Naïve Bayes Text Classification

Reading:

Manning, Raghavan, and Schutze, Text classification and Naive Bayes, pp. 253-270(Chapter 13 in Introduction to Information Retrieval)
Classification, learning, and generalization

- General description of classification:
  
  Given a feature vector representing a possible instance of a class, 
  
  \[ \mathbf{x} = <x_1, x_2, \ldots, x_n>, \]
  
  classify \( \mathbf{x} \) as one of a set of classes \( c \in C \).
Classification, learning, and generalization

• General description of **classification**:
  
  Given a *feature vector* representing a possible instance of a class,  
  \[ x = \langle x_1, x_2, ..., x_n \rangle, \]
  classify \( x \) as one of a set of classes \( c \in C \).

• General description of **supervised learning**:
  
  Given a set of *training examples*  
  \[ \{(x, c(x))_{\text{train}}\}, \]
  where \( c(x) \) is the correct classification of \( x \), construct a hypothesis \( h \) that will correctly classify these training examples (with the goal of generalization).
Classification, learning, and generalization

- General description of **classification**:  
  Given a *feature vector* representing a possible instance of a class, 
  \[ x = <x_1, x_2, ..., x_n>, \]
  classify \( x \) as one of a set of classes \( c \in C \).

- General description of **supervised learning**:  
  Given a set of *training examples*  
  \[ \{(x, c(x))_{\text{train}}\}, \]
  where \( c(x) \) is the correct classification of \( x \), construct a hypothesis \( h \) that will correctly classify these training examples (with the goal of generalization).

- General description of **successful generalization**:  
  Given a set of *test examples* \( \{(x, c(x))_{\text{test}}\} \), not seen before but drawn from the same distribution of the training examples, hypothesis \( h \) will correctly classify these test examples.
Example: Detecting spam
HOLIDAY SPECIAL: SAVE UP TO $10 ON YOUR PURCHASES
(order now and receive by Christmas)

With the holiday season rapidly approaching, we want to remind you of our most generous sale of the year. As a valued customer, we invite you to save up to $10 off your Alibris purchases with three ways to save:

$2 off your order of $20 or more: GIFT2
$5 off your order of $50 or more: GIFT5
$10 off your order of $100 or more: GIFT10

Simply enter the coupon codes above* at checkout. But hurry, this limited time offer expires on December 16, 2003. Visit Alibris now and save!

Save money on shipping too! Now through December 9, 2003, every item listed on our site should be delivered to continental U.S. and Canadian addresses by December 24th via standard shipping (our lowest cost option) or get FREE shipping when you order $49 of In Stock books. Don't delay, start your holiday shopping now.

http://alibris.m0.net/m/S.asp?HB10950943733X2869462X274232X
Can your website answer questions in real time 24 hours a day, 7 days a week? Our clients websites do and we're not talking about some stale FAQ sheet either. Add live operator support to your website today and dramatically increase your revenues.

<a href="http://www.dreamscaper.co.mn@click.net-click.net.ph/click.php?id=3Ddrcomnm">stop sending me emails</a>
I know this proposal letter may come to you as a surprise considering the fact that we have not had any formal acquaintance before. But all the same I would want you for the sake of God to give this an immediate attention in view of the fact that the security of our live and possession is at stake.

I am Mr. JAMES KEN, 28 years old, from war-ravaged SIERRA LEONE but presently domiciled in Abidjan, Ivory coast, with my sister JANET who is 18 years old. My father Mr. KEN, who before his untimely assassination by the rebels was the Director of SIERRA LEONE Diamond corporation (SLDC). He was killed in our government residential house along with two of my other brothers, two house maids and one government attached security guard. Fortunately for me, younger sister and mother, we were on a week end visit to our home town as we got the news of the tragedy. We immediately managed to run into neighboring Ivory coast for refuge. But unfortunately. As Fate would have it, we lost our dear mother (may soul rest in peace) as a result of what the Doctor called cardiac arrest.
As we were coming into this country, we had some documents of a deposit of $11,700,000 USD (eleven million seven hundred thousand USD) made by my late father in a security and trust company. According to my father, he intended to use this fund for his international business transaction after his tenure in office but was unfortunately murdered. We had located the security company where the money is deposited with the help of an attorney and established ownership. Please right now, with the bitter experiences we had in our country and the war still going on especially in diamond area which incidentally is where we hail from. Coupled with the incessant political upheavals and hostilities in this country, Ivory Coast, we desire seriously to leave here and live the rest of our life into a more peaceful and politically stable country like yours. Hence this proposal and request. We therefore wish you can help us in the following regards:

1) To provide us with a good bank account to transfer the money into.
2) To help us invest the money into a lucrative business.
3) To assist my sister Janet get a college admission to further her education.
Please I know that, this letter may sound strange and incredible to you but the CNN and the BBC African bulletin normally have it as their major news features. Therefore for the sake of God and humanity give an immediate positive consideration and reply to me via our e-mail address. I will willingly agree to any suitable percentage of the money you will propose as your compensation for your assistance with regards to the above. Please in view of our sensitive refugee status and as we are still conscious of our father's enemies I would like you to give this a highly confidential approach.

Best Regards.
JAMES KEN.
Spamassassin results

X-Spam-Report: ---- Start SpamAssassin results
   6.70 points, 4 required;
   * 0.4 -- BODY: Offers a limited time offer
   * 0.1 -- BODY: Free Offer
   * 0.4 -- BODY: Stop with the offers, coupons, discounts etc!
   * 0.1 -- BODY: HTML font color is red
   * 0.1 -- BODY: Image tag with an ID code to identify you
   * 2.8 -- BODY: Bayesian classifier says spam probability is 80 to 90%
     [score: 0.8204]
   * 0.8 -- BODY: HTML font color is green
   * 0.3 -- BODY: FONT Size +2 and up or 3 and up
   * 0.1 -- BODY: HTML font color not within safe 6x6x6 palette
   * 0.1 -- BODY: HTML font color is blue
   * 0.3 -- BODY: Message is 70% to 80% HTML
   * 1.2 -- Date: is 6 to 12 hours after Received: date
---- End of SpamAssassin results
Conditional Probability

Recall the definition of conditional probability:

\[ P(X \mid Y) = \frac{P(X \cap Y)}{P(Y)} \]
In general:
\[ P(X \mid Y)P(Y) = P(X \cap Y) = P(Y \mid X)P(X) \]

Thus,
\[ P(X \mid Y) = \frac{P(Y \mid X)P(X)}{P(Y)} \]
Bayes Theorem applied to document classification

Let $C$ be a set of possible class of documents and $d$ be a document.

To classify $d$, calculate $P(c \mid d)$ for all $c \in C$ and return the class $c$ for which $P(c\mid d)$ is maximum.

We can calculate $P(c \mid d)$ using Bayes theorem:

$$P(c \mid d) = \frac{P(d \mid c)P(c)}{P(d)}$$

where $P(c)$ is the prior probability of class $c$, and $P(d)$ is the prior probability of document $d$. 
To classify document $d$, calculate $c_{\text{MAP}}$ (the “maximum a posteriori” class) as follows:

$$P(c \mid d) = \frac{P(d \mid c)P(c)}{P(d)}$$

$$c_{\text{MAP}} = \arg\max_{c \in C} [P(c \mid d)] = \arg\max_{c \in C} \left[ \frac{P(d \mid c)P(c)}{P(d)} \right]$$

$$= \arg\max_{c \in C} [P(d \mid c)P(c)]$$
Representation of documents

Let \( d = (t_1, t_2, t_3, \ldots, t_n) \), where \( t_k \) is a term (from a complete vocabulary) used in document \( d \) (in sequence).

Example (from reading): Suppose the document is “The rain in Spain falls mainly in the plain.”

Then \( d = (\text{rain, spain, fall, mainly, in, plain}) \)

(Here, a pre-processing step deleted punctuation, “stop” words, did “stemming”, and put everything in lower case.)
Naïve Bayes Multinomial Model

(you can ignore the “Bernoulli” model that is also described in the reading)

We have:

\[ c_{MAP} = \arg \max_{c \in C} \left[ P(d \mid c)P(c) \right] \]

For our example, supposed we have \( C = \{\text{Weather, Not Weather}\} \), and suppose we have the following training examples:

<table>
<thead>
<tr>
<th>Document</th>
<th>d</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(in, weather, rain, today, in, portland)</td>
<td>Weather</td>
</tr>
<tr>
<td>2</td>
<td>(yesterday, all, my, trouble, were, far, away)</td>
<td>Not Weather</td>
</tr>
<tr>
<td>3</td>
<td>(yesterday, mainly, rain, fall, fog, spain)</td>
<td>Weather</td>
</tr>
<tr>
<td>4</td>
<td>(today, spain, day, in, portland)</td>
<td>Not Weather</td>
</tr>
</tbody>
</table>

Suppose our vocabulary consisted of 1000 possible terms.

For our test document, we’d calculate

\[ c_{MAP} = \arg \max_{c \in C} \left[ P((rain, spain, fall, mainly, in, plain) \mid c)P(c) \right] \]
Problem:

\[ c_{\text{MAP}} = \arg\max_{c \in \mathcal{C}} [ P((\text{rain}, \text{spain}, \text{fall}, \text{mainly}, \text{in}, \text{plain}) \mid c) P(c) ] \]

We can't calculate \( P((\text{rain}, \text{spain}, \text{fall}, \text{mainly}, \text{in}, \text{plain}) \mid c) \) from the training data!

Naïve Bayes independence assumption:

\[
P((\text{rain}, \text{spain}, \text{fall}, \text{mainly}, \text{in}, \text{plain}) \mid c) \\
= P(\text{rain} \mid c) P(\text{spain} \mid c) P(\text{fall} \mid c) P(\text{mainly} \mid c) P(\text{in} \mid c) P(\text{plain} \mid c)
\]

which we can calculate from the training data! (Well, almost…)

Let \( P(t_k \mid c) = \)

\[
P(t_k \mid c) = \frac{\text{Number of occurrences of term } t_k \text{ in training documents of class } c}{\text{sum of lengths of documents of class } c}
\]
Calculating a model from the training data

Training examples:

<table>
<thead>
<tr>
<th>Document</th>
<th>$d$</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(in, weather, rain, today, in, portland)</td>
<td>Weather</td>
</tr>
<tr>
<td>2</td>
<td>(yesterday, mainly, rain, fall, fog, spain)</td>
<td>Weather</td>
</tr>
<tr>
<td>3</td>
<td>(yesterday, all, my, trouble, were, mainly, far, away)</td>
<td>Not Weather</td>
</tr>
<tr>
<td>4</td>
<td>(today, spain, day, in, portland)</td>
<td>Not Weather</td>
</tr>
</tbody>
</table>

Model from training data:

\[
P(rain \mid Weather) = \frac{2}{12} \quad \text{P}(rain \mid \text{Not Weather}) = \frac{0}{13}
\]
\[
P(spain \mid Weather) = \frac{1}{12} \quad \text{P}(spain \mid \text{Not Weather}) = \frac{1}{13}
\]
\[
P(fall \mid Weather) = \frac{1}{12} \quad \text{P}(fall \mid \text{Not Weather}) = \frac{0}{13}
\]
\[
P(mainly \mid Weather) = \frac{1}{12} \quad \text{P}(mainly \mid \text{Not Weather}) = \frac{1}{13}
\]
\[
P(in \mid Weather) = \frac{1}{12} \quad \text{P}(in \mid \text{Not Weather}) = \frac{1}{13}
\]
\[
P(plain \mid Weather) = 0/12 \quad \text{P}(plain \mid \text{Not Weather}) = 0/13
\]
Predicting the class from the model

Model from training data:

\[ P(rain \mid Weather) = \frac{2}{12} \]
\[ P(spain \mid Weather) = \frac{1}{12} \]
\[ P(fall \mid Weather) = \frac{1}{12} \]
\[ P(mainly \mid Weather) = \frac{1}{12} \]
\[ P(in \mid Weather) = \frac{1}{12} \]
\[ P(plain \mid Weather) = 0 \]

\[ P(rain \mid Not Weather) = 0 \]
\[ P(spain \mid Not Weather) = \frac{1}{13} \]
\[ P(fall \mid Not Weather) = 0 \]
\[ P(mainly \mid Not Weather) = \frac{1}{13} \]
\[ P(in \mid Not Weather) = 1 \]
\[ P(plain \mid Not Weather) = 0 \]

Test document:
\[ d = (\text{rain}, \text{spain}, \text{fall}, \text{mainly}, \text{in}, \text{plain}) \]

\[ c_{MAP} = \arg \max_c [P(c) P(rain \mid c) P(spain \mid c) P(fall \mid c) P(mainly \mid c) P(in \mid c) P(plain \mid c)] \]

Denote \textbf{Weather} as \( W \), \textbf{Not Weather} as \( \overline{W} \)

\[ P(W)P(rain \mid W)P(spain \mid W)P(fall \mid W)P(mainly \mid W)P(in \mid W)P(plain \mid W) \]
\[ = (1/2)(2/12)(1/12)(1/12)(1/12)(1/12)(0) = 0 \]

\[ P(\overline{W})P(rain \mid \overline{W})P(spain \mid \overline{W})P(fall \mid \overline{W})P(mainly \mid \overline{W})P(in \mid \overline{W})P(plain \mid \overline{W}) \]
\[ = (1/2)(0)(1/13)(0)(1/13)(1/13)(0) = 0 \]
Predicting the class from the model

Problem: Get zeros from sparse data

One solution: Add 1 to each count for all words in vocabulary (“Add-1 smoothing)

Let’s assume we have a total vocabulary of 1000 terms

Test document:  
\[d = (\text{rain, spain, fall, mainly, in, plain})\]

\[c_{MAP} = \arg\max_c [P(c)P(\text{rain} \mid c)P(\text{spain} \mid c)P(\text{fall} \mid c)P(\text{mainly} \mid c)P(\text{in} \mid c)P(\text{plain} \mid c)]\]

Denote Weather as W, Not Weather as \(\bar{W}\)

\[P(W)P(\text{rain} \mid W)P(\text{spain} \mid W)P(\text{fall} \mid W)P(\text{mainly} \mid W)P(\text{in} \mid W)P(\text{plain} \mid W)\]
\[= (1/2)(2/12)(1/12)(1/12)(1/12)(1/12)(1/12)(0) = 0\]

\[P(\bar{W})P(\text{rain} \mid \bar{W})P(\text{spain} \mid \bar{W})P(\text{fall} \mid \bar{W})P(\text{mainly} \mid \bar{W})P(\text{in} \mid \bar{W})P(\text{plain} \mid \bar{W})\]
\[= (1/2)(0)(1/13)(0)(1/13)(1/13)(0) = 0\]
Predicting the class from the model

**Model from training data:**

\[
P(\text{rain} \mid W) = \frac{2+1}{(12+1000)} \quad \quad P(\text{rain} \mid \overline{W}) = \frac{0+1}{(13+1000)}
\]

\[
P(\text{spain} \mid W) = \frac{1+1}{(12+1000)} \quad \quad P(\text{spain} \mid \overline{W}) = \frac{1+1}{(13+1000)}
\]

\[
P(\text{fall} \mid W) = \frac{1+1}{(12+1000)} \quad \quad P(\text{fall} \mid \overline{W}) = \frac{0+1}{(13+1000)}
\]

\[
P(\text{mainly} \mid W) = \frac{1+1}{(12+1000)} \quad \quad P(\text{mainly} \mid \overline{W}) = \frac{1+1}{(13+1000)}
\]

\[
P(\text{in} \mid W) = \frac{1+1}{(12+1000)} \quad \quad P(\text{in} \mid \overline{W}) = \frac{1+1}{(13+1000)}
\]

\[
P(\text{plain} \mid W) = \frac{0+1}{(12+1000)} \quad \quad P(\text{plain} \mid \overline{W}) = \frac{0+1}{(13+1000)}
\]
Predicting the class from the model

Model from training data:

\[
P(rain \mid W) = \frac{2+1}{12+1000} \quad P(rain \mid \overline{W}) = \frac{0+1}{13+1000}
\]

\[
P(spain \mid W) = \frac{1+1}{12+1000} \quad P(spain \mid \overline{W}) = \frac{1+1}{13+1000}
\]

\[
P(fall \mid W) = \frac{1+1}{12+1000} \quad P(fall \mid \overline{W}) = \frac{0+1}{13+1000}
\]

\[
P(mainly \mid W) = \frac{1+1}{12+1000} \quad P(mainly \mid \overline{W}) = \frac{1+1}{13+1000}
\]

\[
P(in \mid W) = \frac{1+1}{12+1000} \quad P(in \mid \overline{W}) = \frac{1+1}{13+1000}
\]

\[
P(plain \mid W) = \frac{0+1}{12+1000} \quad P(plain \mid \overline{W}) = \frac{0+1}{13+1000}
\]

\[
P(W)P(rain \mid W)P(spain \mid W)P(fall \mid W)P(mainly \mid W)P(in \mid W)P(plain \mid W)
\]
\[
= (1/2)(3/1012)(2/1012)(2/1012)(2/1012)(2/1012)(1/1012) = 2.23 \times 10^{-17}
\]

\[
P(\overline{W})P(rain \mid \overline{W})P(spain \mid \overline{W})P(fall \mid \overline{W})P(mainly \mid \overline{W})P(in \mid \overline{W})P(plain \mid \overline{W})
\]
\[
\]

So, \( d \) is classified as \textbf{Weather}. 
Predicting the class from the model

Model from training data:

\[
P(\text{rain} \mid W) = \frac{2+1}{12+1000} \quad P(\text{rain} \mid \overline{W}) = \frac{0+1}{13+1000}
\]

\[
P(\text{spain} \mid W) = \frac{1+1}{12+1000} \quad P(\text{spain} \mid \overline{W}) = \frac{1+1}{13+1000}
\]

\[
P(\text{fall} \mid W) = \frac{1+1}{12+1000} \quad P(\text{fall} \mid \overline{W}) = \frac{0+1}{13+1000}
\]

\[
P(\text{mainly} \mid W) = \frac{1+1}{12+1000} \quad P(\text{mainly} \mid \overline{W}) = \frac{1+1}{13+1000}
\]

\[
P(\text{in} \mid W) = \frac{1+1}{12+1000} \quad P(\text{in} \mid \overline{W}) = \frac{1+1}{13+1000}
\]

\[
P(\text{plain} \mid W) = \frac{0+1}{12+1000} \quad P(\text{plain} \mid \overline{W}) = \frac{0+1}{13+1000}
\]

However, another problem: floating-point underflow

Common solution: use logs of values

\[
P(W)P(\text{rain} \mid W)P(\text{spain} \mid W)P(\text{fall} \mid W)P(\text{mainly} \mid W)P(\text{in} \mid W)P(\text{plain} \mid W)
\]

\[
\]

\[
P(\overline{W})P(\text{rain} \mid \overline{W})P(\text{spain} \mid \overline{W})P(\text{fall} \mid \overline{W})P(\text{mainly} \mid \overline{W})P(\text{in} \mid \overline{W})P(\text{plain} \mid \overline{W})
\]

\[
\]

So, \(d\) is classified as \textbf{Weather}.

More generally, here is the formula for the Naïve Bayes Classifier:

\[
\text{class}_{\text{NB}} = \arg\max_{c \in C} P(c) \prod_{i} P(t_i \mid c)
\]

\[
\text{class}_{\text{NB}} = \arg\max_{c \in C} \left[ \log P(c) + \sum_{i} \log P(t_i \mid c) \right]
\]
Summary:
Naïve Bayes classifier for text classification

\[
\text{class}_{NB} = \arg \max_{c \in C} \left[ \log P(c) + \sum_i \log P(t_i | c) \right]
\]

(use with Add-1 Smoothing)
Summary:
Naïve Bayes classifier for text classification

$$class_{NB} = \arg \max_{c \in C} \left[ \log P(c) + \sum \log P(t_i | c) \right]$$

(use with Add-1 Smoothing)

Question 1: Is independence assumption a good assumption?

Question 2: If not, why does Naïve Bayes work so well in practice?