PROBABILISTIC MODELS

David Kauchak
CS158 – Fall 2013

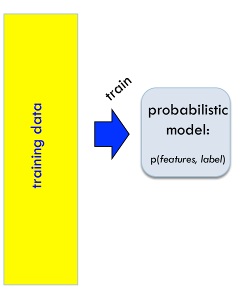
Admin

Assignment 6

Office hours today:

- 2:30-3:15
- 4-4:45

Probabilistic Modeling



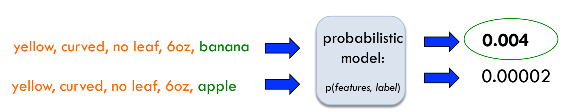
Model the data with a probabilistic model

specifically, learn p(*features, label*)

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

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

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?

Some maths

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

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

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

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

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

Step 1: pick a model

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

So, far we have made NO assumptions about the data

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

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

Full distribution tables

x_1	x_2	x_3	…	y	p( )
0	0	0	…	0	*
0	0	0	…	1	*
1	0	0	…	0	*
1	0	0	…	1	*
0	1	0	…	0	*
0	1	0	…	1	*
			…		

Wine problem:

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

2^{7000}

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 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 die
- the price of tea in England is independent of the whether or not you pass ML

independent or dependent?

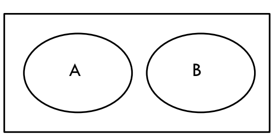
Catching a cold and having a cat-allergy

Miles per gallon and driving habits

Height and longevity of life

Independent variables

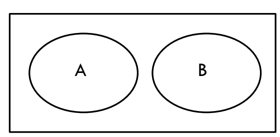
How does independence affect our probability equations/properties?



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 given 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) ≠ 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* (i.e. are conditionally independent given the label)

For the wine problem?

Naïve Bayes assumption

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

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

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

For example, the probability of "pinot" occurring is independent of whether or not "wine" occurs given that the review is about "chardonnay"

Is this assumption true?

Naïve Bayes assumption

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

For most applications, this is not true!

For example, the fact that "pinot" occurs will probably make it *more likely* that "noir" occurs (or take a compound phrase like "San Francisco")

However, this is often a reasonable approximation:

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

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:

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

biased coin toss!

Other features:

Could use a 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

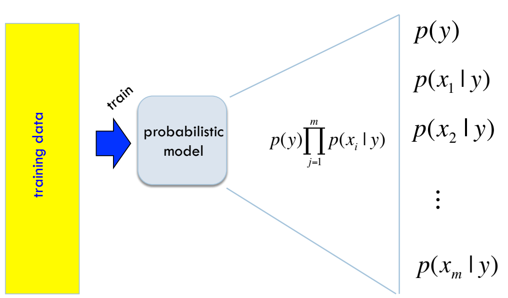
	Probabilistic models
Step 1: pick a model	Which model do we use, i.e. how do we calculate p(*feature, label*)?
Step 2: figure out how to estimate the probabilities for the model	How do train the model, i.e. how to we we estimate the probabilities for the model?
Step 3 (optional): deal with overfitting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

Obtaining probabilities

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

- How do we calculate $p(x_i \mid y)$ 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 pinot noir review?

We don't know!

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

$$\frac{\text{number of reviews labeled pinot noir}}{\text{total number of reviews}}$$

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 and 40 times you get tails.

What is the MLE estimate 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

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

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 ?

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

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 ?

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

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

60 heads with p(head) = 0.6

40 tails with p(tail) = 0.4

MLE example

Can we do any better?

$$likelihood(data) = \prod_i p(x_i)$$

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

60 heads with p(head) = 0.6

40 tails with p(tail) = 0.4

What about p(head) = 0.5?

MLE example

Can we do any better?

$$likelihood(data) = \prod_i p(x_i)$$

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

60 heads with p(head) = 0.6

40 tails with p(tail) = 0.4

$0.50^{60} * 0.50^{40} = 7.888609052210118e\text{-}31$

60 heads with p(head) = 0.5

40 tails with p(tail) = 0.5

MLE example

Can we do any better?

$$likelihood(data) = \prod_i p(x_i)$$

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

60 heads with p(head) = 0.6

40 tails with p(tail) = 0.4

What about p(head) = 0.7?

MLE example

Can we do any better?

$$likelihood(data) = \prod_i p(x_i)$$

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

60 heads with p(head) = 0.6

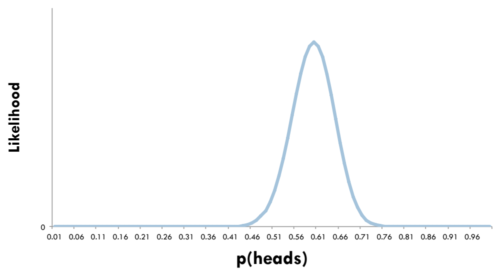
40 tails with p(tail) = 0.4

$0.70^{60} * 0.30^{40} = 6.176359828759916e\text{-}31$

60 heads with p(head) = 0.7

40 tails with p(tail) = 0.3

MLE Example



Maximum Likelihood Estimation (MLE)

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

$$MLE = \arg\max_\theta \prod_{i=1}^{n} p_\theta(x_i)$$

Often easier to work with log-likelihood:

$$MLE = \text{argmax}_\theta \log\left(\prod_{i=1}^{n} p_\theta(x_i)\right)$$

Why is this ok?

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

Calculating MLE

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

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

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.

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

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

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

$$MLE = \arg\max_\theta 60\log(\theta) + 40\log(1-\theta)$$

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.

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

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

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

$$40\theta = 60 - 60\theta$$

$$100\theta = 60$$

$$\theta = \frac{60}{100}$$

Yay!

Calculating MLE

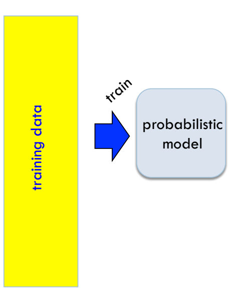
You flip a coin n times. a times you get heads and b times you get tails.

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

$$\ldots$$

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

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

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

$\dfrac{\text{number of examples with label}}{\text{total number of examples}}$

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

$\dfrac{\text{number of examples with the label with feature}}{\text{number of examples with label}}$

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

Naïve Bayes classification

yellow, curved, no leaf, 6oz, banana → NB Model p(features, label) → 0.004

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

Given an unlabeled example: yellow, curved, no leaf, 6oz predict the label

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

Probabilistic models

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

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

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

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

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

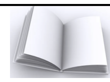
We also can use generative stories to help develop a model

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

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

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

Some maths

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

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

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

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

Some more maths

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

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

(because x_i are binary)

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

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

And…

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

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

What does this look like?

And…

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

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

b + $x_i * w_i$

$\boxed{w\ x + b}$

Linear model !!!

What are the weights?

NB as a linear model

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

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

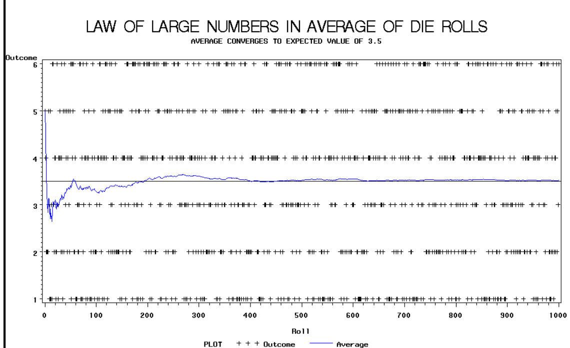
Probabilistic models

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?

Coin experiment



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

Then probability of us finding one is ~10%

- 90% of the time we won't find one and won't know anything (or assume p(parasitic) = 0)
- 10% of the time we would find one and incorrectly assume 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?