Naïve Bayes

Will Monroe
August 11, 2017

with materials by Mehran Sahami and Chris Piech

image: Raneko

Naïve Bayes
Announcement: Problem Set #6

Due the last day of class, **Wednesday, August 16** (before class).

6 problems
(#6 involves serious coding!)

**Congressional voting**

**Heart disease diagnosis**

**No late days!**
Announcements: Final exam

A week from tomorrow:
Saturday, August 19, 12:15-3:15pm
Two pages (both sides) of notes
Comprehensive—material that was on the midterm will also be tested

Review session:
Wednesday, August 16, 2:30-3:20pm
in Huang 18 (location change!)
Maximum likelihood estimation

Choose parameters that maximize the likelihood (joint probability given parameters) of the example data.

\[ \hat{\theta} = \arg \max_\theta LL(\theta) \]
How to: MLE

1. Compute the likelihood.
\[ L(\theta) = P(X_1, \ldots, X_m | \theta) \]

2. Take its log.
\[ LL(\theta) = \log L(\theta) \]

3. Maximize this as a function of the parameters.
\[ \frac{d}{d\theta} LL(\theta) = 0 \]
The maximum likelihood $p$ for Bernoulli random variables is the sample mean.

$$\hat{p} = \frac{1}{m} \sum_{i=1}^{m} X_i$$
The maximum likelihood $\mu$ for normal random variables is the sample mean, and the maximum likelihood $\sigma^2$ is the “uncorrected” mean square deviation.

$$\hat{\mu} = \frac{1}{m} \sum_{i=1}^{m} X_i$$
$$\hat{\sigma}^2 = \frac{1}{m} \sum_{i=1}^{m} (X_i - \hat{\mu})^2$$
The maximum likelihood $a$ and $b$ for uniform random variables are the minimum and maximum of the data.

\[
\hat{a} = \min_{i} X_i \quad \hat{b} = \max_{i} X_i
\]
Maximum a posteriori estimation

Choose the most likely parameters given the example data. You’ll need a prior probability over the parameters.

\[
\hat{\theta} = \arg \max_{\theta} P(\theta | X_1, \ldots, X_n) = \arg \max_{\theta} \left[ LL(\theta) + \log P(\theta) \right]
\]
Laplace smoothing

Also known as **add-one** smoothing: assume you’ve seen one “imaginary” occurrence of each possible outcome.

\[
p_i = \frac{\#(X = i) + k}{n + mk} \]

\[
p_i = \frac{\#(X = i) + 1}{n + m}
\]
Classification

The most basic machine learning task:
predict a label from a vector of features.

\[ \hat{y} = \arg \max_y P(Y = y | \tilde{X} = \tilde{x}) \]
A cat detector

Features (Inputs)

<table>
<thead>
<tr>
<th>Feature</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ears up?</td>
<td>$x_1 = 1$</td>
</tr>
<tr>
<td>Whiskers?</td>
<td>$x_2 = 1$</td>
</tr>
<tr>
<td>&gt; Duck?</td>
<td>$x_3 = 0$</td>
</tr>
</tbody>
</table>

Prediction function

$\hat{y} = g(\tilde{x})$

Label (Output, Class)

Cat? Cat.

$y = 1$
How to make predictions

\[ \hat{y} = g(\vec{x}) \]
How to make predictions

\[ \hat{y} = g(\vec{x}) = \arg \max_y P(Y = y | \vec{X} = \vec{x}) \]

\( (Y | \vec{X}) \sim \text{Ber}(p = ???) \)
Training data

<table>
<thead>
<tr>
<th>Ears up?</th>
<th>Whiskers?</th>
<th>&gt; Duck?</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_1^{(1)} = 1$</td>
<td>$x_2^{(1)} = 0$</td>
<td>$x_3^{(1)} = 0$</td>
</tr>
<tr>
<td>$y^{(1)} = 0$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$x_1^{(2)} = 1$</td>
<td>$x_2^{(2)} = 1$</td>
<td>$x_3^{(2)} = 0$</td>
</tr>
<tr>
<td>$y^{(2)} = 1$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$x_1^{(3)} = 0$</td>
<td>$x_2^{(3)} = 0$</td>
<td>$x_3^{(3)} = 1$</td>
</tr>
<tr>
<td>$y^{(3)} = 0$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$x_1^{(4)} = 1$</td>
<td>$x_2^{(4)} = 1$</td>
<td>$x_3^{(4)} = 1$</td>
</tr>
<tr>
<td>$y^{(4)} = 1$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Three secret ingredients

1. Maximum likelihood or maximum a posteriori for conditional probabilities.

\[
\hat{P}(X_j = x_j | Y = y) = \frac{\#(X_j = x_j, Y = y) + 1}{\#(Y = y) + 2}
\]

2. “Naïve Bayes assumption”: features are independent conditioned on the label.

\[
\hat{P}(\bar{X} = \bar{x} | Y = y) = \prod_j \hat{P}(X_j = x_j | Y = y)
\]

3. (Take logs for numerical stability.)
Working with one feature

\[ y^{(1)} = 0 \quad x^{(1)} = 0 \]
\[ y^{(2)} = 1 \quad x^{(2)} = 1 \]
\[ y^{(3)} = 0 \quad x^{(3)} = 0 \]
\[ y^{(4)} = 1 \quad x^{(4)} = 1 \]

\[ \hat{y} = g(\bar{x}) = \arg \max_y \hat{P}(Y = y|X = x) \]
\[ P(Y = y|X = x) = \begin{cases} 
1-p_0 & x=0, y=0 \\
p_0 & x=0, y=1 \\
1-p_1 & x=1, y=0 \\
p_1 & x=1, y=1 
\end{cases} \]

\( (Y|X=0) \sim \text{Ber}(p_0) \)
\( (Y|X=1) \sim \text{Ber}(p_1) \)
Three secret ingredients

1. Maximum likelihood or maximum a posteriori for conditional probabilities.

\[
\hat{P}(X_j = x_j | Y = y) = \frac{\#(X_j = x_j, Y = y) + 1}{\#(Y = y) + 2}
\]

2. “Naïve Bayes assumption”: features are independent conditioned on the label.

\[
\hat{P}(\mathbf{X} = \mathbf{x} | Y = y) = \prod_j \hat{P}(X_j = x_j | Y = y)
\]

3. (Take logs for numerical stability.)
Working with one feature

Cat? (label)  Whiskers?

\[ y^{(1)} = 0 \quad x^{(1)} = 0 \]
\[ y^{(2)} = 1 \quad x^{(2)} = 1 \]
\[ y^{(3)} = 0 \quad x^{(3)} = 0 \]
\[ y^{(4)} = 1 \quad x^{(4)} = 1 \]

\[ \hat{y} = g(\bar{x}) = \arg \max_y \hat{P}(Y = y | X = x) \]

\[ P(Y = y | X = x) = \begin{cases} 
1 - p_0 & x = 0, y = 0 \\
p_0 & x = 0, y = 1 \\
1 - p_1 & x = 1, y = 0 \\
p_1 & x = 1, y = 1 
\end{cases} \]

\[ (Y|X = 0) \sim \text{Ber}(p_0) \]
\[ \hat{p}_0 = \frac{\#(X = 0, Y = 1)}{\#(X = 0)} \]

\[ (Y|X = 1) \sim \text{Ber}(p_1) \]
\[ \hat{p}_1 = \frac{\#(X = 1, Y = 1)}{\#(X = 1)} \]
Big (O) problems

Cat? (label)

$y^{(1)} = 0$

Ears up?

$x_{1}^{(1)} = 1$

Whiskers?

$x_{2}^{(1)} = 0$

$> \text{Duck?}$

$x_{3}^{(1)} = 0$

$\hat{y} = g(\vec{x}) = \arg \max_{y} \hat{P}(Y = y | \vec{X} = \vec{x})$

$P(Y = y | \vec{X} = \vec{x}) = \begin{pmatrix}
p_{000} & 1 - p_{000} & p_{100} & 1 - p_{100} \\
p_{001} & 1 - p_{001} & p_{101} & 1 - p_{101} \\
p_{010} & 1 - p_{010} & p_{110} & 1 - p_{110} \\
p_{011} & 1 - p_{011} & p_{111} & 1 - p_{111} \\
\end{pmatrix}$

$n \text{ features} \Rightarrow 2^n \text{ probability estimates}$
Naïve Bayes

A classification algorithm using the assumption that features are **conditionally independent** given the label.

\[ \hat{y} = \arg \max_y \hat{P}(Y = y) \prod_j \hat{P}(X_j = x_j | Y = y) \]
A strong assumption

\[ \hat{y} = g(\vec{x}) = \arg\max_y \hat{P}(Y = y|\vec{X} = \vec{x}) \]

\[ = \arg\max_y \frac{\hat{P}(Y = y) \hat{P}(\vec{X} = \vec{x}|Y = y)}{\hat{P}(\vec{X} = \vec{x})} \]

denominator is constant (with respect to \(y\))

\[ = \arg\max_y \hat{P}(Y = y) \hat{P}(\vec{X} = \vec{x}|Y = y) \]

Bayes!
Three secret ingredients

1. Maximum likelihood or maximum a posteriori for conditional probabilities.

\[ \hat{P}(X_j = x_j | Y = y) = \frac{\#(X_j = x_j, Y = y) + 1}{\#(Y = y) + 2} \]

2. “Naïve Bayes assumption”: features are independent conditioned on the label.

\[ \hat{P}(\tilde{X} = \tilde{x} | Y = y) = \prod_j \hat{P}(X_j = x_j | Y = y) \]

3. (Take logs for numerical stability.)
A strong assumption

\[ \hat{y} = g(\hat{x}) = \arg \max_y \hat{P}(Y = y | \hat{X} = \hat{x}) \]

\[ = \arg \max_y \frac{\hat{P}(Y = y) \hat{P}(\hat{X} = \hat{x} | Y = y)}{\hat{P}(\hat{X} = \hat{x})} \]

Bayes!

denominator is constant (with respect to \( y \))

\[ = \arg \max_y \hat{P}(Y = y) \hat{P}(\hat{X} = \hat{x} | Y = y) \]

“Naïve” Bayes!

\[ = \arg \max_y \hat{P}(Y = y) \prod_j \hat{P}(X_j = x_j | Y = y) \]

2n parameters (not \( 2^n \))
Three secret ingredients

1. Maximum likelihood or maximum a posteriori for conditional probabilities.

\[
\hat{P}(X_j = x_j | Y = y) = \frac{\#(X_j = x_j, Y = y) + 1}{\#(Y = y) + 2}
\]

2. “Naïve Bayes assumption”: features are independent conditioned on the label.

\[
\hat{P}(\tilde{X} = \tilde{x} | Y = y) = \prod_j \hat{P}(X_j = x_j | Y = y)
\]

3. (Take logs for numerical stability.)
A strong assumption

\[ \hat{y} = g(\bar{x}) = \arg \max_y \hat{P}(Y = y | \bar{X} = \bar{x}) \]

\[ = \arg \max_y \frac{\hat{P}(Y = y) \hat{P}(\bar{X} = \bar{x} | Y = y)}{\hat{P}(\bar{X} = \bar{x})} \]

Bayes!

denominator is constant (with respect to \( y \))

\[ = \arg \max_y \hat{P}(Y = y) \hat{P}(\bar{X} = \bar{x} | Y = y) \]

“Naïve” Bayes!

\[ = \arg \max_y \hat{P}(Y = y) \prod_j \hat{P}(X_j = x_j | Y = y) \]

2n parameters (not \( 2^n \))

take the log

\[ = \arg \max_y \left( \log \hat{P}(Y = y) + \sum_j \log \hat{P}(X_j = x_j | Y = y) \right) \]
A real-life cat detector

https://googleblog.blogspot.com/2012/06/usinag-large-scale-brain-simulations-for.html
Break time!
Movie recommendations

STAR WARS

THUMBS UP

THE LORD OF THE RINGS
THE RETURN OF THE KING

HP

THUMBS DOWN

?
Movie recommendations

\[
\hat{P}_{x_1|y} = \frac{\#(X_1 = x_1, Y = y)}{\#(Y = y)} \quad \hat{P}_{x_2|y} = \frac{\#(X_2 = x_2, Y = y)}{\#(Y = y)} \quad \hat{P}_y = \frac{\#(Y = y)}{n}
\]

\[
\hat{y} = g(\tilde{x}) = \arg\max_y \left( \log \hat{P}(Y = y) + \sum_j \log \hat{P}(X_j = x_j | Y = y) \right)
\]

\[
= \arg\max_y \left( \log \hat{P}_y + \log \hat{P}_{x_1=1|y} + \log \hat{P}_{x_2=0|y} \right)
\]

\[
\begin{pmatrix}
-0.84 + -0.26 + -0.96 \\
-0.56 + -0.27 + -0.89
\end{pmatrix}
\]

\[
= 1
\]
Movie recommendations: Laplace

(Laplace is not normally used for the prior)

<table>
<thead>
<tr>
<th>$X_1$</th>
<th>0</th>
<th>1</th>
<th>MLE estimates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>3</td>
<td>10</td>
<td>0.27 0.73</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>13</td>
<td>0.26 0.74</td>
</tr>
<tr>
<td></td>
<td>-1.31 -0.31</td>
<td>-1.35 -0.30</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$X_2$</th>
<th>0</th>
<th>1</th>
<th>MLE estimates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>5</td>
<td>8</td>
<td>0.40 0.60</td>
</tr>
<tr>
<td>1</td>
<td>7</td>
<td>10</td>
<td>0.42 0.58</td>
</tr>
<tr>
<td></td>
<td>-0.92 -0.51</td>
<td>-0.87 -0.54</td>
<td></td>
</tr>
</tbody>
</table>

\[
\hat{p}_{x_1|y} = \frac{\#(X_1 = x_1, Y = y) + 1}{\#(Y = y) + 2}
\]
\[
\hat{p}_{x_2|y} = \frac{\#(X_2 = x_2, Y = y) + 1}{\#(Y = y) + 2}
\]
\[
\hat{p}_y = \frac{\#(Y = y)}{n}
\]

\[
\hat{y} = g(\tilde{x}) = \arg\max_y \left( \log \hat{P}(Y = y) + \sum_j \log \hat{P}(X_j = x_j | Y = y) \right)
\]

\[
= \arg\max_y \left( \log \hat{p}_y + \log \hat{p}_{x_1=1|y} + \log \hat{p}_{x_2=0|y} \right) \begin{pmatrix} -0.84 & -0.31 & -0.92 \\ -0.56 & -0.30 & -0.87 \end{pmatrix}
\]

\[
= 1
\]
Hi all, I have a question about Chebyshev’s inequality. Is it possible to reverse the sign and...

Dear cs109-sum1617-staff: You have (12) messages from Singles in your Area! Click HERE...

Spam classification with Naive Bayes

<table>
<thead>
<tr>
<th>Spam? (label)</th>
<th>“a”</th>
<th>“Brazilian”</th>
<th>“Chebyshev”</th>
<th>...</th>
<th>“Viagra”</th>
<th>“Xavier”</th>
<th>“Z-value”</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td></td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Hi images: (left) Virginia State Parks; (right) Renee Comet
Hi all, I have a question about Chebyshev's inequality. Is it possible to reverse the sign and...

Dear cs109-sum1617-staff: You have (12) messages from Singles in your Area! Click HERE...

Spam classification with Naive Bayes

\[ \hat{P}(\text{"Viagra"} \mid \text{SPAM}) = \frac{\#(\text{SPAM}, \text{"Viagra"})}{\#(\text{SPAM})} \]

<table>
<thead>
<tr>
<th>Spam? (label)</th>
<th>&quot;a&quot;</th>
<th>&quot;Brazilian&quot;</th>
<th>&quot;Chebyshev&quot;</th>
<th>...</th>
<th>&quot;Viagra&quot;</th>
<th>&quot;Xavier&quot;</th>
<th>&quot;Z-value&quot;</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

images: (left) Virginia State Parks; (right) Renee Comet
Classifying a new email

Buy Brazilian Viagra! A special discount just for you, cs109-sum1617-staff...

<table>
<thead>
<tr>
<th>Spam? (label)</th>
<th>“a”</th>
<th>“Brazilian”</th>
<th>“Chebyshev”</th>
<th>...</th>
<th>“Viagra”</th>
<th>“Xavier”</th>
<th>“Z-value”</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td></td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

\[
\hat{y} = g(\hat{x}) = \arg \max_y \left( \log \hat{P}(Y = y) + \sum_j \log \hat{P}(X_j = x_j | Y = y) \right)
\]

\[
\log \hat{P}(\text{SPAM}) + \log \hat{P}(\text{"a" | SPAM}) + \log \hat{P}(\text{"Brazilian" | SPAM}) + \log \hat{P}(\text{"Chebyshev" | SPAM}) + \cdots = -8.2
\]

\[
\log \hat{P}(\text{HAM}) + \log \hat{P}(\text{"a" | HAM}) + \log \hat{P}(\text{"Brazilian" | HAM}) + \log \hat{P}(\text{"Chebyshev" | HAM}) + \cdots = -12.1
\]

\[
\hat{y} = 1
\]
Two envelopes

$X$

$2X$

$Y = \text{amount in envelope chosen}$

$E[W|\text{stay}] = Y$

$E[W|\text{switch}] = \frac{Y}{2} \cdot 0.5 + 2Y \cdot 0.5$

$= \frac{5}{4}Y$