Naïve Bayes

Daniel Lowd
Based on slides by Vibhav Gogate

Text classification

• Classify e-mails
  – \( Y = \{ \text{Spam, NotSpam} \} \)
• Classify news articles
  – \( Y = \{ \text{what is the topic of the article?} \} \)
• Classify webpages
  – \( Y = \{ \text{Student, professor, project, ...} \} \)
• What to use for features, \( X \)?

What to use for features, \( X \)?

Article from rec.sport.hockey

Path: /var/spool/postfix/spool/queue/incoming/200601/76915
From: zbevry@msn.com (John Doe)
Subject: Re: This year's biggest and worst (opinion)
Date: Fri, 9 Dec 2006 03:09:03 GMT

I can rely on the Kings, but the most obvious candidate for pleasant surprises is Alexei Yashin. He is back in his best form and is the most consistent player in the league. The team has also made a number of trades to strengthen their defense. Overall, I think the Kings will make it to the playoffs.

Bag of Words Model

Silly assumption: Assume that the order of the words doesn’t matter.

Equivalently:

{assume, doesn’t, matter, of, order, that, the, words}

Bernoulli model: \( X_i = 1 \) if the \( i \)th dictionary word is present in the document.

Multinomial model: \( X_i = n \) if the word is present in the document \( n \) times.

Bag of Words Features

<table>
<thead>
<tr>
<th>Feature</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>about</td>
<td>2</td>
</tr>
<tr>
<td>Africa</td>
<td>1</td>
</tr>
<tr>
<td>apple</td>
<td>0</td>
</tr>
<tr>
<td>access</td>
<td>0</td>
</tr>
<tr>
<td>gas</td>
<td>1</td>
</tr>
<tr>
<td>oil</td>
<td>1</td>
</tr>
<tr>
<td>Zaire</td>
<td>0</td>
</tr>
</tbody>
</table>

Bayesian Categorization

\[ P(y_i \mid X) \sim P(y_i) P(X \mid y_i) \]

Need to know:

• Priors: \( P(y_1) \), \( P(y_2) \), ...
• Conditionals: \( P(x \mid y_1) \), \( P(x \mid y_2) \), ...

\( P(y_i) \) are easily estimated from data.

Exponentially many \( x \) (even with bag of words assumption)
Need to Simplify Somehow

- Too many probabilities
  \[ P(X_1 \land X_2 \land X_3 \mid y) \]
- Can we assume some are the same?
  \[ P(x_1 \land x_2 \mid y) = P(x_1 \mid y) P(x_2 \mid y) \]

The Naïve Bayes Classifier

- Given:
  - Prior \( P(y) \)
  - \( n \) conditionally independent features \( X \) given the class \( Y \)
  - For each \( X \), we have likelihood \( P(X \mid Y) \)
- Decision rule:
  \[ y^* = h_{NB}(x) = \arg \max_y P(y) P(x_1, \ldots, x_n \mid y) \]
  \[ = \arg \max_y P(y) \prod_i P(x_i \mid y) \]

Naïve Bayes

- Naïve Bayes assumption:
  - Features are independent given class:
    \[ P(X_1, X_2 \mid Y) = P(X_1 \mid Y) P(X_2 \mid Y) \]
  - More generally:
    \[ P(X_1 \ldots X_n \mid Y) = \prod_i P(X_i \mid Y) \]
- How many parameters now?
  - Suppose \( X \) is composed of \( n \) binary features

MLE for the parameters of NB

- Given dataset, count occurrences for all pairs
  \[ \text{Count}(X = x, Y = y) \]
- MLE for discrete NB, simply:
  - Prior:
    \[ P(Y = y) = \frac{\text{Count}(Y = y)}{\sum_y \text{Count}(Y = y)} \]
  - Likelihood:
    \[ P(X_i = x \mid Y = y) = \frac{\text{Count}(X_i = x, Y = y)}{\sum_{x'} \text{Count}(X_i = x', Y = y)} \]

NÀIVE BAYES CALCULATIONS

<table>
<thead>
<tr>
<th>Day</th>
<th>Outlook</th>
<th>Temperature</th>
<th>Humidity</th>
<th>Wind</th>
<th>PlayTennis</th>
</tr>
</thead>
<tbody>
<tr>
<td>D1</td>
<td>Sunny</td>
<td>Hot</td>
<td>High</td>
<td>Weak</td>
<td>No</td>
</tr>
<tr>
<td>D2</td>
<td>Sunny</td>
<td>Hot</td>
<td>High</td>
<td>Strong</td>
<td>No</td>
</tr>
<tr>
<td>D3</td>
<td>Overcast</td>
<td>Hot</td>
<td>High</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D4</td>
<td>Rain</td>
<td>Mild</td>
<td>High</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D5</td>
<td>Rain</td>
<td>Cool</td>
<td>Normal</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D6</td>
<td>Rain</td>
<td>Cool</td>
<td>Normal</td>
<td>Strong</td>
<td>No</td>
</tr>
<tr>
<td>D7</td>
<td>Overcast</td>
<td>Cool</td>
<td>Normal</td>
<td>Strong</td>
<td>Yes</td>
</tr>
<tr>
<td>D8</td>
<td>Sunny</td>
<td>Mild</td>
<td>High</td>
<td>Weak</td>
<td>No</td>
</tr>
<tr>
<td>D9</td>
<td>Sunny</td>
<td>Cool</td>
<td>Normal</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D10</td>
<td>Rain</td>
<td>Mild</td>
<td>Normal</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D11</td>
<td>Sunny</td>
<td>Mild</td>
<td>Normal</td>
<td>Strong</td>
<td>Yes</td>
</tr>
<tr>
<td>D12</td>
<td>Overcast</td>
<td>Mild</td>
<td>High</td>
<td>Strong</td>
<td>Yes</td>
</tr>
<tr>
<td>D13</td>
<td>Overcast</td>
<td>Hot</td>
<td>Normal</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D14</td>
<td>Rain</td>
<td>Mild</td>
<td>High</td>
<td>Strong</td>
<td>No</td>
</tr>
</tbody>
</table>
Subtleties of NB Classifier: #2
Insufficient Training Data
- What if you never see a training instance where X1=a and Y=b
  - You never saw Y=spam, X1=Obama
  - P(X=Obama|Y=spam)=0
- Thus no matter what values x2, x3,...,xn take:
  - P(x1=Obama, x2=a2,...,xn=an|Y=Obama)=0
  - Why?

\[ y^* = h_{NB}(x) = \arg \max_y P(y)P(x_1, \ldots, x_n \mid y) \]
\[ = \arg \max_y P(y) \prod_i P(x_i \mid y) \]

For Binary Features: We already know the answer!

\[ P(\theta \mid D) = \frac{\theta^{y_H+n_H-1}(1-\theta)^{y_T+n_T-1}}{B(y_H+n_H, y_T+n_T)} \sim \text{Beta}(\theta_H+n_H, \theta_T+n_T), \]
- MAP: use most likely parameter
\[ \hat{\theta} = \arg \max_y P(\theta \mid D) = \frac{\alpha_H+n_H-1}{\alpha_H+n_H+\alpha_T+n_T-2} \]
- Beta prior equivalent to extra observations for each feature
  - As N → 1, prior is “forgotten”
  - But, for small sample size, prior is important!

That’s Great for Binomial
- Works for Spam / Ham
- What about multiple classes
  - E.g., given a Wikipedia page, predicting type

Probabilities: Important Detail!
- P(spam | X1 ... Xn) = \prod P(spam | X_i)
  Any more potential problems here?
- We are multiplying lots of small numbers
  Danger of underflow!
  - 0.5^{18} = 7 \times 10^{-18}
- Solution? Use logs and add!
  - P_1 \times P_2 = e^{log(P_1)+log(P_2)}
  - Always keep in log form

Multinomials: Laplace Smoothing
- Laplace’s estimate:
  - Pretend you saw every outcome k extra times
    \[ P_{LAP}(x) = \frac{c(x)+k}{N+k|X|} \]
    \[ P_{LAP}(y) = \begin{cases} 2/3 & |y|=2 \\ 3/5 & |y|=3 \\ 102/203 & |y|=10 \end{cases} \]
  - What’s Laplace with k = 0?
  - k is the strength of the prior
  - Can derive this as a MAP estimate for multinomial with Dirichlet priors
- Laplace for conditionals:
  - Smooth each condition independently:
    \[ P_{LAP}(x \mid y) = \frac{c(x,y)+k}{c(y)+k|X|} \]

Naïve Bayes Posterior Probabilities
- Classification results of naïve Bayes
  - I.e. the class with maximum posterior probability...
  - Usually fairly accurate (?!?!?)
- However, due to the inadequacy of the conditional independence assumption...
  - Actual posterior-probability estimates not accurate.
  - Output probabilities generally very close to 0 or 1.
NB with Bag of Words for Text Classification

• Learning phase:
  – Prior $P(Y_m)$
  – Count how many documents from each topic (prior)
  – $P(X_i | Y_m)$
  • Let $B$ be the bag of words created from the union of all docs
  • Let $B_m$ be a bag of words formed from all the docs in topic $m$
  • Let $\#(i, B)$ be the number of times word $i$ is in bag $B$
  - $P(X_i | Y_m)$ is proportional to $(\#(i, B_m)+1)$

• Test phase:
  – For each document
    • Use naïve Bayes decision rule
      $$h_{NB}(x) = \arg \max_y P(y) \prod_{i=1}^{LengthDoc} P(x_i | y)$$

Twenty News Groups results

<table>
<thead>
<tr>
<th>Group</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>comp.graphics</td>
<td>comp.graphics, windows, mpc</td>
</tr>
<tr>
<td>comp.os.ms-windows, mpc</td>
<td>comp.sys.ibm.pc.hardware</td>
</tr>
<tr>
<td>comp.windows.x</td>
<td>rec.sport.hockey</td>
</tr>
<tr>
<td>alt.atheism</td>
<td>sci.space</td>
</tr>
<tr>
<td>soc.religion.religion</td>
<td>sci.crypt</td>
</tr>
<tr>
<td>talk.religion.misce</td>
<td>cs.soc</td>
</tr>
<tr>
<td>talk.politics.misce</td>
<td>talk.politics.political</td>
</tr>
<tr>
<td>talk.politics.misce</td>
<td>talk.politics.misce</td>
</tr>
<tr>
<td>talk.politics.misce</td>
<td>talk.politics.misce</td>
</tr>
<tr>
<td>talk.politics.misce</td>
<td>talk.politics.misce</td>
</tr>
</tbody>
</table>

Naïve Bayes: 89% classification accuracy

Learning curve for Twenty News Groups

Accuracy vs. Training set size (1/3 withheld for test)

Bayesian Learning

What if Features are Continuous?

Eg., Character Recognition: $X_i$ is $i$th pixel

$$P(Y | X) \propto P(X | Y) P(Y)$$

Gaussian Naive Bayes

Sometimes Assume Variance

- is independent of $Y$ (i.e., $\sigma_i$),
- or independent of $X_i$ (i.e., $\sigma_i$),
- or both (i.e., $\sigma_i$)

$$P(Y | X) \propto P(X | Y) P(Y)$$

$$P(X_i | Y=y_k) = N(\mu_{ik}, \sigma_{ik})$$

$$N(\mu_{ik}, \sigma_{ik}) = \frac{1}{\sigma_{ik}\sqrt{2\pi}} e^{-\frac{(x-\mu_{ik})^2}{2\sigma_{ik}^2}}$$
Learning Gaussian Parameters
Maximum Likelihood Estimates:
• Mean:
  \[ \hat{\mu}_{M LE} = \frac{1}{N} \sum_{i=1}^{N} x_i \]
• Variance:
  \[ \hat{\sigma}^2_{M LE} = \frac{1}{N} \sum_{i=1}^{N} (x_i - \hat{\mu})^2 \]

Learning Gaussian Parameters
Maximum Likelihood Estimates:
• Mean:
  \[ \hat{\mu}_{M LE} = \frac{1}{\sum_j \delta(Y^j = y_k)} \sum_j x_i^j \delta(Y^j = y_k) \]
• Variance:
  \[ \hat{\sigma}^2_{M LE} = \frac{1}{\sum_j \delta(Y^j = y_k) - 1} \sum_j (x_i^j - \hat{\mu})^2 \delta(Y^j = y_k) \]

What you need to know about Naïve Bayes
• Naïve Bayes classifier
  – What’s the assumption
  – Why we use it
  – How do we learn it
  – Why is Bayesian estimation important
• Text classification
  – Bag of words model
• Gaussian NB
  – Features are still conditionally independent
  – Each feature has a Gaussian distribution given class