Naïve Bayes

Vibhav Gogate
The University of Texas at Dallas
Supervised Learning of Classifiers

Find $f$

• Given: Training set $\{(x_i, y_i) \mid i = 1 \ldots n\}$

• Find: A good approximation to $f : X \rightarrow Y$

Examples: what are $X$ and $Y$?

• Spam Detection
  – Map email to {Spam,Ham}

• Digit recognition
  – Map pixels to {0,1,2,3,4,5,6,7,8,9}

• Stock Prediction
  – Map new, historic prices, etc. to (the real numbers)
Bayesian Categorization/Classification

- Let the set of categories be \( \{c_1, c_2, \ldots, c_n\} \)
- Let \( E \) be description of an instance.
- Determine category of \( E \) by determining for each \( c_i \)

\[
P(c_i \mid E) = \frac{P(c_i)P(E \mid c_i)}{P(E)}
\]

- \( P(E) \) can be ignored (normalization constant)

\[
P(c_i \mid E) \sim P(c_i)P(E \mid c_i)
\]

- Select the class with the max. probability.
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, $X$?
Features $\mathbf{X}$ are word sequence in document $X_i$ for $i^{th}$ word in article

Article from rec.sport.hockey

Path: cantaloupe.srv.cs.cmu.edu!das-news.harvard.e
From: xxx@yyy.zzz.edu (John Doe)
Subject: Re: This year’s biggest and worst (opinic
Date: 5 Apr 93 09:53:39 GMT

I can only comment on the Kings, but the most obvious candidate for pleasant surprise is Alex Zhitnik. He came highly touted as a defensive defenseman, but he’s clearly much more than that. Great skater and hard shot (though wish he were more accurate). In fact, he pretty much allowed the Kings to trade away that huge defensive liability Paul Coffey. Kelly Hrudey is only the biggest disappointment if you thought he was any good to begin with. But, at best, he’s only a mediocre goaltender. A better choice would be Tomas Sandstrom, though not through any fault of his own, but because some thugs in Toronto decided
Features for Text Classification

- $X$ is sequence of words in document
- $X$ (and hence $P(X|Y)$) is huge!!!
  - Article at least 1000 words, $X=\{X_1,...,X_{1000}\}$
  - $X_i$ represents $i^{th}$ word in document, i.e., the domain of $X_i$ is entire vocabulary, e.g., Webster Dictionary (or more), 10,000 words, etc.
- $10,000^{1000} = 10^{4000}$
- Atoms in Universe: $10^{80}$
  - We may have a problem...
Bag of Words Model

Typical additional assumption –

- **Position in document doesn’t matter:**
  - $P(X_i=x_i | Y=y) = P(X_k=x_i | Y=y)$
  - (all positions have the same distribution)

- Ignore the order of words
- Sounds really silly, but often works very well!

- From now on:
  - $X_i =$ Boolean: “word $i$ is in document”
  - $X = X_1 \land \ldots \land X_n$
Bag of Words Approach

Our energy exploration, production, and distribution operations span the globe, with activities in more than 100 countries.

At TOTAL, we draw our greatest strength from our fast-growing oil and gas reserves. Our strategic emphasis on natural gas provides a strong position in a rapidly expanding market.

Our expanding refining and marketing operations in Asia and the Mediterranean Rim complement already solid positions in Europe, Africa, and the U.S.

Our growing specialty chemicals sector adds balance and profit to the core energy business.
Bayesian Categorization

\[ P(y_1 \mid X) \sim P(y_i)P(X \mid y_i) \]

• Need to know:
  – Priors: \( P(y_i) \)
  – Conditionals: \( P(X \mid y_i) \)
• \( P(y_i) \) are easily estimated from data.
  – If \( n_i \) of the examples in \( D \) are in \( y_i \), then \( P(y_i) = n_i / |D| \)
• Conditionals:
  – \( X = X_1 \land \ldots \land X_n \)
  – Estimate \( P(X_1 \land \ldots \land X_n \mid y_i) \)
• Too many possible instances to estimate!
  – \((\text{exponential in } n)\)
  – Even \textbf{with} bag of words assumption!
Need to Simplify Somehow

• Too many probabilities
  – \( P(x_1 \land x_2 \land x_3 \mid y_i) \)

• Can we assume some are the same?
  – \( P(x_1 \land x_2 \mid y_i) = P(x_1 \mid y_i) \cdot P(x_2 \mid y_i) \)

\[
\begin{align*}
P(x_1 \land x_2 \land x_3 \mid \text{spam}) \\
P(x_1 \land x_2 \land \neg x_3 \mid \text{spam}) \\
P(x_1 \land \neg x_2 \land x_3 \mid \text{spam}) \\
\quad \text{...} \\
P(\neg x_1 \land \neg x_2 \land \neg x_3 \mid \neg \text{spam})
\end{align*}
\]
Conditional Independence

- X is **conditionally independent** of Y given Z, if the probability distribution for X is independent of the value of Y, given the value of Z

\[(\forall i, j, k) P(X = i | Y = j, Z = k) = P(X = i | Z = k)\]

- e.g.,
  \[P(\text{Thunder} | \text{Rain, Lightning}) = P(\text{Thunder} | \text{Lightning})\]

- Equivalent to:
  \[P(X, Y | Z) = P(X | Z)P(Y | Z)\]
Naïve Bayes

• Naïve Bayes assumption:
  – Features are independent given class:

$$P(X_1, X_2 | Y) = P(X_1 | X_2, Y)P(X_2 | Y)$$

$$= P(X_1 | Y)P(X_2 | Y)$$

  – More generally:

$$P(X_1 \ldots X_n | Y) = \prod_i P(X_i | Y)$$

• How many parameters now?
  – Suppose $X$ is composed of $n$ binary features
The Naïve Bayes Classifier

- **Given:**
  - Prior $P(Y)$
  - $n$ conditionally independent features $X$ given the class $Y$
  - For each $X_i$, we have likelihood $P(X_i | Y)$

- **Decision rule:**
  $$y^* = h_{NB}(x) = \arg \max_y P(y)P(x_1, \ldots, x_n | y)$$
  $$= \arg \max_y P(y) \prod_i P(x_i | y)$$
MLE for the parameters of NB

• Given dataset, count occurrences for all pairs
  – Count($X_i=x_i, Y=y$)------ How many pairs?

• MLE for discrete NB, simply:
  – Prior:
    \[ P(Y = y) = \frac{\text{Count}(Y = y)}{\sum_{y'} \text{Count}(Y = y')} \]
  – Likelihood:
    \[ P(X_i = x | Y = y) = \frac{\text{Count}(X_i = x, Y = y)}{\sum_{x'} \text{Count}(X_i = x', Y = y)} \]
## NAÏVE 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: #1
Violating the NB Assumption

• Usually, features are not conditionally independent:

\[ P(X_1 \ldots X_n | Y) \neq \prod_i P(X_i | Y) \]

• The naïve Bayes assumption is often violated, yet it performs surprisingly well in many cases.

• Plausible reason: Only need the probability of the correct class to be the largest!
  
  – Example: two-way classification; just need to figure out the correct side of 0.5 and not the actual probability (0.51 is the same as 0.99).
Subtleties of NB Classifier: #2
Insufficient Training Data

- What if you never see a training instance where $X_1=a$ and $Y=b$
  - You never saw, $Y=$Spam, $X_1=${Enlargement}
  - $P(X_1=$Enlargement$| Y=$Spam$)=0$

- Thus no matter what values $X_2, X_3, \ldots, X_n$ take:
  - $P(X_1=$Enlargement, $X_2=a_2, \ldots, X_n=a_n| Y=$Spam$)=0$
  - 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!

\[ P(\theta \mid \mathcal{D}) = \frac{\theta^{\beta_H + \alpha_H - 1}(1 - \theta)^{\beta_T + \alpha_T - 1}}{B(\beta_H + \alpha_H, \beta_T + \alpha_T)} \sim \text{Beta}(\beta_H + \alpha_H, \beta_T + \alpha_T) \]

- **MAP**: use most likely parameter

\[ \hat{\theta} = \arg \max_{\theta} P(\theta \mid \mathcal{D}) = \frac{\alpha_H + \beta_H - 1}{\alpha_H + \beta_H + \alpha_T + \beta_T - 2} \]

- **Beta prior** equivalent to extra observations for each feature
- As \( N \to 1 \), prior is “forgotten”
- **But, for small sample size, prior is important!**
That’s Great for Binomial

• Works for Spam / Ham
• What about multiple classes
  – Eg, given a wikipedia page, predicting type
Multinomials: Laplace Smoothing

• Laplace’s estimate:
  – Pretend you saw every outcome $k$ extra times
    $$P_{LAP,k}(x) = \frac{c(x) + k}{N + k|X|}$$
  – 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:
    $$P_{LAP,k}(x|y) = \frac{c(x, y) + k}{c(y) + k|X|}$$

\[
\begin{align*}
P_{LAP,0}(X) &= \left\langle \frac{2}{3}, \frac{1}{3} \right\rangle \\
P_{LAP,1}(X) &= \left\langle \frac{3}{5}, \frac{2}{5} \right\rangle \\
P_{LAP,100}(X) &= \left\langle \frac{102}{203}, \frac{101}{203} \right\rangle
\end{align*}
\]
Probabilities: Important Detail!

- $P(\text{spam} \mid X_1 \ldots X_n) = \prod_i P(\text{spam} \mid X_i)$

Any more potential problems here?

- We are multiplying lots of small numbers
  Danger of underflow!
  - $0.5^{57} = 7 \times 10^{-18}$

- Solution? Use logs and add!
  - $p_1 \times p_2 = e^{\log(p_1)+\log(p_2)}$
  - Always keep in log form
NB for Text Classification: Learning

• Learning phase: \( P(Y_m) \) and \( P(X_i|Y_m) \)

Prior: \( P(Y_m) \)

\[
P(Y_m) = \frac{N_m}{N}
\]

where \( N_m \) is the number of documents having class label \( m \) and \( N \) is the total number of documents.

Class conditional probabilities: \( P(X_i|Y_m) \)

\[
P(X_i|Y_m) = \frac{\text{Count}(X_i, Y_m) + 1}{\sum_{j=1}^{V}(\text{Count}(X_j, Y_m) + 1)}
\]

where \( V \) is the size of the vocabulary (number of distinct words) in all documents and \( \text{Count}(X_i, Y_m) \) is the number of times the word \( X_i \) appears in documents of class \( Y_m \).
NB for Text Classification: Classification

• Given a new document having length “L”

\[
\arg\max_Y P(Y) \prod_{i=1}^{L} P(X_i | Y)
\]
Example: (Borrowed from Dan Jurafsky)

<table>
<thead>
<tr>
<th>Doc</th>
<th>Words</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>1   Chinese Beijing Chinese</td>
<td>c</td>
</tr>
<tr>
<td></td>
<td>2   Chinese Chinese Shanghai</td>
<td>c</td>
</tr>
<tr>
<td></td>
<td>3   Chinese Macao</td>
<td>c</td>
</tr>
<tr>
<td></td>
<td>4   Tokyo Japan Chinese</td>
<td>j</td>
</tr>
<tr>
<td>Test</td>
<td>5   Chinese Chinese Chinese Tokyo Japan</td>
<td>?</td>
</tr>
</tbody>
</table>

Priors:
\[
P(c) = \frac{3}{4}, \quad P(j) = \frac{1}{4}
\]

Choosing a class:
\[
P(c|d5) \propto \frac{3}{4} \times \left(\frac{3}{7}\right)^3 \times \frac{1}{14} \times \frac{1}{14} \\
\approx 0.0003
\]

Conditional Probabilities:
\[
P(\text{Chinese} | c) = \frac{5+1}{8+6} = \frac{6}{14} = \frac{3}{7}
\]
\[
P(\text{Tokyo} | c) = \frac{0+1}{8+6} = \frac{1}{14}
\]
\[
P(\text{Japan} | c) = \frac{0+1}{8+6} = \frac{1}{14}
\]
\[
P(\text{Chinese} | j) = \frac{1+1}{3+6} = \frac{2}{9}
\]
\[
P(\text{Tokyo} | j) = \frac{1+1}{3+6} = \frac{2}{9}
\]
\[
P(\text{Japan} | j) = \frac{1+1}{3+6} = \frac{2}{9}
\]
Bayesian Learning

What if Features are Continuous?

Eg., Character Recognition: \( X_i \) is \( i^{th} \) pixel

\[
P(Y \mid X) \propto P(X \mid Y) P(Y)
\]
Bayesian Learning

What if Features are Continuous?

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

\[
P(Y \mid X) \propto P(X \mid Y) P(Y)
\]

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

\[
N(\mu_{ik}, \sigma_{ik}) = \frac{1}{\sigma_{ik} \sqrt{2\pi}} e^{-\frac{(x-\mu_{ik})^2}{2\sigma_{ik}^2}}
\]
Gaussian Naïve Bayes

Sometimes Assume Variance

\( P(X_i = x \mid Y = y_k) = N(\mu_{ik}, \sigma_{ik}) \)

\[ N(\mu_{ik}, \sigma_{ik}) = \frac{1}{\sigma_{ik} \sqrt{2\pi}} e^{-\frac{(x-\mu_{ik})^2}{2\sigma_{ik}^2}} \]

\[ P(Y \mid X) \propto P(X \mid Y) P(Y) \]
Learning Gaussian Parameters

Maximum Likelihood Estimates:

- Mean:
  \[ \hat{\mu}_{MLE} = \frac{1}{N} \sum_{i=1}^{N} x_i \]

- Variance:
  \[ \hat{\sigma}^2_{MLE} = \frac{1}{N} \sum_{i=1}^{N} (x_i - \hat{\mu})^2 \]
Learning Gaussian Parameters

Maximum Likelihood Estimates:

• Mean:

\[ \hat{\mu}_{ik} = \frac{1}{\sum_j \delta(Y^j = y_k)} \sum_j X^j_i \delta(Y^j = y_k) \]

\( \delta(x) = 1 \) if \( x \) true, else 0

• Variance:

\[ \hat{\sigma}^2_{MLE} = \frac{1}{N} \sum_{i=1}^{N} (x_i - \hat{\mu})^2 \]
Learning Gaussian Parameters

Maximum Likelihood Estimates:

• Mean:

\[
\hat{\mu}_{ik} = \frac{1}{\sum_j \delta(Y_j = y_k)} \sum_j X_i^j \delta(Y_j = y_k)
\]

• Variance:

\[
\hat{\sigma}^2_{ik} = \frac{1}{\sum_j \delta(Y_j = y_k) - 1} \sum_j (X_i^j - \hat{\mu}_{ik})^2 \delta(Y_j = y_k)
\]
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