PROBABILISTIC MODELS

David Kauchak
CS451 – Fall 2013

Admin

Assignment 6

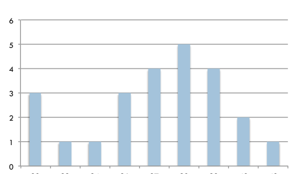
Assignment 7

CS Lunch on Thursday

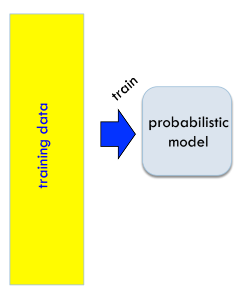
Midterm

Midterm

- mean: 37
- median: 38



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

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?

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

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*

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?

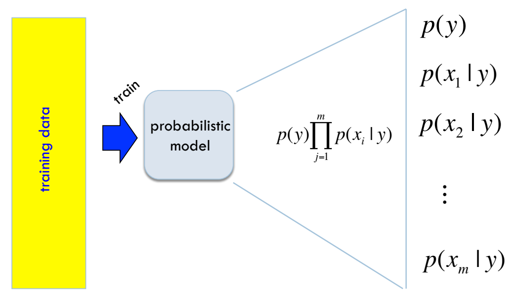
How do we deal with overfitting?

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?

NB: obtaining probabilities



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

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

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

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: sanity check

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

Can we do any better?

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

p(heads) = 0.6

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

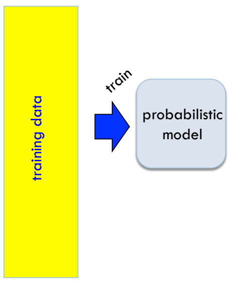
p(heads) = 0.5

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

p(heads) = 0.7

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

MLE estimation for NB



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

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

What are the MLE estimates for these?

Maximum likelihood estimates

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

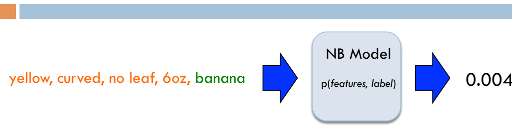
number of examples with label / total number of examples

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

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

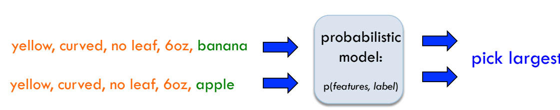


$$p(y)\prod_{j=1}^{m} p(x_i \mid y)$$

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

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

Probabilistic models



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

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

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

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

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

Some maths

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

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

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

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

Some more maths

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

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

(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

$\boxed{w\ x + b}$

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

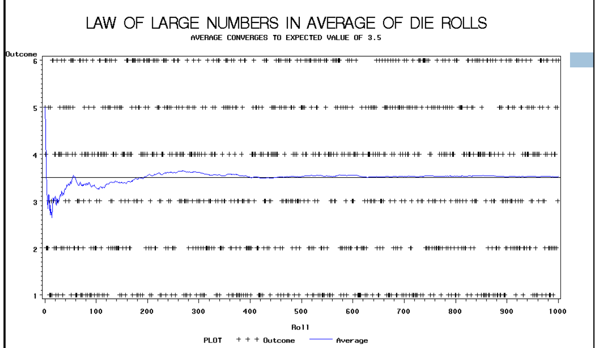
Problems?

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

Basic steps for probabilistic modeling

	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?

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(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?