(y) + possible\_values\_of\_x_i * l}$$



for each label, pretend like we've seen each feature value occur in λ additional examples

Sometimes this is also called smoothing because it is seen as smoothing or interpolating between the MLE and some other distribution

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?

Joint models vs conditional models

We've been trying to model the joint distribution (i.e. the data generating distribution):

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

However, if all we're interested in is classification, why not directly model the conditional distribution:

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

A first try: linear

$$p(y \mid x_1, x_2, ..., x_m) = x_1 w_1 + w_2 x_2 + ... + w_m x_m + b$$

Any problems with this?

- Nothing constrains it to be a probability
- Could still have combination of features and weight that exceeds 1 or is below 0

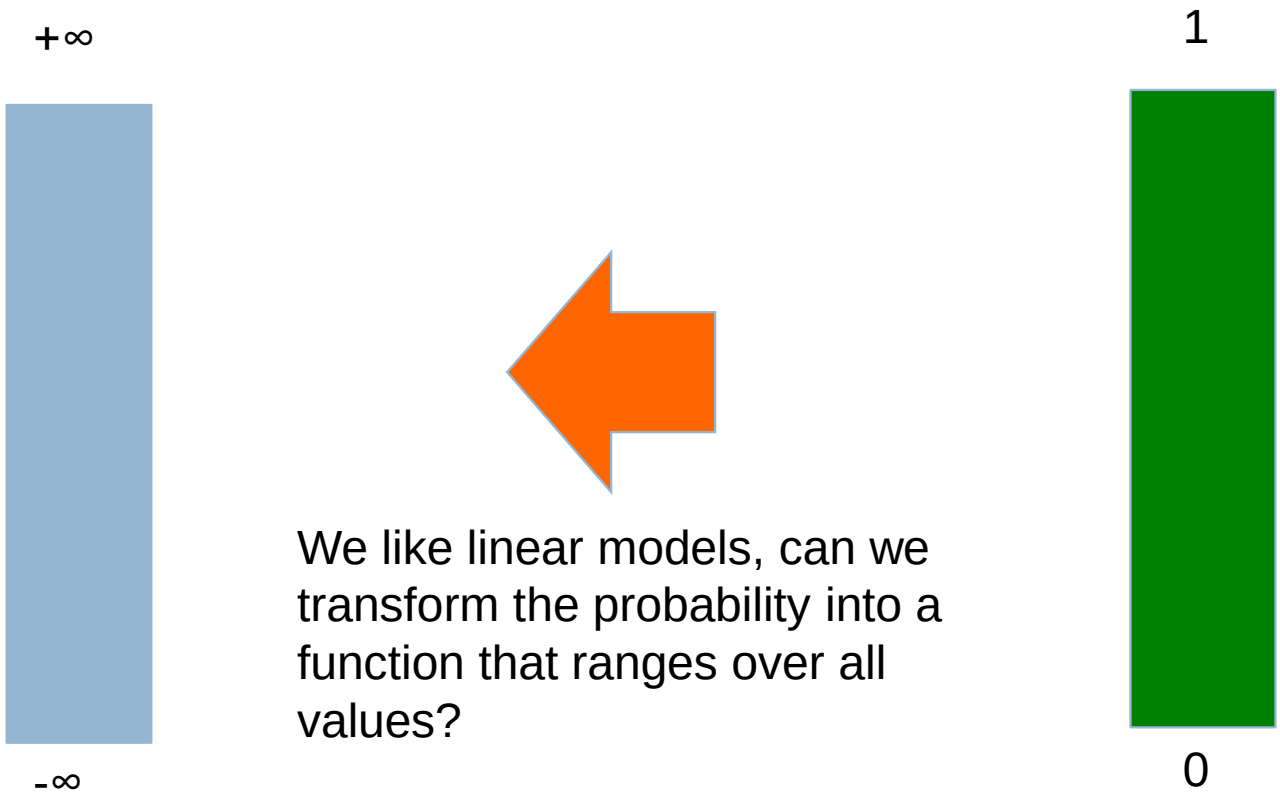
The challenge

$$x_1 w_1 + w_2 x_2 + ... + w_m x_m + b$$

Linear model

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

probability



Odds ratio

Rather than predict the probability, we can predict the ratio of 1/0 (positive/negative)

Predict the **odds** that it is 1 (true): How much more likely is 1 than 0.

Does this help us?

$$\frac{P(1 \mid x_1, x_2, \ldots, x_m)}{P(0 \mid x_1, x_2, \ldots, x_m)} = \frac{P(1 \mid x_1, x_2, \ldots, x_m)}{1 - P(1 \mid x_1, x_2, \ldots, x_m)} = x_1 w_1 + w_2 x_2 + \ldots + w_m x_m + b$$

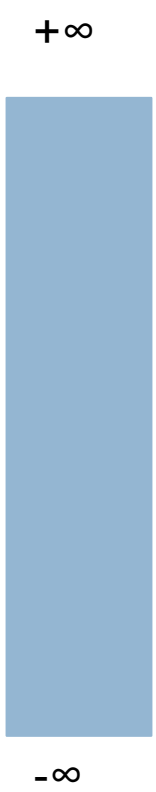
Odds ratio

$$x_1 w_1 + w_2 x_2 + \ldots + w_m x_m + b$$

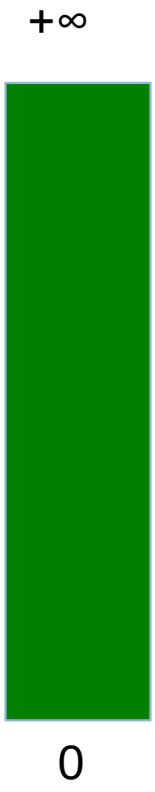
Linear model

$$\frac{P(1 \mid x_1, x_2, \ldots, x_m)}{1 - P(1 \mid x_1, x_2, \ldots, x_m)}$$

odds ratio



Where is the dividing line between class 1 and class 0 being selected?



Odds ratio

$$P(1 \mid x_1, x_2, \ldots, x_m) > P(0 \mid x_1, x_2, \ldots, x_m)$$

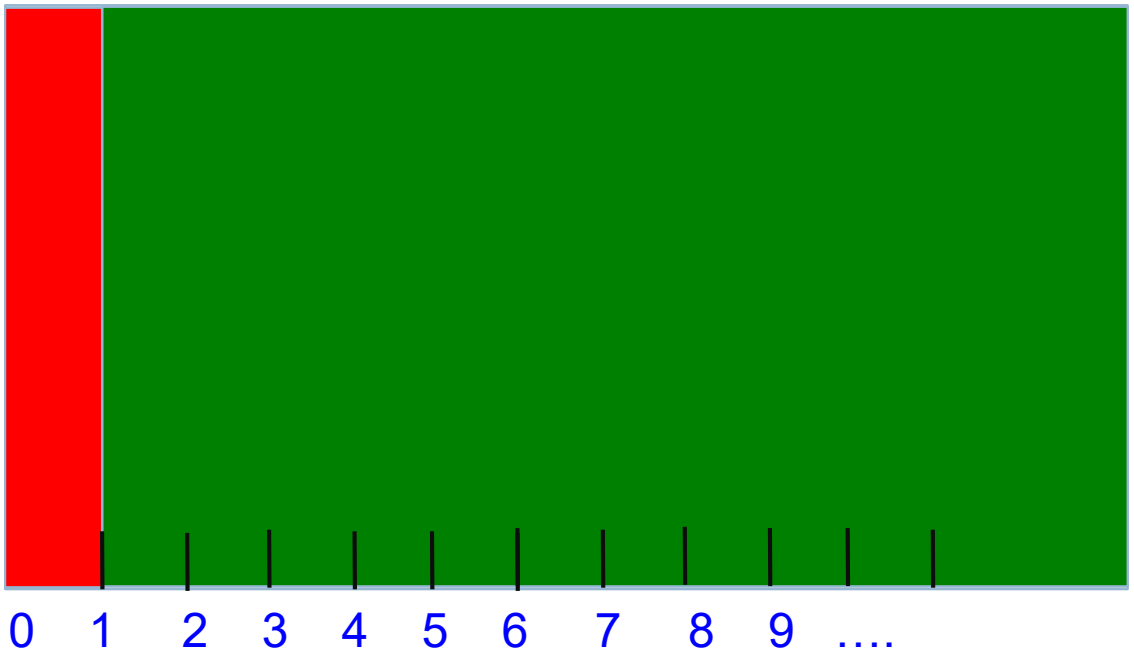
$$\frac{P(1 \mid x_1, x_2, \ldots, x_m)}{1 - P(1 \mid x_1, x_2, \ldots, x_m)}$$

$$P(1 \mid x_1, x_2, \ldots, x_m) > 1 - P(1 \mid x_1, x_2, \ldots, x_m)$$

We're trying to find some transformation that transforms the odds ratio to a number that is -∞ to +∞

Does this suggest another transformation?

odds ratio



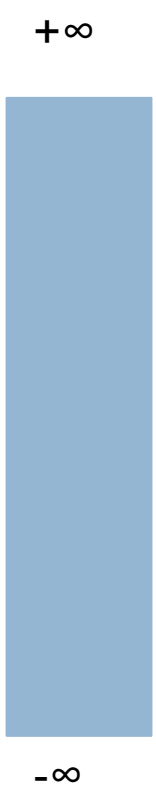
Log odds (logit function)

$$x_1 w_1 + w_2 x_2 + \ldots + w_m x_m + b$$

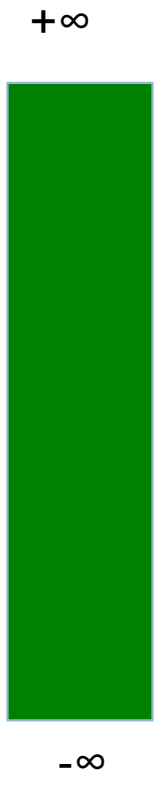
Linear regression

$$\log \frac{P(1 \mid x_1, x_2, \ldots, x_m)}{1 - P(1 \mid x_1, x_2, \ldots, x_m)}$$

odds ratio



How do we get the probability of an example?



Log odds (logit function)

$$\log \frac{P(1 \mid x_1, x_2, \ldots, x_m)}{1 - P(1 \mid x_1, x_2, \ldots, x_m)} = w_1 x_2 + w_2 x_2 + \ldots + w_m x_m + b$$

$$\frac{P(1 \mid x_1, x_2, \ldots, x_m)}{1 - P(1 \mid x_1, x_2, \ldots, x_m)} = e^{w_1 x_2 + w_2 x_2 + \ldots + w_m x_m + b}$$

$$P(1 \mid x_1, x_2, \ldots, x_m) = (1 - P(1 \mid x_1, x_2, \ldots, x_m)) e^{w_1 x_2 + w_2 x_2 + \ldots + w_m x_m + b}$$

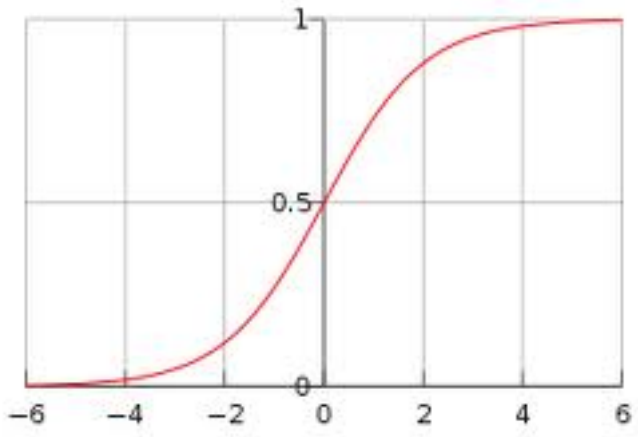
. . .

$$P(1 \mid x_1, x_2, \ldots, x_m) = \frac{1}{1 + e^{-(w_1 x_2 + w_2 x_2 + \ldots + w_m x_m + b)}}$$

anyone recognize this?

Logistic function

$$\text{logistic} = \frac{1}{1+e^{-x}}$$



Logistic regression

How would we classify examples once we had a trained model?

$$\log \frac{P(1 \mid x_1, x_2, \ldots, x_m)}{1 - P(1 \mid x_1, x_2, \ldots, x_m)} = w_1 x_2 + w_2 x_2 + \ldots + w_m x_m + b$$

If the sum > 0 then p(1)/p(0) > 1, so positive

if the sum < 0 then p(1)/p(0) < 1, so negative

Still a *linear* classifier (decision boundary is a line)

Training logistic regression models

How should we learn the parameters for logistic regression (i.e. the w's)?

$$\log \frac{P(1 \mid x_1, x_2, \ldots, x_m)}{1 - P(1 \mid x_1, x_2, \ldots, x_m)} = w_1 x_2 + w_2 x_2 + \ldots + w_m x_m + b$$

parameters

$$P(1 \mid x_1, x_2, \ldots, x_m) = \frac{1}{1 + e^{-(w_1 x_2 + w_2 x_2 + \ldots + w_m x_m + b)}}$$

MLE logistic regression

Find the parameters that maximize the likelihood (or log-likelihood) of the data:

$$\log\text{-}\ likelihood = \sum_{i=1}^{n} \log(p(x_i))$$

$$= \sum_{i=1}^{n} \log\left(\frac{1}{1+e^{-y_i(w_1x_2+w_2x_2+\ldots+w_mx_m+b)}}\right)$$

assume labels 1, -1

$$= \sum_{i=1}^{n} -\log(1+e^{-y_i(w_1x_2+w_2x_2+\ldots+w_mx_m+b)})$$

MLE logistic regression

$$\log\text{-}\ likelihood = \sum_{i=1}^{n} -\log(1+e^{-y_i(w_1x_2+w_2x_2+\ldots+w_mx_m+b)})$$

We want to maximize, i.e.

$$MLE(data) = \text{argmax}_{w,b} \log\text{-}\ likelihood(data)$$

$$= \text{argmax}_{w,b} \sum_{i=1}^{n} -\log(1+e^{-y_i(w_1x_2+w_2x_2+\ldots+w_mx_m+b)})$$

$$= \arg\min_{w,b} \sum_{i=1}^{n} \log(1+e^{-y_i(w_1x_2+w_2x_2+\ldots+w_mx_m+b)})$$

Look familiar? Hint: anybody read the book?

MLE logistic regression

$$\text{argmin}_{w,b} \sum_{i=1}^{n} \log(1+e^{-y_i(w_1x_2+w_2x_2+\ldots+w_mx_m+b)})$$

Surrogate loss functions:

Zero/one:	$\ell^{(0/1)}(y,\hat{y}) = 1[y\hat{y} \le 0]$
Hinge:	$\ell^{(\text{hin})}(y,\hat{y}) = \max\{0, 1-y\hat{y}\}$
Logistic:	$\ell^{(\log)}(y,\hat{y}) = \frac{1}{\log 2}\log(1+\exp[-y\hat{y}])$
Exponential:	$\ell^{(\exp)}(y,\hat{y}) = \exp[-y\hat{y}]$
Squared:	$\ell^{(\text{sqr})}(y,\hat{y}) = (y-\hat{y})^2$

logistic regression: three views

$$\log \frac{P(1 \mid x_1, x_2, \ldots, x_m)}{1 - P(1 \mid x_1, x_2, \ldots, x_m)} = w_0 + w_1 x_2 + w_2 x_2 + \ldots + w_m x_m$$

linear classifier

$$P(1 \mid x_1, x_2, \ldots, x_m) = \frac{1}{1 + e^{-(w_0 + w_1 x_2 + w_2 x_2 + \ldots + w_m x_m)}}$$

conditional model logistic

$$\text{argmin}_{w,b} \sum_{i=1}^{n} \log(1 + e^{-y_i (w_1 x_2 + w_2 x_2 + \ldots + w_m x_m + b)})$$

linear model minimizing logistic loss

Overfitting

$$\text{argmin}_{w,b} \sum_{i=1}^{n} \log(1 + e^{-y_i(w_1 x_2 + w_2 x_2 + \ldots + w_m x_m + b)})$$

If we minimize this loss function, in practice, the results aren't great and we tend to overfit

Solution?

Regularization/prior

$$\text{argmin}_{w,b} \sum_{i=1}^{n} \log(1+e^{-y_i(w_1x_2+w_2x_2+\ldots+w_mx_m+b)}) + \lambda\, regularizer(w,b)$$

or

$$\text{argmin}_{w,b} \sum_{i=1}^{n} \log(1+e^{-y_i(w_1x_2+w_2x_2+\ldots+w_mx_m+b)}) - \log(p(w,b))$$

What are some of the regularizers we know?

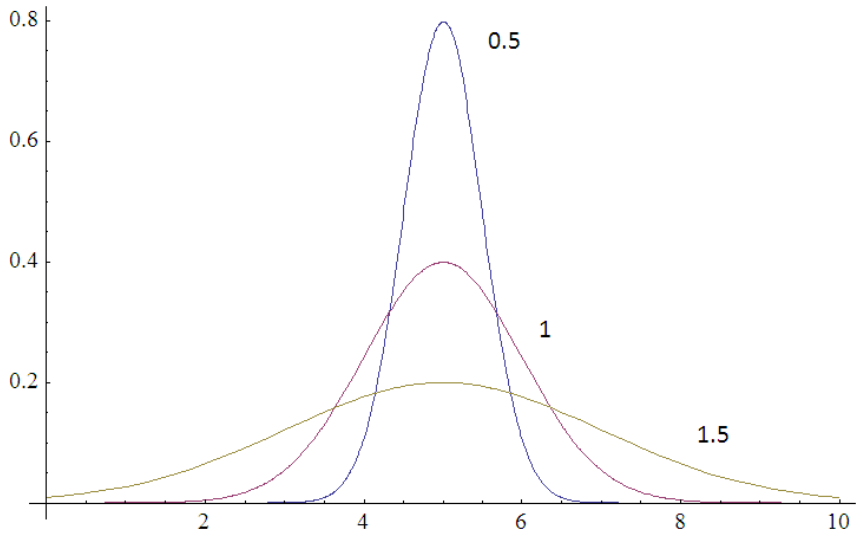
Regularization/prior

L2 regularization:

$$\text{argmin}_{w,b} \sum_{i=1}^{n} \log(1+e^{-y_i(w_1x_2+w_2x_2+\ldots+w_mx_m+b)}) + \lambda \|w\|^2$$

Gaussian prior:

p(w,b) ~



Regularization/prior

L2 regularization:

$$\text{argmin}_{w,b} \sum_{i=1}^{n} \log(1+e^{-y_i(w_1x_2+w_2x_2+\ldots+w_mx_{m+b})}) + \lambda \|w\|^2$$

Gaussian prior:

$$\text{argmin}_{w,b} \sum_{i=1}^{n} \log(1+e^{-y_i(w_1x_2+w_2x_2+\ldots+w_mx_{m+b})}) + \frac{1}{2\sigma^2}\|w\|^2$$

Does the λ make sense? $\lambda = \frac{1}{2\sigma^2}$

Regularization/prior

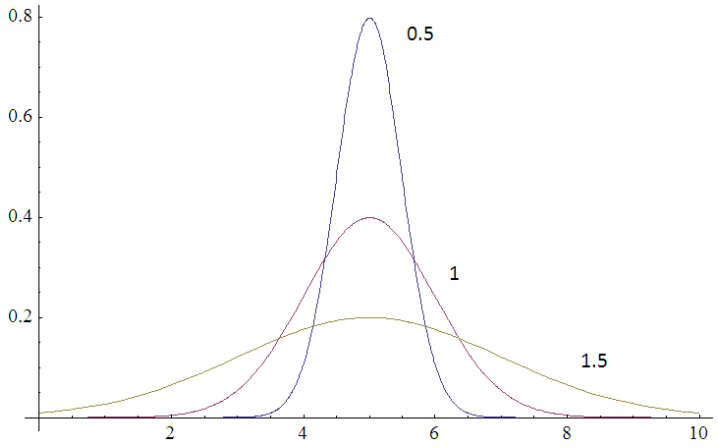
L2 regularization:

$$\text{argmin}_{w,b} \sum_{i=1}^{n} \log(1+e^{-y_i(w_1x_2+w_2x_2+\ldots+w_mx_{m+b})}) + \lambda \|w\|^2$$

Gaussian prior:

$$\text{argmin}_{w,b} \sum_{i=1}^{n} \log(1+e^{-y_i(w_1x_2+w_2x_2+\ldots+w_mx_{m+b})}) + \frac{1}{2\sigma^2}\|w\|^2$$

$$\lambda = \frac{1}{2\sigma^2}$$



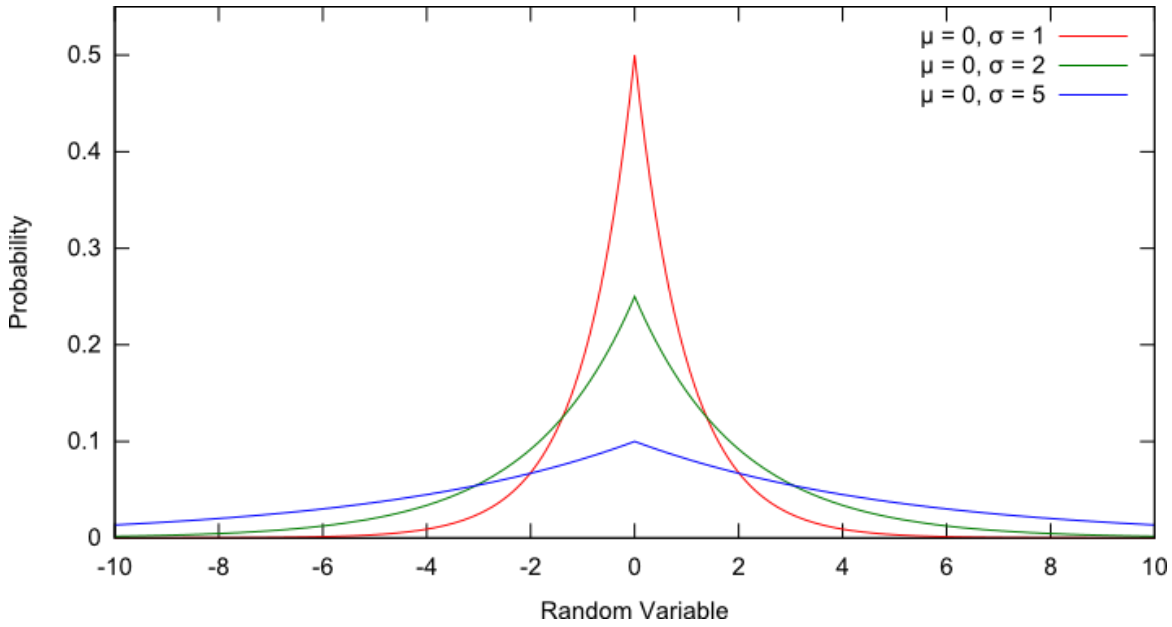
Regularization/prior

L1 regularization:

$$\text{argmin}_{w,b} \sum_{i=1}^{n} \log(1+e^{-y_i(w_1x_2+w_2x_2+\ldots+w_mx_m+b)}) + \lambda \|w\|$$

Laplacian prior:

p(w,b) ~



Regularization/prior

L1 regularization:

$$\text{argmin}_{w,b} \sum_{i=1}^{n} \log(1+e^{-y_i(w_1x_2+w_2x_2+\ldots+w_mx_m+b)}) + \lambda \|w\|$$

Laplacian prior:

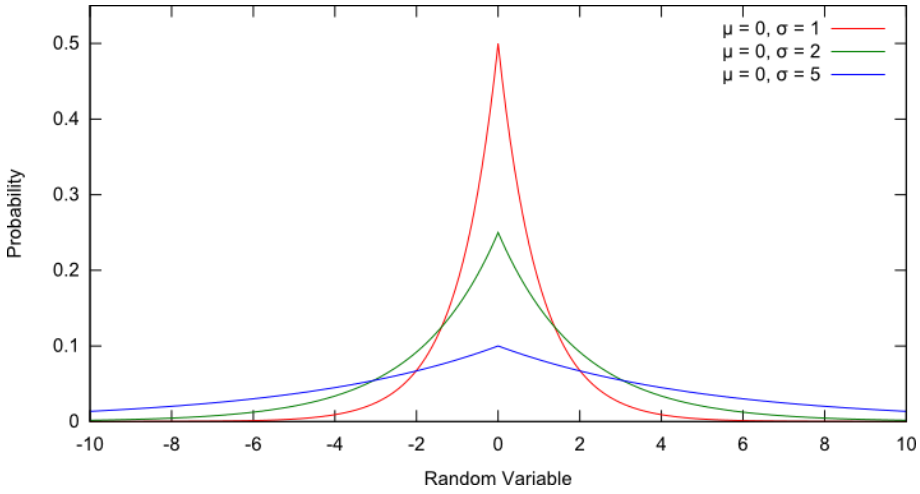
$$\text{argmin}_{w,b} \sum_{i=1}^{n} \log(1+e^{-y_i(w_1x_2+w_2x_2+\ldots+w_mx_m+b)}) + \frac{1}{s}\|w\|$$

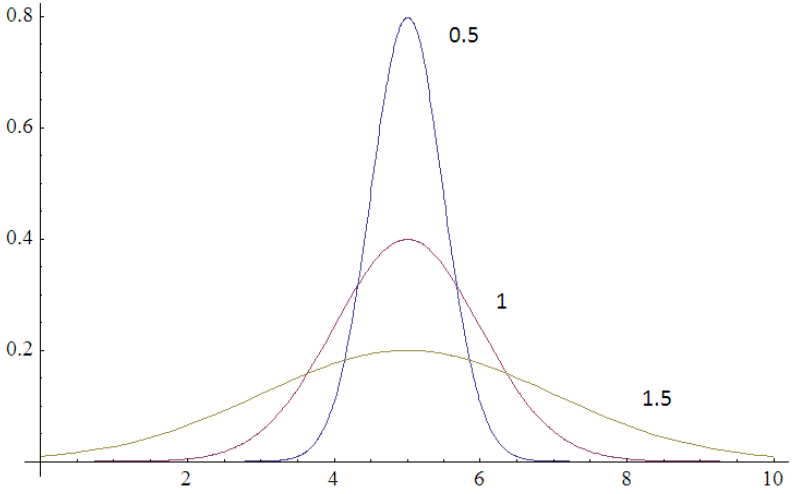
$$\lambda = \frac{1}{2s^2}$$

L1 vs. L2

L1 = Laplacian prior

L2 = Gaussian prior

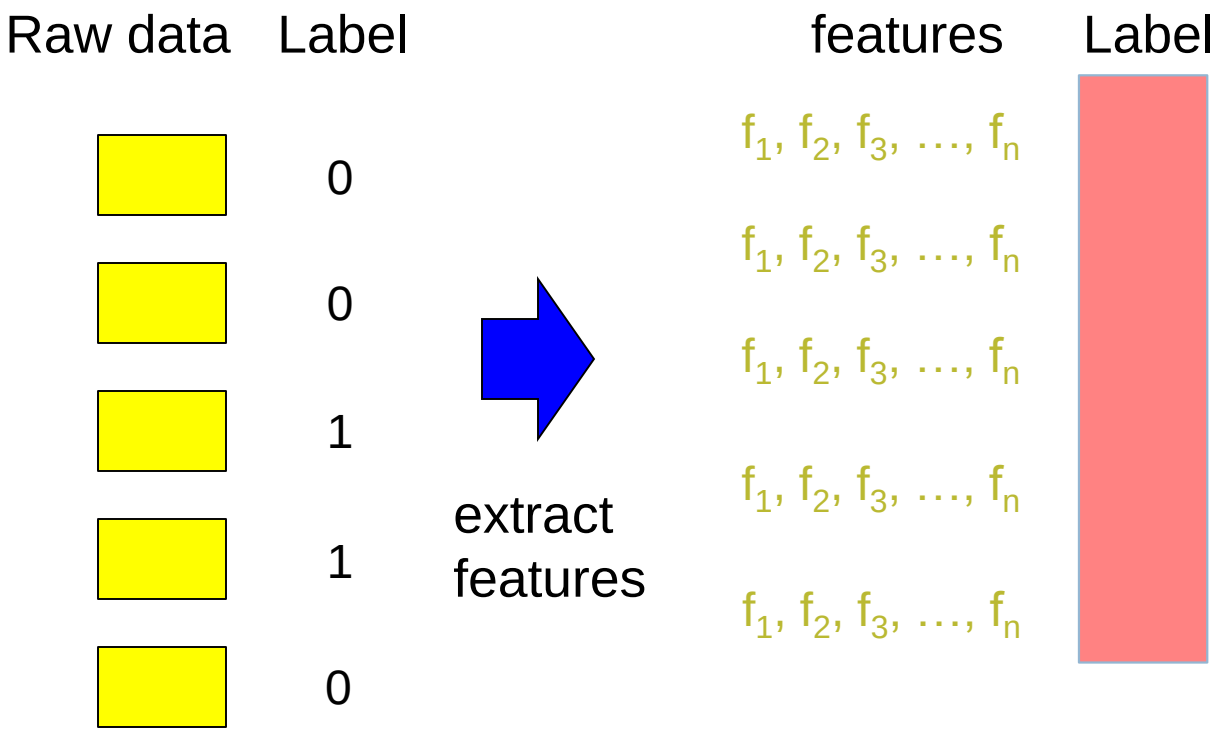




Logistic regression

Why is it called logistic regression?

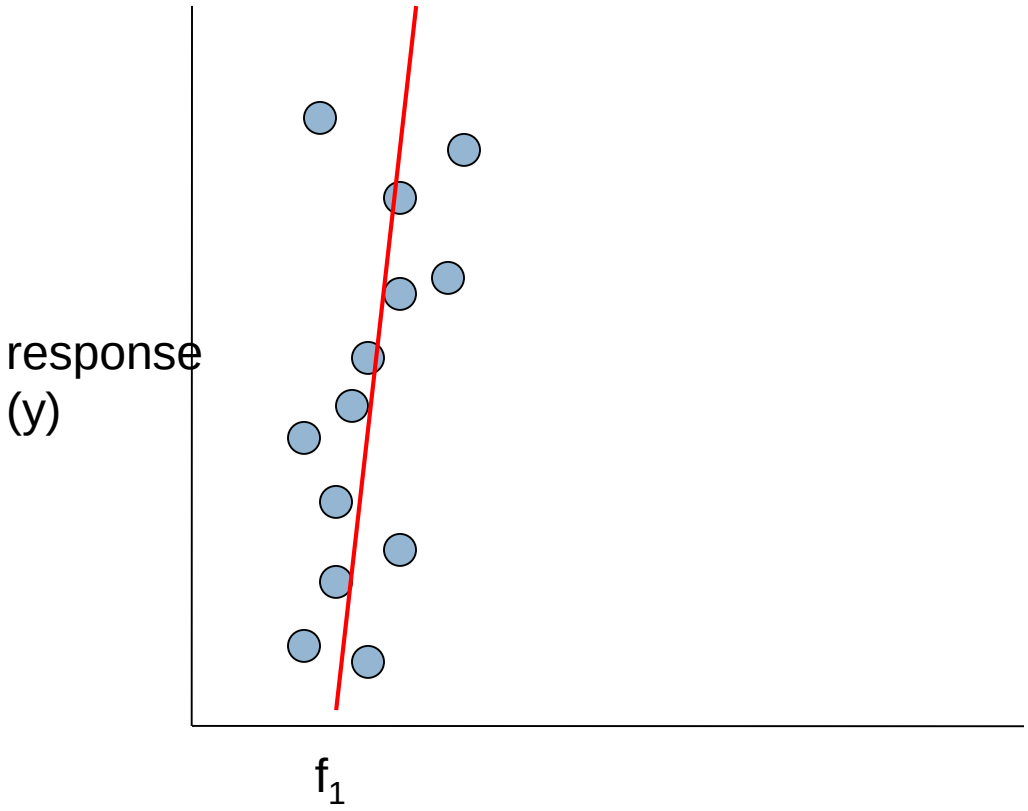
A digression: regression vs. classification



classification: discrete (some finite set of labels)

regression: real value

linear regression



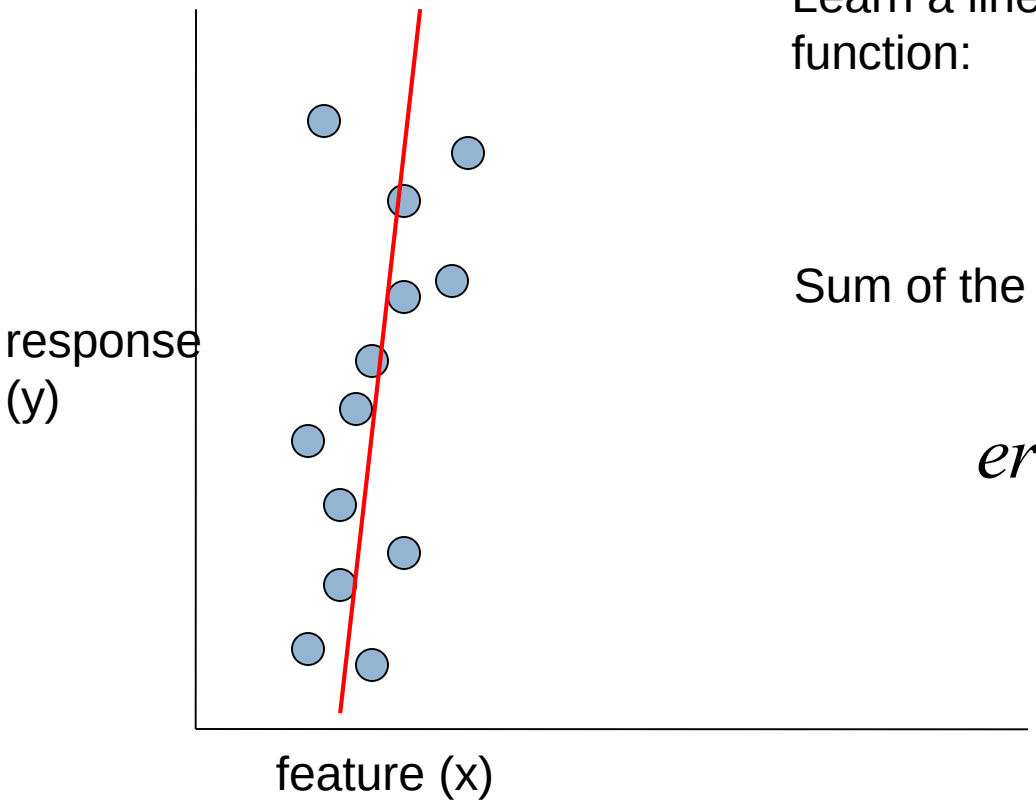
Given some points, find the ***line*** that best fits/explains the data

Our model is a line, i.e. we're assuming a linear relationship between the feature and the label value

$$h(y) = w_1 x_1 + b$$

How can we find this line?

Linear regression



Learn a line *h* that minimizes some loss/error function:

$$error(h) = ?$$

Sum of the individual errors:

$$error(h) = \sum_{i=1}^{n} |y_i - h(f_i)|$$

0/1 loss!

Error minimization

How do we find the minimum of an equation?

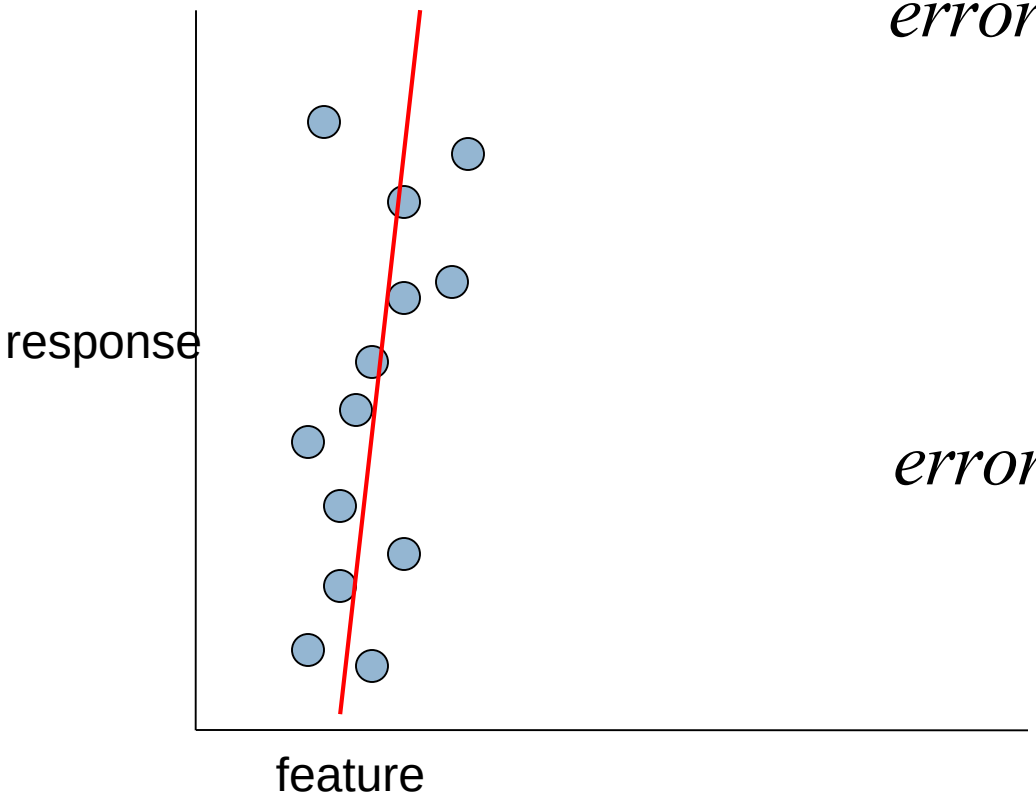
$$error(h) = \sum_{i=1}^{n} |y_i - h(f_i)|$$

Take the derivative, set to 0 and solve (going to be a min or a max)

Any problems here?

Ideas?

Linear regression

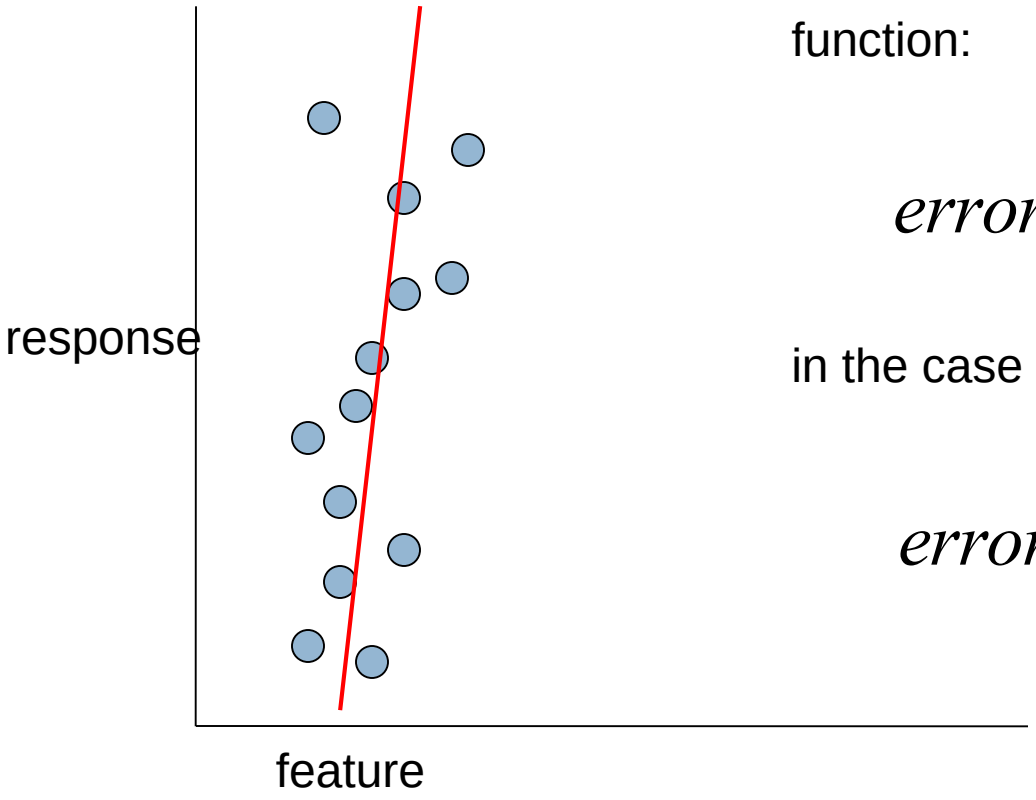


$$error(h) = \sum_{i=1}^{n} |y_i - h(f_i)|$$

$$error(h) = \sum_{i=1}^{n} (y_i - h(f_i))^2$$

squared error is convex!

Linear regression



Learn a line *h* that minimizes an error function:

$$error(h) = \sum_{i=1}^{n} (y_i - h(f_i))^2$$

in the case of a 2d line:

$$error(h) = \sum_{i=1}^{n} (y_i - (w_1 x_1 + w_0))^2$$

function for a line

Linear regression

We'd like to *minimize* the error

Find w_1 and w_0 such that the error is minimized

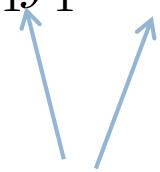
$$error(h) = \sum_{i=1}^{n} (y_i - (w_1 f_i + w_0))^2$$

We can solve this in closed form

Multiple linear regression

If we have m features, then we have a line in *m* dimensions

$$h(\bar{f}) = w_0 + w_1 f_1 + w_2 f_2 + \ldots + w_m f_m$$



weights

Multiple linear regression

We can still calculate the squared error like before

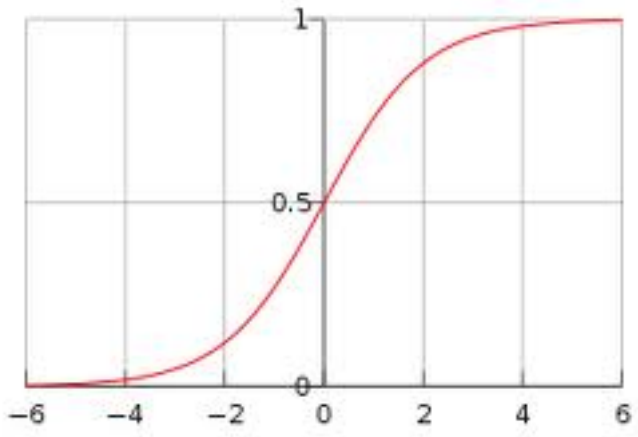
$$h(\bar{f}) = w_0 + w_1 f_1 + w_2 f_2 + \ldots + w_m f_m$$

$$error(h) = \sum_{i=1}^{n} (y_i - (w_0 + w_1 f_1 + w_2 f_2 + \ldots + w_m f_m))^2$$

Still can solve this exactly!

Logistic function

$$\text{logistic} = \frac{1}{1+e^{-x}}$$



Logistic regression

Find the best fit of the data based on a logistic