

MACHINE LEARNING  
BASICS

David Kauchak  
CS159 Spring 2023

1

Admin

Assignment 6

No office hours on Friday

2

Zoom

3

Quiz #3

40 minutes

Open book and notes

Text Similarity (2/27) through Machine Translation (4/3)

Will continue class after (though I'll try and keep it brief)

4

### What do you want to see?

Large language models

ChatGPT

Generative models

“modern NLP”

NLP research

5

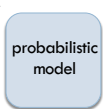
### Machine Learning is...

Machine learning is about predicting the future based on the past.  
-- Hal Daume III



6

### Probabilistic Modeling



Model the data with a probabilistic model

specifically, learn  $p(\text{features}, \text{label})$

$p(\text{features}, \text{label})$  tells us how likely these features and this example are

7

### An example: classifying fruit

Training data

examples

label

red, round, leaf, 3oz, ...

apple

green, round, no leaf, 4oz, ...

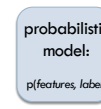
apple

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

banana

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

banana



8

## Probabilistic models

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



9

## 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

10

## Probabilistic models: big questions

1. Which model do we use, i.e. how do we calculate  $p(\text{feature}, \text{label})$ ?
2. How do train the model, i.e. how do we *estimate the probabilities* for the model?
3. How do we deal with overfitting (i.e. smoothing)?

11

## 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  $p(\text{feature}, \text{label})$ ?

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

How do we deal with overfitting?

12

### Some math

$$\begin{aligned}
 p(\text{features}, \text{label}) &= p(x_1, x_2, \dots, x_m, y) \\
 &= p(y) p(x_1, x_2, \dots, x_m | y) \\
 &= p(y) p(x_1 | y) p(x_2, \dots, x_m | y, x_1) \\
 &= p(y) p(x_1 | y) p(x_2 | y, x_1) p(x_3, \dots, x_m | y, x_1, x_2) \\
 &= p(y) \prod_{j=1}^m p(x_j | y, x_1, \dots, x_{j-1})
 \end{aligned}$$

13

### Step 1: pick a model

$$p(\text{features}, \text{label}) = p(y) \prod_{j=1}^m p(x_j | y, x_1, \dots, x_{j-1})$$

So, far we have made NO assumptions about the data

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

How many entries would the probability distribution table have if we tried to represent all possible values and we had 7000 binary features?

14

### Full distribution tables

$x_1$	$x_2$	$x_3$	...	$y$	$p(\cdot)$
0	0	0	...	0	*
0	0	0	...	1	*
1	0	0	...	0	*
1	0	0	...	1	*
0	1	0	...	0	*
0	1	0	...	1	*
			...		

All possible combination of features!

Table size:  $2^{7000} = ?$

15

2<sup>7000</sup>

```

1621696755662202026466665085478377095191112430363743256235982084151527023162702352987080237879
4460004651996019099530984538652557892546513204107022110253564658647431585227076599373340842842
724200122818782600729310826170431944842663920778412509996860169436066600112098175792966787
819625237700653294757256478055809293844652718640216108862600816097132874749204352087401101862
690842327501724665231129395523305905454421455477250950906507889478094683592939574112569473438
6191215296848474344406741204174020887540371869421701550220735398381224299258743537536161041593
4359455766656170179090417259702533652662682021808493892812699709528570890637557541434487608
82483699419938024151975145101251270438290872809195384763028578118540240995889964192277601255
360491156240349994714416090573084242931396211953679373012944795600248333570738998392029910322
346598038953069042980174009801732521069130797124201696339723021835300758978451952584853710885
8195631737000743805167411189134617501484521767984296782842287373127422122022517597535994839257
02987790706353347902449354353866605125910795672914312162977887848185522928196541766009803989
979916814047493842157435158026038115106828460678973048382922034604277576550737656754730702714
46622348768570962126107476270520304948890720897859369047063428548331668656573271744606581835
60906484958001276175461457216176955575199211750751406775104496728590822558547771447242334900
76402632176089211355256124119453870268029904001838585057671936968975936612135688883680023840
932567380777501891470304962150996983853975207154939633923720287592041517294937079077853625108
3200928396048072795488706954662168804465211249307629009199071774235503913511744153297374799300
8995583051888113334798464113680049994037374546003528811232632821866113104550728992296946
915601858083982074170460683212438815202609584696588161375826382921029547343888832163627122302
921229795384868355483537106034077891774170263636562027269554375177807413134551018100094688094
0781122057380335371124632958916237089580476224595091825301636909236240671411644331656159828058
3720783439888562390892028440902553829376
    
```

Any problems with this?

16

## Full distribution tables

$x_1$	$x_2$	$x_3$	...	$y$	$p(\cdot)$
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?

17

## Step 1: pick a model

$$p(\text{features}, \text{label}) = p(y) \prod_{j=1}^m p(x_j | y, x_1, \dots, x_{j-1})$$

So, far we have made NO assumptions about the data

Model selection involves making assumptions about the data

We've done this before, n-gram language model, parsing, etc.

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

18

## Naïve Bayes assumption

$$p(\text{features}, \text{label}) = p(y) \prod_{j=1}^m p(x_j | y, x_1, \dots, x_{j-1})$$

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

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

19

## Naïve Bayes model

$$\begin{aligned} p(\text{features}, \text{label}) &= p(y) \prod_{j=1}^m p(x_j | y, x_1, \dots, x_{j-1}) \\ &= p(y) \prod_{j=1}^m p(x_j | y) \quad \text{naïve Bayes assumption} \end{aligned}$$

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

How do we model this?

- for binary features (e.g., "banana" occurs in the text)
- for discrete features (e.g., "banana" occurs  $x_i$  times)
- for real valued features (e.g, the text contains  $x_i$  proportion of verbs)

20

### p(x | y)

---

Binary features (aka, Bernoulli Naïve Bayes) :

$$p(x_j | y) = \begin{cases} \theta_j & \text{if } x_j = 1 \\ 1 - \theta_j & \text{otherwise} \end{cases} \quad \text{biased coin toss!}$$

21

### Basic steps for probabilistic modeling

---

<p>Step 1: pick a model</p> <div style="border: 1px solid #4F81BD; padding: 5px; margin: 5px 0;"> <p>Step 2: figure out how to estimate the probabilities for the model</p> </div> <p>Step 3 (optional): deal with overfitting</p>	<p><b>Probabilistic models</b></p> <p>Which model do we use, i.e. how do we calculate p(feature, label)?</p> <p>How do we train the model, i.e. how do we estimate the probabilities for the model?</p> <p>How do we deal with overfitting?</p>
------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

22

### Obtaining probabilities

---

The diagram shows a yellow vertical bar labeled "training data" with a blue arrow labeled "train" pointing to a rounded rectangle labeled "probabilistic model". From the right side of the model, a large blue triangle expands to the right, containing the following text:

- $p(y)$
- $p(x_1 | y)$
- $p(x_2 | y)$
- $\vdots$
- $p(x_m | y)$

Below the triangle, the joint probability is given as  $p(y) \prod_{j=1}^m p(x_j | y)$ . At the bottom right, a note says "(m = number of features)".

23

### MLE estimation for Bernoulli NB

---

The diagram shows a yellow vertical bar labeled "training data" with a blue arrow labeled "train" pointing to a rounded rectangle labeled "probabilistic model". From the right side of the model, a large blue triangle expands to the right, containing the following text:

- $p(y) \prod_{i=1}^m p(x_i | y)$
- $p(y)$
- $p(x_j | y)$

Below the triangle, a red text prompt asks: "What are the MLE estimates for these?"

24

### Maximum likelihood estimates

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

number of examples with label  
total number of examples

$$p(x_j | y) = \frac{\text{count}(x_j, y)}{\text{count}(y)}$$

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?

25

### Text classification

$$p(y) = \frac{\text{count}(y)}{n}$$
$$p(w_j | y) = \frac{\text{count}(w_j, y)}{\text{count}(y)}$$

**Unigram features:**  
 $w_i$ , whether or not word  $w_i$  occurs in the text

What are these counts for text classification with unigram features?

26

### Text classification

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

number of texts with label  
total number of texts

$$p(w_j | y) = \frac{\text{count}(w_j, y)}{\text{count}(y)}$$

number of texts with the label with word  $w_j$   
number of texts with label

27

### Naïve Bayes classification

yellow, curved, no leaf, 6oz, banana

NB Model  
 $p(\text{features}, \text{label})$

➔

0.004

$$p(y) \prod_{j=1}^m p(x_j | y)$$


---

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

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

28

### NB classification

probabilistic model:  $p(\text{features}, \text{label})$

yellow, curved, no leaf, 6oz, banana  $\rightarrow p(y=1) \prod_{j=1}^m p(x_j | y=1)$   $\rightarrow$  pick largest

yellow, curved, no leaf, 6oz, apple  $\rightarrow p(y=2) \prod_{j=1}^m p(x_j | y=2)$   $\rightarrow$  pick largest

---

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

29

### NB classification

probabilistic model:  $p(\text{features}, \text{label})$

yellow, curved, no leaf, 6oz, banana  $\rightarrow p(y=1) \prod_{j=1}^m p(x_j | y=1)$   $\rightarrow$  pick largest


yellow, curved, no leaf, 6oz, apple  $\rightarrow p(y=2) \prod_{j=1}^m p(x_j | y=2)$   $\rightarrow$  pick largest

Notice that each label has its own separate set of parameters, i.e.  $p(x_j | y)$

30

### Bernoulli NB for text classification

probabilistic model:  $p(\text{features}, \text{label})$

  $(1, 1, 1, 0, 0, 1, 0, 0, \dots)$   
 $\hat{x}_1 \hat{x}_2 \hat{x}_3 \hat{x}_4 \hat{x}_5 \hat{x}_6 \hat{x}_7 \hat{x}_8$

$\rightarrow p(y=1) \prod_{j=1}^m p(w_j | y=1)$   $\rightarrow$  pick largest


$\rightarrow p(y=2) \prod_{j=1}^m p(w_j | y=2)$   $\rightarrow$  pick largest

---

How good is this model for text classification?

31

### Bernoulli NB for text classification

  $(1, 1, 1, 0, 0, 1, 0, 0, \dots)$   
 $w_1 w_2 w_3 w_4 w_5 w_6 w_7 w_8$

$\rightarrow p(y=1) \prod_{j=1}^m p(w_j | y=1)$   $\rightarrow$  pick largest

$\rightarrow p(y=2) \prod_{j=1}^m p(w_j | y=2)$   $\rightarrow$  pick largest

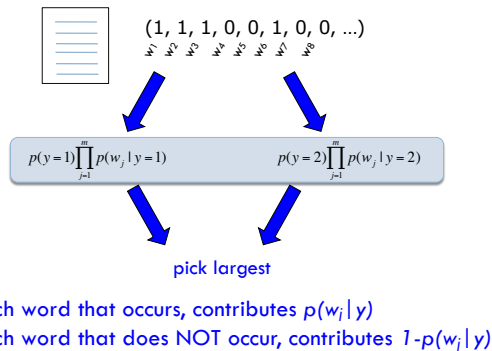
---

For text classification, what is this computation?  
Does it make sense?

32



## Bernoulli NB for text classification



33

## 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** an example

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

34

## Bernoulli NB generative story



$$p(y) \prod_{j=1}^m p(x_j | y)$$

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 does this mean for text classification, assuming unigram features?

35

## Bernoulli NB generative story



$$p(y) \prod_{j=1}^m p(w_j | y)$$

1. Pick a label according to  $p(y)$ 
  - roll a biased, num\_labels-sided die
2. For each word in your vocabulary:
  - Flip a *biased* coin:
    - if heads, include the word in the text
    - if tails, don't include the word

36

## Bernoulli NB

$$p(y) \prod_{j=1}^m p(x_j | y)$$

Pros/cons?

37

## Bernoulli NB

### Pros

- Easy to implement
- Fast!
- Can be done on large data sets

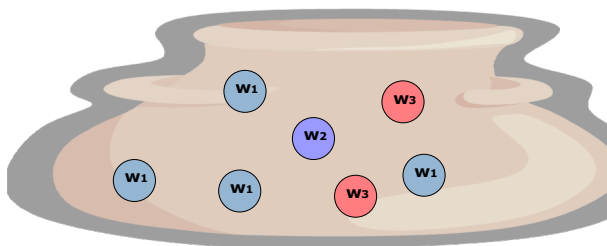
### Cons

- Naïve Bayes assumption is generally not true
- Performance isn't as good as other models
- For text classification (and other sparse feature domains) the  $p(x_i=0 | y)$  can be problematic

38


## Another generative story

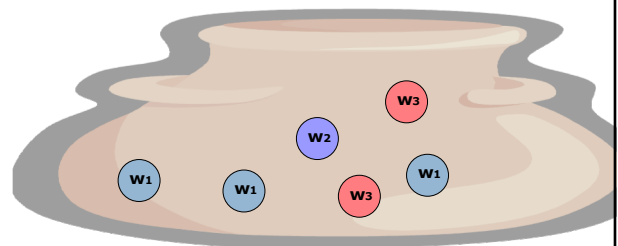
Randomly draw words from a "bag of words" until document length is reached



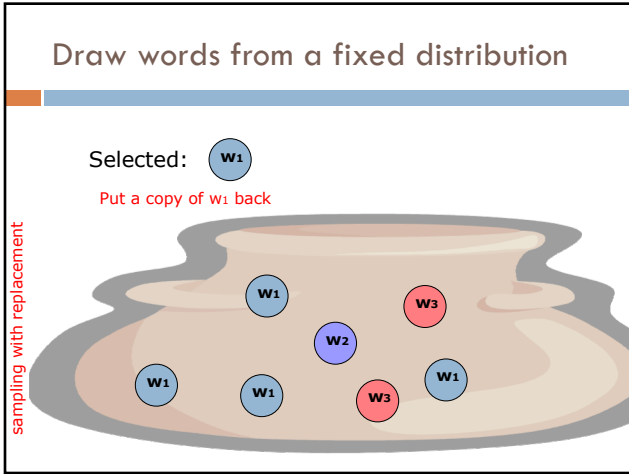
39

## Draw words from a fixed distribution

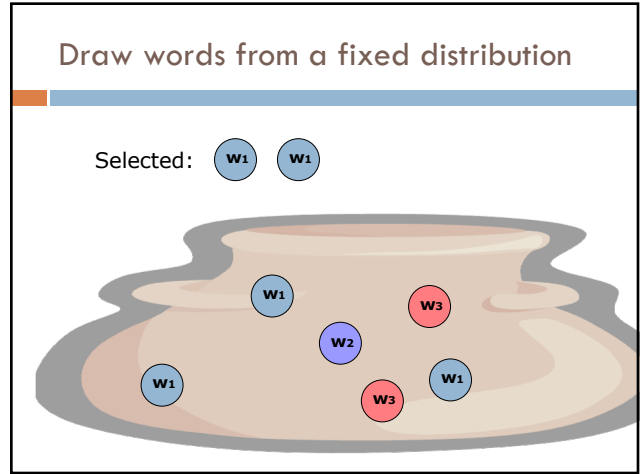
Selected: 



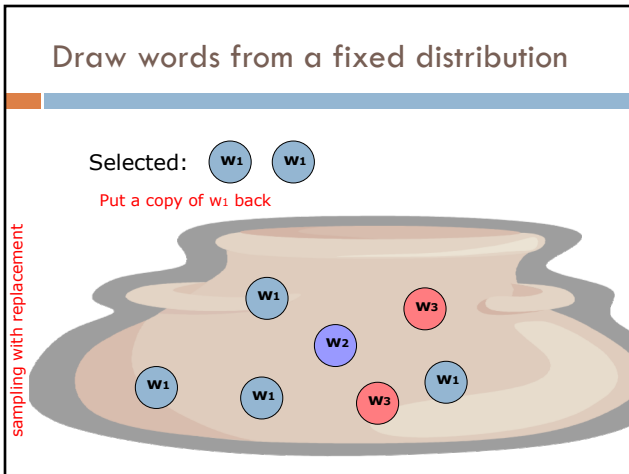
40



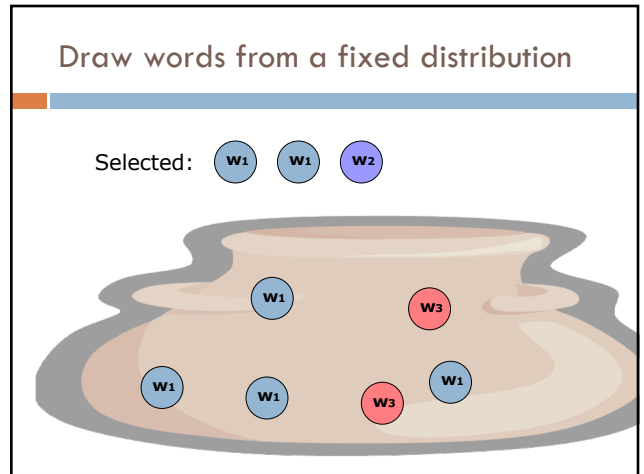
41



42



43



44

Draw words from a fixed distribution

Selected:  $w_1$   $w_1$   $w_2$

Put a copy of  $w_2$  back

sampling with replacement

45

Draw words from a fixed distribution

Selected:  $w_1$   $w_1$   $w_2$  ...

46

Draw words from a fixed distribution

Is this a NB model, i.e. does it assume each individual word occurrence is independent?

47

Draw words from a fixed distribution

Yes! Doesn't matter what words were drawn previously, still the same probability of getting any particular word

48

### Draw words from a fixed distribution

Does this model handle multiple word occurrences?

49

### Draw words from a fixed distribution

Selected:  $w_1$   $w_1$   $w_2$  ...

50

### NB generative story

#### Bernoulli NB

- Pick a label according to  $p(y)$ 
  - roll a biased, num\_labels-sided die
- For each word in your vocabulary:
  - Flip a biased coin:
    - if heads, include the word in the text
    - if tails, don't include the word

#### Multinomial NB

- Pick a label according to  $p(y)$ 
  - roll a biased, num\_labels-sided die
- Keep drawing words from  $p(\text{words} | y)$  until text length has been reached.

51

### Probabilities

#### Bernoulli NB

- Pick a label according to  $p(y)$ 
  - roll a biased, num\_labels-sided die
- For each word in your vocabulary:
  - Flip a biased coin:
    - if heads, include the word in the text
    - if tails, don't include the word

$$p(y) \prod_{j=1}^m p(x_j | y)$$

(1, 1, 1, 0, 0, 1, 0, 0, ...)

#### Multinomial NB


- Pick a label according to  $p(y)$ 
  - roll a biased, num\_labels-sided die
- Keep drawing words from  $p(\text{words} | y)$  until document length has been reached

?

(4, 1, 2, 0, 0, 7, 0, 0, ...)

52

A digression: rolling dice




What's the probability of getting a 3 for a single roll of this dice?

$1/6$

53

A digression: rolling dice




What is the probability distribution over possible single rolls?

$1/6$	$1/6$	$1/6$	$1/6$	$1/6$	$1/6$
1	2	3	4	5	6

54


A digression: rolling dice



What if I told you 1 was twice as likely as the others?

$2/7$	$1/7$	$1/7$	$1/7$	$1/7$	$1/7$
1	2	3	4	5	6

55

A digression: rolling dice 

What if I rolled 400 times and got the following number?

1: 100  
 2: 50  
 3: 50  
 4: 100  
 5: 50  
 6: 50

$1/4$	$1/8$	$1/8$	$1/4$	$1/8$	$1/8$
1	2	3	4	5	6

56

### A digression: rolling dice

1. What is the probability of rolling a 1 and a 5 (in any order)?
2. Two 1s and a 5 (in any order)?
3. Five 1s and two 5s (in any order)?

1/4	1/8	1/8	1/4	1/8	1/8
1	2	3	4	5	6

57

### A digression: rolling dice

1. What is the probability of rolling a 1 and a 5 (in any order)?  
 $(1/4 * 1/8) * 2 = 1/16$   
prob. of those two rolls      number of ways that can happen (1,5 and 5,1)
1. Two 1s and a 5 (in any order)?  
 $((1/4)^2 * 1/8) * 3 = 3/128$
2. Five 1s and two 5s (in any order)?  
 $((1/4)^5 * (1/8)^2) * 21 = 21/524,288 = 0.00004$       **General formula?**

1/4	1/8	1/8	1/4	1/8	1/8
1	2	3	4	5	6

58

### Multinomial distribution

Multinomial distribution: independent draws over  $m$  possible categories

If we have frequency counts  $x_1, x_2, \dots, x_m$  over each of the categories, the probability is:

$$p(x_1, x_2, \dots, x_m | \theta_1, \theta_2, \dots, \theta_m) = \frac{n!}{\prod_{j=1}^m x_j!} \prod_{j=1}^m \theta_j^{x_j}$$

number of different ways to get those counts
probability of particular counts

$\theta_1$	$\theta_2$	$\theta_3$	$\theta_4$	$\theta_5$	$\theta_6$	...
1	2	3	4	5	6	...

59

### Multinomial distribution

$$p(x_1, x_2, \dots, x_m | \theta_1, \theta_2, \dots, \theta_m) = \frac{n!}{\prod_{j=1}^m x_j!} \prod_{j=1}^m \theta_j^{x_j}$$

What are  $\theta_j$ ?

Are there any constraints on the values that they can take?

$\theta_1$	$\theta_2$	$\theta_3$	$\theta_4$	$\theta_5$	$\theta_6$	...
1	2	3	4	5	6	...

60

### Multinomial distribution

$$p(x_1, x_2, \dots, x_m | \theta_1, \theta_2, \dots, \theta_m) = \frac{n!}{\prod_{j=1}^m x_j!} \prod_{j=1}^m \theta_j^{x_j}$$

$\theta_j$ : probability of rolling "j"

$$\theta_j \geq 0$$

$$\sum_{j=1}^m \theta_j = 1$$

$\theta_1$	$\theta_2$	$\theta_3$	$\theta_4$	$\theta_5$	$\theta_6$	...
1	2	3	4	5	6	

61

### Back to words...

Why the digression?

$$p(x_1, x_2, \dots, x_m | \theta_1, \theta_2, \dots, \theta_m) = \frac{n!}{\prod_{j=1}^m x_j!} \prod_{j=1}^m \theta_j^{x_j}$$

Drawing words from a bag is the same as rolling a die!

number of sides = number of words in the vocabulary

62

### Back to words...

Why the digression?

$$p(x_1, x_2, \dots, x_m | \theta_1, \theta_2, \dots, \theta_m) = \frac{n!}{\prod_{j=1}^m x_j!} \prod_{j=1}^m \theta_j^{x_j}$$

$$p(\text{features}, \text{label}) = p(y) \frac{n!}{\prod_{j=1}^m x_j!} \prod_{j=1}^m (\theta_y)^{x_j}$$

$\theta_y$  for class y


63

### Basic steps for probabilistic modeling

Model each class as a multinomial:

$$p(\text{features}, \text{label}) = p(y) \frac{n!}{\prod_{j=1}^m x_j!} \prod_{j=1}^m (\theta_y)^{x_j}$$

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



How do we train the model, i.e. estimate  $\theta_j$  for each class?

64



### A digression: rolling dice





What if I rolled 400 times and got the following number?



1: 100  
2: 50  
3: 50  
4: 100  
5: 50  
6: 50

1/4	1/8	1/8	1/4	1/8	1/8
1	2	3	4	5	6

65

### Training a multinomial





label1    

label2  

1/4	1/8	1/8	1/4	1/8	1/8
1	2	3	4	5	6

66

### Training a multinomial

label1    

For each label, y:

w1: 100 times  
w2: 50 times  
w3: 10 times  
w4: ...


$$\theta_j = \frac{\text{count}(w_j, y)}{\sum_{k=1}^m \text{count}(w_k, y)}$$

=  $\frac{\text{number of times word } w_j \text{ occurs in label } y \text{ docs}}{\text{total number of words in label } y \text{ docs}}$

1/4	1/8	1/8	1/4	1/8	1/8
1	2	3	4	5	6

67

### Classifying with a multinomial

 (10, 2, 6, 0, 0, 1, 0, 0, ...)

$w_1$   $w_2$   $w_3$   $w_4$   $w_5$   $w_6$   $w_7$   $w_8$

$p(y=1) = \frac{n!}{\prod_{j=1}^m x_j!} \prod_{j=1}^m (\theta_j)^{x_j}$        $p(y=2) = \frac{n!}{\prod_{j=1}^m x_j!} \prod_{j=1}^m (\theta_j)^{x_j}$

Any way I can make this simpler?

pick largest

68

### Classifying with a multinomial

(10, 2, 6, 0, 0, 1, 0, 0, ...)

$w_1, w_2, w_3, w_4, w_5, w_6, w_7, w_8, \dots$

$p(y=1) \prod_{j=1}^m (\theta_j)^{x_j^y}$        $p(y=2) \prod_{j=1}^m (\theta_j)^{x_j^y}$

$\frac{n!}{\prod_{j=1}^m x_j!}$  is a constant!

pick largest

69

### Multinomial finalized

**Training:**

- Calculate  $p(\text{label})$
- For each label, calculate  $\theta$ s

$$\theta_j = \frac{\text{count}(w_j, y)}{\sum_{k=1}^m \text{count}(w_k, y)}$$

**Classification:**

- Get word counts
- For each label you had in training, calculate:

$$p(y) \prod_{j=1}^m \theta_j^{x_j^y}$$

and pick the largest

70

### Multinomial vs. Bernoulli?

Handles word frequency

Given enough data, tends to performs better

Yahoo Science

<http://www.cs.cmu.edu/~kriyam/papers/multinomial-aaaiw98.pdf>

71

### Multinomial vs. Bernoulli?

Handles word frequency

Given enough data, tends to performs better

Newsgroups

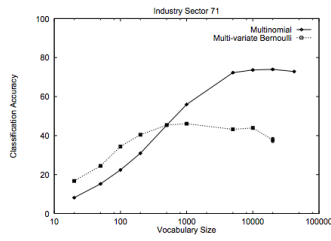
<http://www.cs.cmu.edu/~kriyam/papers/multinomial-aaaiw98.pdf>

72

## Multinomial vs. Bernoulli?

Handles word frequency

Given enough data, tends to perform better



<http://www.cs.cmu.edu/~lsgiam/papers/multinomial-coast-98.pdf>

73

## 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

74

## 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

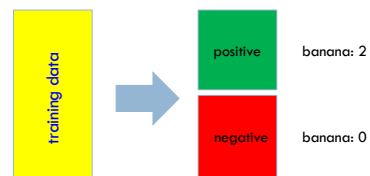
Which model do we use, i.e. how do we calculate  $p(\text{feature}, \text{label})$ ?

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

How do we deal with overfitting?

75

## Unseen events



$$\theta_j = \frac{\text{count}(w_j, y)}{\sum_{k=1}^m \text{count}(w_k, y)}$$

What will  $\theta_{\text{banana}}$  be for the negative class?

76

### Unseen events

training data → positive banana: 2  
negative banana: 0

$$\theta_j = \frac{\text{count}(w_j, y)}{\sum_{k=1}^m \text{count}(w_k, y)}$$

What will  $\theta_{\text{banana}}$  be for the negative class?  
O! Is this a problem?

77

### Unseen events

training data → positive banana: 2  
negative banana: 0

$p(\text{"I ate a bad banana"}, \text{negative}) = ?$

78

### Unseen events

training data → positive banana: 2  
negative banana: 0

$p(\text{"I ate a bad banana"}, \text{negative}) = 0$   
 $p(\text{"... banana ..."}, \text{negative}) = 0$

Solution?

79

### Add lambda smoothing

training data → positive banana: 2  
negative banana: 0

$$\theta_j = \frac{\text{count}(w_j, y)}{\sum_{k=1}^m \text{count}(w_k, y)}$$

$$\theta_j = \frac{\text{count}(w_j, y) + \lambda}{\lambda m + \sum_{k=1}^m \text{count}(w_k, y)}$$

for each label, pretend like we've seen each feature/word occur in  $\lambda$  additional examples

80

### Different than...

training data

positive banana: 0

negative banana: 0

How is this problem different?

81

### Different than...

training data

positive banana: 0

negative banana: 0

$p(\text{"I ate a bad banana"}, \text{positive})$   $\rightarrow$   $p(\text{"I ate a bad"}, \text{positive})$

$p(\text{"I ate a bad banana"}, \text{negative})$   $\rightarrow$   $p(\text{"I ate a bad"}, \text{negative})$

Out of vocabulary. Many ways to solve... for our implementation, we'll just ignore them.

82

### 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?

83

### Training revisited

From a probability standpoint, MLE training is selecting the  $\Theta$  that maximizes:

$$p(\theta | \text{data})$$

i.e.

$$\text{argmax}_{\theta} p(\theta | \text{data})$$

We pick the most likely model parameters given the data

84

## Estimating revisited

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

To do this, we need to break down our probability

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

(Hint: Bayes rule)

85

## Estimating revisited

What are each of these probabilities?

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

86

## 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)

87

## Priors

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

Does  $p(data)$  matter for the  $\operatorname{argmax}$ ?

88

## Priors

likelihood of the data  
under the model

probability of different parameters,  
call the **prior**

$$\theta = \operatorname{argmax}_{\theta} p(\text{data} | \theta)p(\theta)$$

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

89

## Priors

likelihood of the data  
under the model

probability of different parameters,  
call the **prior**

$$\theta = \operatorname{argmax}_{\theta} p(\text{data} | \theta)p(\theta)$$

- Assumes a **uniform prior**, i.e. all  $\Theta$  are equally likely!
- Relies solely on the **likelihood**

90

## A better approach

$$\theta = \operatorname{argmax}_{\theta} p(\text{data} | \theta)p(\theta)$$

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

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

91

## Another view on the prior

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

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

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

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

92

## What about smoothing?

training data

$$\theta_j = \frac{\text{count}(w_j, y)}{\sum_{k=1}^m \text{count}(w_k, y)}$$



$$\theta_j = \frac{\text{count}(w_j, y) + \lambda}{\lambda m + \sum_{k=1}^m \text{count}(w_k, y)}$$

for each label, pretend like we've seen each feature/word occur in additional examples

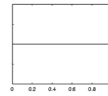
Sometimes this is also called **smoothing** because it is seen as smoothing or interpolating between the MLE and some other distribution

93

## Prior for NB

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

Uniform prior



Dirichlet prior



$\lambda = 0$  → increasing

$$p(w_j | y) = \theta_j = \frac{\text{count}(w_j, y)}{\sum_{k=1}^m \text{count}(w_k, y)}$$

$$\theta_j = \frac{\text{count}(w_j, y) + \lambda}{\sum_{k=1}^m (\text{count}(w_k, y) + \lambda)} = \frac{\text{count}(w_j, y) + \lambda}{\lambda m + \sum_{k=1}^m \text{count}(w_k, y)}$$

94