Probabilistic Modeling

Model the data with a probabilistic model specifically, learn \( p(\text{features, label}) \)

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

Probabilistic models

Probabilistic models define a probability distribution over features and labels:

- yellow, curved, no leaf, 6oz, banana
- yellow, curved, no leaf, 6oz, apple

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

Office hours today:
- 2:30-3:15
- 4:45-5:15
Basic steps for probabilistic modeling

<table>
<thead>
<tr>
<th>Step 1: pick a model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Step 2: figure out how to estimate the probabilities for the model</td>
</tr>
<tr>
<td>Step 3 (optional): deal with overfitting</td>
</tr>
</tbody>
</table>

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 estimate the probabilities for the model?

How do we deal with overfitting?

Probabilistic models

Step 1: pick a model

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

Step 3 (optional): deal with overfitting

Some maths

$p(features, label) = p(x_1, x_2, ..., x_n, y) = p(y)p(x_1, x_2, ..., x_n \mid y) = p(y)p(x_1 \mid y)p(x_2, ..., x_n \mid y, x_1) = p(y)p(x_1 \mid y)p(x_2 \mid y, x_1)p(x_3, ..., x_n \mid y, x_2) = p(y)\prod_{j=1}^{n} p(x_j \mid y, x_1, ..., x_{j-1})$
Wine problem:

- all possible combination of features
- ~7000 binary features

Sample space size: $2^{7000} = \infty$

Any problems with this?

Step 1: pick a model

$$p(\text{features},\text{label}) = p(y) \prod_{i \geq 1} p(x_i | y, x_1, ..., x_{i-1})$$

So, far we have made NO assumptions about the data

Model selection involves making assumptions about the data

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

These assumptions allow us to represent the data more compactly and to estimate the parameters of the model
An aside: independence

Two variables are independent if one has nothing 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 whether or not you pass ML.

Independent variables

How does independence affect our probability equations/properties?

If A and B are independent (written \( \text{...} \)):
- \( P(A, B) = ? \)
- \( P(A | B) = ? \)
- \( P(B | A) = ? \)

Independent or dependent?

Catching a cold and having a cat-allergy

Miles per gallon and driving habits

Height and longevity of life

How does independence help us?

If A and B are independent (written \( \text{...} \)):
- \( P(A, B) = P(A)P(B) \)
- \( P(A | B) = P(A) \)
- \( P(B | A) = P(B) \)
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 \( P(A,B) \neq P(A)P(B) \)

Naïve Bayes assumption

\[
p(\text{features, label}) = p(y) \prod_{j=1}^{m} p(x_j | y, x_1, ..., x_{i-1}) \]

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

What does this assume?

Naïve Bayes assumption

\[
p(\text{features, label}) = p(y) \prod_{j=1}^{m} p(x_j | y, x_1, ..., x_{i-1}) \]

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

Assumes feature \( i \) is independent of the other features given the label (i.e. are conditionally independent given the label)

For the wine problem?
**Naïve Bayes assumption**

\[ p(x_i | y, x_1, x_2, \ldots, x_{i-1}) = p(x_i | y) \]

Assumes feature i is independent of 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?

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:

\[ p(x_i | y, x_1, x_2, \ldots, x_{i-1}) \approx p(x_i | y) \]

**Naïve Bayes model**

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

\[ = p(y) \prod_{j=1}^{m} p(x_j | y) \quad \text{naive bayes assumption} \]

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

How do we model this?
- for binary features
- for discrete features, i.e. counts
- for real valued features

**p(x | y)**

Binary features:

\[ p(x_i | y) = \begin{cases} 
\theta_i & \text{if } x_i = 1 \\
1 - \theta_i & \text{otherwise} 
\end{cases} \quad \text{biased coin toss!} \]

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)
- ...

10/30/16
Basic steps for probabilistic modeling

<table>
<thead>
<tr>
<th>Step</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Pick a model</td>
</tr>
<tr>
<td>2</td>
<td>Figure out how to estimate the probabilities for the model</td>
</tr>
<tr>
<td>3 (optional)</td>
<td>Deal with overfitting</td>
</tr>
</tbody>
</table>

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?

Obtaining probabilities

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

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

Obtaining probabilities:

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

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?

\[ p(\text{head}) = 0.60 \]

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.

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

for each example

how probable is it under the model

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

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 \( \Theta = p(\text{head}) = 0.6 \)?

\[
0.60^{60} \times 0.40^{40} = 5.908465121038621 \times 10^{-30}
\]

60 heads with \( p(\text{head}) = 0.6 \)

40 tails with \( p(\text{tail}) = 0.4 \)
MLE example

Can we do any better? $\text{likelihood(data)} = \prod p(x_i)$

- $0.60^{60} \times 0.40^{40} = 5.908465121038621 \times 10^{-30}$

  60 heads with $p(\text{head}) = 0.6$
  40 tails with $p(\text{tail}) = 0.4$

What about $p(\text{head}) = 0.5$?

- $0.50^{60} \times 0.50^{40} = 7.88860952210118 \times 10^{-31}$

  60 heads with $p(\text{head}) = 0.5$
  40 tails with $p(\text{tail}) = 0.5$

What about $p(\text{head}) = 0.7$?

- $0.70^{60} \times 0.30^{40} = 6.176359828759916 \times 10^{-31}$

  60 heads with $p(\text{head}) = 0.7$
  40 tails with $p(\text{tail}) = 0.3$
The maximum likelihood estimate for a model parameter is the one that maximize the likelihood of the training data.

\[ MLE = \arg\max_{\theta} \prod_{i} p(x_i) \]

Often easier to work with log-likelihood:

\[ MLE = \arg\max_{\theta} \log\left(\prod_{i} p(x_i)\right) = \arg\max_{\theta} \sum_{i} \log(p(x_i)) \]

Calculating MLE

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

\[ \text{log-likelihood} = \sum \log(p(x_i)) \]

\[ = 60 \log(p(\text{heads})) + 40 \log(p(\text{tails})) \]

\[ = 60 \log(\theta) + 40 \log(1 - \theta) \]

\[ MLE = \arg \max_{\theta} 60 \log(\theta) + 40 \log(1 - \theta) \]

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.

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

\[
\frac{60}{\theta} - \frac{40}{1 - \theta} = 0
\]

\[
\frac{40}{\theta} - \frac{60}{1 - \theta}
\]

\[
40\theta = 60 - 60\theta
\]

\[
100\theta = 60
\]

\[
\theta = \frac{60}{100}
\]

Yay!

Calculating MLE

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

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

\[
\theta = \frac{a}{a + b}
\]

MLE estimation for NB

Maximum likelihood estimates

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

\[ p(x_1, y) = \frac{\text{count}(x_1, y)}{\text{count}(y)} \]

What does training a NB model then involve?

How difficult is this to calculate?
Naïve Bayes classification

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

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

Probabilistic models

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

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.

What is the generative story for the NB model?
NB generative story

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

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?

Some maths

\[ label = \log \arg \max_{\text{labels}} \log(p(y)) + \sum_{i=1}^{m} \log(p(x_i | y)) \]

\[ = \arg \max_{\text{labels}} \log(p(y)) + \sum_{i=1}^{m} \log(p(x_i | y)) \]

\[ = \arg \max_{\text{labels}} \log(p(y)) + \sum_{i=1}^{m} \log(p(x_i | y)) + \sum_{i=1}^{m} \log(1 - p(x_i | y)) \]

\[ p(x_i | y) = \begin{cases} 0 & \text{if } x_i = 1 \\ 1 & \text{if } x_i \neq 1 \end{cases} \]

Some more maths

\[ labels = \arg \max_{\text{labels}} \log(p(y)) + \sum_{i=1}^{m} \log(p(x_i | y)) + \sum_{i=1}^{m} \log(1 - p(x_i | y)) \]

\[ = \arg \max_{\text{labels}} \log(p(y)) + \sum_{i=1}^{m} \log(p(x_i | y)) + (1 - x_i) \log(1 - p(x_i | y)) \]

(because $x_i$ are binary)

\[ = \arg \max_{\text{labels}} \log(p(y)) + \sum_{i=1}^{m} \log(p(x_i | y)) - x_i \log(1 - p(x_i | y)) + \log(1 - p(x_i | y)) \]

\[ = \arg \max_{\text{labels}} \log(p(y)) + \sum_{i=1}^{m} \log \left( \frac{p(x_i | y)}{1 - p(x_i | y)} \right) + \log(1 - p(x_i | y)) \]
And…

\[
\text{labels} = \arg \max_{\alpha \in \text{labels}} \log(p(y)) + \sum_{i} \log \left( \frac{p(x_i | y)}{1 - p(x_i | y)} \right) + \log(1 - p(x_i | y))
\]

\[
= \arg \max_{\alpha \in \text{labels}} \log(p(y)) + \sum_{i} \log(1 - p(x_i | y)) + \sum_{i} \log \left( \frac{p(x_i | y)}{1 - p(x_i | y)} \right)
\]

What does this look like?

And…

\[
\text{labels} = \arg \max_{\alpha \in \text{labels}} \log(p(y)) + \sum_{i} \log \left( \frac{p(x_i | y)}{1 - p(x_i | y)} \right) + \log(1 - p(x_i | y))
\]

\[
= \arg \max_{\alpha \in \text{labels}} \log(p(y)) + \sum_{i} \log(1 - p(x_i | y)) + \sum_{i} \log \left( \frac{p(x_i | y)}{1 - p(x_i | y)} \right)
\]

\[
b + x_i \cdot w_i
\]

Linear model !!!

What are the weights?

NB as a linear model

\[
w_i = \log \left( \frac{p(x_i | y)}{1 - p(x_i | y)} \right)
\]

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

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

How do we deal with overfitting?

---

Coin experiment

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

\[ p(\text{not\_parasitic}) = 0.99999 \]

\[ p(\text{not\_parasitic})^{10000} = 0.905 \] 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 \( p(\text{parasitic}) = 0 \))
- 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?