Text Classification using Naive Bayes

Hiroshi Shimodaira*

January-March 2018

Text classification is the task of classifying documents by their content: that is, by the words of which they are comprised. Perhaps the best-known current text classification problem is email *spam filtering*: classifying email messages into spam and non-spam (ham).

7.1 Document models

Text classifiers often don't use any kind of deep representation about language: often a document is represented as a *bag of words*. (A bag is like a set that allows repeating elements.) This is an extremely simple representation: it only knows which words are included in the document (and how many times each word occurs), and throws away the word order!

Consider a document $\mathcal{D}$, whose class is given by C. In the case of email spam filtering there are two classes $C = S$ (spam) and $C = H$ (ham). We classify $\mathcal{D}$ as the class which has the highest posterior probability $P(C\,|\,\mathcal{D})$, which can be re-expressed using Bayes' Theorem:

$$P(C\,|\,\mathcal{D}) = \frac{P(\mathcal{D}\,|\,C)\,P(C)}{P(\mathcal{D})} \propto P(\mathcal{D}\,|\,C)\,P(C)\,. \tag{7.1}$$

We shall look at two probabilistic models of documents, both of which represent documents as a bag of words, using the Naive Bayes assumption. Both models represent documents using feature vectors whose components correspond to word types. If we have a vocabulary V, containing $|V|$ word types, then the feature vector dimension $D = |V|$.

Bernoulli document model: a document is represented by a feature vector with binary elements taking value 1 if the corresponding word is present in the document and 0 if the word is not present.

Multinomial document model: a document is represented by a feature vector with integer elements whose value is the frequency of that word in the document.

Example: Consider the vocabulary:

$$V = \{blue, red, dog, cat, biscuit, apple\}\,.$$

In this case $|V| = D = 6$. Now consider the (short) document "the blue dog ate a blue biscuit". If $\mathbf{d}^B$ is the Bernoulli feature vector for this document, and $\mathbf{d}^M$ is the multinomial feature vector, then we would have:

$$\mathbf{d}^B = (1, 0, 1, 0, 1, 0)^T$$
$$\mathbf{d}^M = (2, 0, 1, 0, 1, 0)^T$$

To classify a document we use Equation (7.1), which requires estimating the likelihoods of the document given the class, $P(\mathcal{D}|C)$ and the class prior probabilities $P(C)$. To estimate the likelihood, $P(\mathcal{D}|C)$, we use the Naive Bayes assumption applied to whichever of the two document models we are using.

7.2 The Bernoulli document model

As mentioned above, in the Bernoulli model a document is represented by a binary vector, which represents a point in the space of words. If we have a vocabulary V containing a set of $|V|$ words, then the t'th element of a document vector corresponds to word w_t in the vocabulary. Let $\mathbf{b}$ be the feature vector for the document $\mathcal{D}$; then the t'th element of $\mathbf{b}$, written b_t, is either 0 or 1 representing the absence or presence of word w_t in the document.

Let $P(w_t|C_k)$ be the probability of word w_t occurring in a document of class k; the probability of w_t not occurring in a document of this class is given by $(1-P(w_t|C_k))$. If we make the naive Bayes assumption, that the probability of each word occurring in the document is independent of the occurrences of the other words, then we can write the document likelihood $P(\mathcal{D}\,|\,C)$ in terms of the individual word likelihoods $P(w_t|C_k)$:

$$P(\mathcal{D}|C_k) \;=\; P(\mathbf{b}|C_k) = \prod_{t=1}^{|V|} \left[b_t\, P(w_t|C_k) + (1-b_t)\,(1-P(w_t|C_k)) \right]\,. \tag{7.2}$$

This product goes over all words in the vocabulary. If word w_t is present, then $b_t = 1$ and the required probability is $P(w_t\,|\,C_k)$; if word w_t is not present, then $b_t = 0$ and the required probability is $1 - P(w_t|C_k)$. We can imagine this as a model for generating document feature vectors of class k, in which the document feature vector is modelled as a collection of $|V|$ weighted coin tosses, the t th having a probability of success equal to $P(w_t|C_k)$.

The *parameters* of the likelihoods are the probabilities of each word given the document class $P(w_t|C_k)$; the model is also parameterised by the prior probabilities, $P(C_k)$. We can learn (estimate) these parameters from a training set of documents labelled with class k. Let $n_k(w_t)$ be the number of documents of class k in which w_t is observed; and let N_k be the total number of documents of that class. Then we can estimate the parameters of the word likelihoods as:[1]

$$\hat{P}(w_t\,|\,C_k) = \frac{n_k(w_t)}{N_k}\,, \tag{7.3}$$

the relative frequency of documents of class k that contain word w_t. If there are N documents in total in the training set, then the prior probability of class k may be estimated as the relative frequency of documents of class k:

$$\hat{P}(C_k) = \frac{N_k}{N}\,. \tag{7.4}$$

Thus given a training set of documents (each labelled with a class), and a set of K classes, we can estimate a Bernoulli text classification model as follows:

[1] We sometimes use $\hat{P}(\,)$ to explicitly denote the estimated probability as opposed to the (theoretical) true one, $P(\,)$.

1. Define the vocabulary V; the number of words in the vocabulary defines the dimension of the feature vectors

2. Count the following in the training set:

   - N the total number of documents
   - N_k the number of documents labelled with class k, for $k=1,\ldots,K$
   - $n_k(w_t)$ the number of documents of class k containing word w_t for $k=1,\ldots,K, t=1,\ldots,|V|$

3. Estimate the likelihoods $P(w_t \mid C_k)$ using Equation (7.3)

4. Estimate the priors $P(C_k)$ using Equation (7.4)

To classify an unlabelled document $\mathcal{D}$, we estimate the posterior probability for each class k combining Equation (7.1) and Equation (7.2):

$$\begin{aligned} P(C_k\,|\,\mathbf{b}) &\propto P(\mathbf{b}\,|\,C_k)\,P(C_k) \\ &\propto P(C_k)\prod_{t=1}^{|V|}\left[b_t\,P(w_t|C_k) + (1-b_t)\,(1-P(w_t|C_k))\right]. \end{aligned} \tag{7.5}$$

Example

Consider a set of documents, each of which is related either to *Sports* (S) or to *Informatics* (I). Given a training set of 11 documents, we would like to estimate a Naive Bayes classifier, using the Bernoulli document model, to classify unlabelled documents as S or I.

We define a vocabulary of eight words:

$$V = \begin{bmatrix} w_1 = \text{goal}, \\ w_2 = \text{tutor}, \\ w_3 = \text{variance}, \\ w_4 = \text{speed}, \\ w_5 = \text{drink}, \\ w_6 = \text{defence}, \\ w_7 = \text{performance}, \\ w_8 = \text{field} \end{bmatrix}$$

Thus each document is represented as an 8-dimensional binary vector.

The training data is presented below as a matrix for each class, in which each row represents an 8-dimensional document vector

$$\mathbf{B}^{\text{Sport}} = \begin{pmatrix} 1&0&0&0&1&1&1&1 \\ 0&0&1&0&1&1&0&0 \\ 0&1&0&1&0&1&1&0 \\ 1&0&0&1&0&1&0&1 \\ 1&0&0&0&1&0&1&1 \\ 0&0&1&1&0&0&1&1 \end{pmatrix}$$

$$\mathbf{B}^{\text{Inf}} = \begin{pmatrix} 0&1&1&0&0&0&1&0 \\ 1&1&0&1&0&0&1&1 \\ 0&1&1&0&0&1&0&0 \\ 0&0&0&0&0&0&0&0 \\ 0&0&1&0&1&0&1&0 \end{pmatrix}$$

Classify the following into Sports or Informatics using a Naive Bayes classifier.

1. $\mathbf{b}_1 = (1,0,0,1,1,1,0,1)^T$
2. $\mathbf{b}_2 = (0,1,1,0,1,0,1,0)^T$

Solution:

The total number of documents in the training set $N=11$; $N_S=6$, $N_I=5$.

Using (7.4), we can estimate the prior probabilities from the training data as:

$$\hat{P}(S) = \frac{6}{11}; \qquad \hat{P}(I) = \frac{5}{11}$$

The document count $n_k(w)$ in the training data set, and estimated word likelihoods using (7.3) are given as:

	$n_S(w)$	$\hat{P}(w\|S)$	$n_I(w)$	$\hat{P}(w\|I)$
w_1	3	3/6	1	1/5
w_2	1	1/6	3	4/5
w_3	2	2/6	3	3/5
w_4	3	3/6	1	1/5
w_5	3	3/6	1	1/5
w_6	4	4/6	1	2/5
w_7	4	4/6	3	3/5
w_8	4	4/6	1	1/5

We use (7.5) to compute the posterior probabilities of the two test vectors and hence classify them.

1. $\mathbf{b}_1 = (1,0,0,1,1,1,0,1)^T$

$$\begin{aligned} \hat{P}(S\,|\,\mathbf{b}_1) &\propto \hat{P}(S)\prod_{t=1}^{8}\left[b_{1t}\hat{P}(w_t\,|\,S) + (1-b_{1t})(1-\hat{P}(w_t\,|\,S))\right] \\ &\propto \frac{6}{11}\left(\frac{1}{2}\times\frac{5}{6}\times\frac{2}{3}\times\frac{1}{2}\times\frac{1}{2}\times\frac{2}{3}\times\frac{1}{3}\times\frac{2}{3}\right) = \frac{5}{891} \approx 5.6\times10^{-3} \\ \hat{P}(I\,|\,\mathbf{b}_1) &\propto \hat{P}(I)\prod_{t=1}^{8}\left[b_{1t}\hat{P}(w_t\,|\,I) + (1-b_{1t})(1-\hat{P}(w_t\,|\,I))\right] \\ &\propto \frac{5}{11}\left(\frac{1}{5}\times\frac{2}{5}\times\frac{2}{5}\times\frac{1}{5}\times\frac{1}{5}\times\frac{1}{5}\times\frac{2}{5}\times\frac{1}{5}\right) = \frac{8}{859375} \approx 9.3\times10^{-6} \end{aligned}$$

Classify this document as S.

2. $\mathbf{b}_2 = (0, 1, 1, 0, 1, 0, 1, 0)^T$

$$\hat{P}(S\,|\,\mathbf{b}_2) \propto \hat{P}(S)\prod_{t=1}^{8}\left[b_{2t}\hat{P}(w_t\,|\,S) + (1-b_{2t})(1-\hat{P}(w_t\,|\,S))\right]$$
$$\propto \frac{6}{11}\left(\frac{1}{2}\times\frac{1}{6}\times\frac{1}{3}\times\frac{1}{2}\times\frac{1}{2}\times\frac{1}{3}\times\frac{2}{3}\times\frac{1}{3}\right) = \frac{12}{42768} \approx 2.8\times 10^{-4}$$

$$\hat{P}(I\,|\,\mathbf{b}_1) \propto \hat{P}(I)\prod_{t=1}^{8}\left[b_{1t}\hat{P}(w_t\,|\,I) + (1-b_{1t})(1-\hat{P}(w_t\,|\,I))\right]$$
$$\propto \frac{5}{11}\left(\frac{4}{5}\times\frac{3}{5}\times\frac{3}{5}\times\frac{4}{5}\times\frac{1}{5}\times\frac{4}{5}\times\frac{3}{5}\times\frac{4}{5}\right) = \frac{34560}{4296875} \approx 8.0\times 10^{-3}$$

Classify as I.

7.3 The multinomial distribution *(NE)*

Before discussing the multinomial document model, it is important to be familiar with the multinomial distribution.

We first need to be able to count the number of distinct arrangements of a set of items, when some of the items are *indistinguishable*. For example: Using all the letters, how many distinct sequences can you make from the word "Mississippi"? There are 11 letters to permute, but "i" and "s" occur four times and "p" twice. If these letters were distinct (e.g., if they were labelled i_1, i_2, etc.) then there would be 11! permutations. However of these permutations there are 4! that are the same if the subscripts are removed from the "i"s. This means that we can reduce the size of the total sample space by a factor of 4! to take account of the four occurrences of "i". Likewise there is a factor of 4! for "s" and a factor of 2! for "p" (and a factor of 1! for "m"). This gives the total number distinct permutations as:

$$\frac{11!}{4!\,4!\,2!\,1!} = 34650$$

Generally if we have n items of D types, with n_1 of type 1, n_2 of type 2 and n_D of type D (such that $n_1 + n_2 + \ldots + n_D = n$), then the number of distinct permutations is given by:

$$\frac{n!}{n_1!\,n_2!\ldots n_D!}$$

These numbers are called the *multinomial coefficients*.

Now suppose a population contains items of $D \geq 2$ different types and that the proportion of items that are of type t is p_t ($t = 1, \ldots, D$), with

$$\sum_{t=1}^{D} p_t = 1 \qquad p_t > 0,\ \text{for all } t.$$

Suppose n items are drawn at random (with replacement) and let x_t denote the number of items of type t. The vector $\mathbf{x} = (x_1, \ldots, x_D)^T$ has a *multinomial distribution* with parameters n and $p_1, \ldots, p_D$, defined by:

$$P(\mathbf{x}) = \frac{n!}{x_1!\,x_2!\ldots x_D!}\,p_1^{x_1}p_2^{x_2}\ldots p_D^{x_D}$$
$$= \frac{n!}{\prod_{t=1}^{D} x_t!}\prod_{t=1}^{D} p_t^{x_t}\,. \tag{7.6}$$

The $\prod_{t=1}^{D} p_t^{x_t}$ product gives the probability of one sequence of outcomes with counts $\mathbf{x}$. The multinomial coefficient, counts the number of such sequences that there are.

7.4 The multinomial document model

In the multinomial document model, the document feature vectors capture the frequency of words, not just their presence or absence. Let $\mathbf{x}$ be the multinomial model feature vector for document $\mathcal{D}$. The t'th element of $\mathbf{x}$, written x_t, is the count of the number of times word w_t occurs in document $\mathcal{D}$. Let $n = \sum_t x_t$ be the total number of words in document $\mathcal{D}$.

Let $P(w_t|C_k)$ again be the probability of word w_t occurring in class k, this time estimated using the word frequency information from the document feature vectors. We again make the naive Bayes assumption, that the probability of each word occurring in the document is independent of the occurrences of the other words. We can then write the document likelihood $P(\mathcal{D}|C_k)$ as a multinomial distribution (Equation7.6), where the number of draws corresponds to the length of the document, and the proportion of drawing item t is the probability of word type t occurring in a document of class k, $P(w_t|C_k)$.

$$P(\mathcal{D}|C) = P(\mathbf{x}|C_k) = \frac{n!}{\prod_{t=1}^{|V|} x_t!}\prod_{t=1}^{|V|} P(w_t|C_k)^{x_t}$$
$$\propto \prod_{t=1}^{|V|} P(w_t|C_k)^{x_t}\,. \tag{7.7}$$

We often won't need the normalisation term $(n!/\prod_t x_t!)$, because it does not depend on the class k. The numerator of the right hand side of this expression can be interpreted as the product of word likelihoods for each word in the document, with repeated words taking part for each repetition.

As for the multinomial model, the parameters of the likelihood are the probabilities of each word given the document class $P(w_t|C_k)$, and the model parameters also include the prior probabilities $P(C_k)$. To estimate these parameters from a training set of documents, $\{\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_N\}$, let z_{ik} be an indicator variable which equals 1 when $\mathcal{D}_i$ belongs to class k, and equals 0 otherwise. If N is again the total number of documents, then we have:

$$\hat{P}(w_t \mid C_k) = \frac{\sum_{i=1}^{N} x_{it}\,z_{ik}}{\sum_{s=1}^{|V|}\sum_{i=1}^{N} x_{is}\,z_{ik}} \tag{7.8}$$
$$= \frac{n_k(w_t)}{\sum_{s=1}^{|V|} n_k(w_s)}\,, \tag{7.9}$$

an estimate of the probability $P(w_t \mid C_k)$ as the relative frequency of w_t in documents of class k with respect to the total number of words in documents of that class, where $n_k(w_t) = \sum_{i=1}^{N} x_{it}\,z_{ik}$.[2]

The prior probability of class k is estimated as before (Equation (7.4).

Thus given a training set of documents (each labelled with a class) and a set of K classes, we can estimate a multinomial text classification model as follows:

1. Define the vocabulary V; the number of words in the vocabulary defines the dimension of the feature vectors.

[2]We previously used the same variable $n_k(w_t)$ in Equation (7.3) for Bernoulli model to denote the number of documents of class k containing word w_t, whereas it now represents the frequency of word w_t in the documents of class k. You should now understand Equation (7.9) is more general than Equation (7.3). The actual difference lies in the fact that the two document models employ different types of feature vector - Bernoulli feature vector and multinomial feature vector.

2. Count the following in the training set:
   - N : the total number of documents
   - N_k : the number of documents labelled with class k, for each class $k = 1, \dots, K$
   - x_{it} : the frequency of word w_t in document $\mathcal{D}_i$, computed for every word w_t in V (Alternatively, $n_k(w_t)$: the frequency of word w_t in the documents of class k)
3. Estimate the likelihoods $P(w_t \mid C_k)$ using (7.8) or (7.9).
4. Estimate the priors $P(C_k)$ using (7.4).

To classify an unlabelled document $\mathcal{D}$, we estimate the posterior probability for each class combining (7.1) and (7.7):

$$\begin{aligned} P(C_k \mid \mathcal{D}) &= P(C_k \mid \mathbf{x}) \\ &\propto P(\mathbf{x} \mid C_k)\, P(C_k) \\ &\propto P(C_k) \prod_{t=1}^{|V|} P(w_t \mid C_k)^{x_t} . \end{aligned} \tag{7.10}$$

Unlike the Bernoulli model, words that do not occur in the document (i.e., for which $x_t = 0$) do not affect the probability (since $p^0 = 1$).

Note that we can rewrite the posterior probability in terms of words u which occur in the document:

$$P(C_k \mid \mathcal{D}) \propto P(C_k) \prod_{j=1}^{\mathrm{len}(\mathcal{D})} P(u_j \mid C_k) \tag{7.11}$$

where u_j is the j'th word in document $\mathcal{D}$. This form will be preferable to the original one in some applications, which you will see in the following example.

Example

Consider the same example in Section 7.2, except that this time we will use the multinomial model instead of Bernoulli model.

Each document $\mathcal{D}_i$ is now represented as an 8-dimensional row vector $\mathbf{m}_i$; element m_{it} is the frequency of word w_t in document $\mathcal{D}_i$.

The training data is presented below as a matrix for each class, in which each row represents an 8-dimensional document vector.

$$\mathbf{M}^{\text{Sport}} = \begin{pmatrix} 2 & 0 & 0 & 0 & 1 & 2 & 3 & 1 \\ 0 & 0 & 1 & 0 & 2 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 2 & 1 & 0 \\ 1 & 0 & 0 & 2 & 0 & 1 & 0 & 1 \\ 2 & 0 & 0 & 0 & 1 & 0 & 1 & 3 \\ 0 & 0 & 1 & 2 & 0 & 0 & 2 & 1 \end{pmatrix}$$

$$\mathbf{M}^{\text{Inf}} = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 2 & 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \end{pmatrix}$$

Classify the following test documents into Sports or Informatics.

1. $\mathcal{D}_1 = w_5\ w_1\ w_6\ w_8\ w_1\ w_2\ w_6$
2. $\mathcal{D}_2 = w_3\ w_5\ w_2\ w_7$

Solution:

Let $n_k(w)$ be the frequency of word w in all documents of class k, giving likelihood estimate,

$$\hat{P}(w \mid S) = \frac{n_S(w)}{\sum_{v \in V} n_S(v)}, \quad \hat{P}(w \mid I) = \frac{n_I(w)}{\sum_{v \in V} n_I(v)},$$

	$n_S(w)$	$\hat{P}(w \mid S)$	$n_I(w)$	$\hat{P}(w \mid I)$
w_1	5	5/36	1	1/16
w_2	1	1/36	4	4/16
w_3	2	2/36	3	3/16
w_4	5	5/36	1	1/16
w_5	4	4/36	1	1/16
w_6	6	6/36	2	2/16
w_7	7	7/36	3	3/16
w_8	6	6/36	1	1/16

We have now estimated the model parameters.

1. $\mathcal{D}_1 = w_5\ w_1\ w_6\ w_8\ w_1\ w_2\ w_6$

$$\begin{aligned} \hat{P}(\mathcal{D}_1 \mid S) &= \hat{P}(w_5 \mid S) \cdot \hat{P}(w_1 \mid S) \cdot \hat{P}(w_6 \mid S) \cdot \hat{P}(w_8 \mid S) \cdot \hat{P}(w_1 \mid S) \cdot \hat{P}(w_2 \mid S) \cdot \hat{P}(w_6 \mid S) \\ &= \frac{4}{36} \times \frac{5}{36} \times \frac{6}{36} \times \frac{6}{36} \times \frac{5}{36} \times \frac{1}{36} \times \frac{6}{36} = \frac{4 \times 5^2 \times 6^3}{36^7} \approx 2.76 \times 10^{-7} \end{aligned}$$

$$\hat{P}(S \mid \mathcal{D}_1) \propto \hat{P}(S)\, \hat{P}(\mathcal{D}_1 \mid S) = \frac{6}{11} \cdot \frac{4 \times 5^2 \times 6^3}{36^7} \approx 1.50 \times 10^{-7}$$

$$\begin{aligned} \hat{P}(\mathcal{D}_1 \mid I) &= \hat{P}(w_5 \mid I) \cdot \hat{P}(w_1 \mid I) \cdot \hat{P}(w_6 \mid I) \cdot \hat{P}(w_8 \mid I) \cdot \hat{P}(w_1 \mid I) \cdot \hat{P}(w_2 \mid I) \cdot \hat{P}(w_6 \mid I) \\ &= \frac{1}{16} \times \frac{1}{16} \times \frac{2}{16} \times \frac{1}{16} \times \frac{1}{16} \times \frac{4}{16} \times \frac{2}{16} = \frac{2^4}{16^7} \approx 5.96 \times 10^{-7} \end{aligned}$$

$$\hat{P}(I \mid \mathcal{D}_1) \propto \hat{P}(I)\, \hat{P}(\mathcal{D}_1 \mid I) = \frac{5}{11} \cdot \frac{2^4}{16^6} \approx 2.71 \times 10^{-8}$$

$\hat{P}(S \mid \mathcal{D}_1) > \hat{P}(I \mid \mathcal{D}_1)$, thus we classify $\mathcal{D}_1$ as S.

We have not normalised by $\hat{P}(\mathcal{D}_1)$, hence the above are joint probabilities, proportional to the posterior probability. To obtain the posterior:

$$\hat{P}(S \mid \mathcal{D}_1) = \frac{\hat{P}(S)\, \hat{P}(\mathcal{D}_1 \mid S)}{\hat{P}(S)\, \hat{P}(\mathcal{D}_1 \mid S) + \hat{P}(I)\, \hat{P}(\mathcal{D}_1 \mid I)} = \frac{1.50 \times 10^{-7}}{1.50 \times 10^{-7} + 2.71 \times 10^{-8}} \approx 0.847$$

$$\hat{P}(I \mid \mathcal{D}_1) = 1 - \hat{P}(S \mid \mathcal{D}_1) \approx 0.153$$

2. $\mathcal{D}_2 = w_3\ w_5\ w_2\ w_7$

$$\hat{P}(\mathcal{D}_2 \,|\, S) = \hat{P}(w_3 \,|\, S) \cdot \hat{P}(w_5 \,|\, S) \cdot \hat{P}(w_2 \,|\, S) \cdot \hat{P}(w_7 \,|\, S)$$
$$= \frac{2}{36} \times \frac{4}{36} \times \frac{1}{36} \times \frac{7}{36} = \frac{2^3 \times 7}{36^4} \approx 3.33 \times 10^{-5}$$
$$\hat{P}(S \,|\, \mathcal{D}_2) \propto \hat{P}(S)\, \hat{P}(\mathcal{D}_2 \,|\, S) = \frac{6}{11} \cdot \frac{2^3 \cdot 7}{36^4} \approx 1.82 \times 10^{-5}$$

$$\hat{P}(\mathcal{D}_2 \,|\, I) = \hat{P}(w_3 \,|\, I) \cdot \hat{P}(w_5 \,|\, I) \cdot \hat{P}(w_2 \,|\, I) \cdot \hat{P}(w_7 \,|\, I)$$
$$= \frac{3}{16} \times \frac{1}{16} \times \frac{4}{16} \times \frac{3}{16} = \frac{3^2 \cdot 4}{16^4} \approx 5.49 \times 10^{-4}$$
$$\hat{P}(I \,|\, \mathcal{D}_2) \propto \hat{P}(I)\, \hat{P}(\mathcal{D}_2 \,|\, I) = \frac{5}{11} \cdot \frac{3^2 \cdot 4}{16^4} \approx 2.50 \times 10^{-4}$$

$\hat{P}(I \,|\, \mathcal{D}_2) > \hat{P}(S \,|\, \mathcal{D}_2)$, thus we classify $\mathcal{D}_2$ as I.

We have not normalised by $\hat{P}(\mathcal{D}_2)$, hence the above are joint probabilities, proportional to the posterior probability. To obtain the posterior:

$$\hat{P}(S \,|\, \mathcal{D}_2) = \frac{\hat{P}(S)\, \hat{P}(\mathcal{D}_2 \,|\, S)}{\hat{P}(S)\, \hat{P}(\mathcal{D}_2 \,|\, S) + \hat{P}(I)\, \hat{P}(\mathcal{D}_2 \,|\, I)} = \frac{1.82 \times 10^{-5}}{1.82 \times 10^{-5} + 2.50 \times 10^{-4}} \approx 0.0679$$
$$\hat{P}(I \,|\, \mathcal{D}_2) = 1 - \hat{P}(S \,|\, \mathcal{D}_2) \approx 0.932$$

7.5 The Zero Probability Problem

A drawback of relative frequency estimates—Equation (7.8) for the multinomial model—is that zero counts result in estimates of zero probability. This is a bad thing because the Naive Bayes equation for the likelihood (7.7) involves taking a product of probabilities: if any one of the terms of the product is zero, then the whole product is zero. This means that the probability of the document belonging to the class in question is zero—which means it is impossible.

Just because a word does not occur in a document class in the training data does not mean that it cannot occur in any document of that class.

The problem is that Equation (7.8) *underestimates* the likelihoods of words that do not occur in the data. Even if word w is not observed for class k in the training set, we would still like $P(w \mid C_k) > 0$. Since probabilities must sum to 1, if unobserved words have underestimated probabilities, then those words that are observed must have overestimated probabilities. Therefore, one way to alleviate the problem is to remove a small amount of probability allocated to observed events and distribute this across the unobserved events. A simple way to do this, sometimes called *Laplace's law of succession* or *add-one smoothing*, adds a count of one to each word type. If there are W word types in total, then Equation (7.8) may be replaced with:

$$P_{\text{Lap}}(w_t \mid C_k) = \frac{1 + \sum_{i=1}^{N} x_{it}\, z_{ik}}{|V| + \sum_{s=1}^{|V|} \sum_{i=1}^{N} x_{is}\, z_{ik}} = \frac{1 + n_k(w_t)}{|V| + \sum_{s=1}^{|V|} n_k(w_s)} \tag{7.12}$$

The denominator was increased to take account of the $|V|$ extra "observations" arising from the "add 1" term, ensuring that the probabilities are still normalised.

7.6 Comparing the two models

The Bernoulli and the multinomial document models are both based on a bag of words. However there are a number of differences, which we summarise here:

1. **Underlying model of text:**
   Bernoulli: a document can be thought of as being generated from a multidimensional Bernoulli distribution: the probability of a word being present can be thought of as a (weighted) coin flip with probability $P(w_t | C)$.
   Multinomial: a document is formed by drawing words from a multinomial distribution: you can think of obtaining the next word in the document by rolling a (weighted) $|V|$-sided dice with probabilities $P(w_t | C)$.

2. **Document representation:**
   Bernoulli: binary vector, elements indicating presence or absence of a word.
   Multinomial: integer vector, elements indicating frequency of occurrence of a word.

3. **Multiple occurrences of words:**
   Bernoulli: ignored.
   Multinomial: taken into account.

4. **Behaviour with document length:**
   Bernoulli: best for short documents.
   Multinomial: longer documents are OK.

5. **Behaviour with "the":**
   Bernoulli: since "the" is present in almost every document, $P(\text{"the"} | C) \approx 1.0$.
   Multinomial: since probabilities are based on relative frequencies of word occurrence in a class, $P(\text{"the"} | C) \approx 0.05$.

7.7 Conclusion

In this chapter we have shown how the Naive Bayes approximation can be used for document classification, by constructing distributions over words. The classifiers require a *document model* to estimate $P(\text{document} \mid \text{class})$. We looked at two document models that we can use with the Naive Bayes approximation:

- **Bernoulli document model:** a document is represented by a binary feature vector, whose elements indicate absence or presence of corresponding word in the document.

- **Multinomial document model:** a document is represented by an integer feature vector, whose elements indicate frequency of corresponding word in the document.

Reference: Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, *Introduction to Information Retrieval*, Cambridge University Press. 2008. (Chapter 13 Text classification and Naive Bayes) `http://www-nlp.stanford.edu/IR-book/`

Exercises

1. Considering the same example in Section 7.4, classify a new test document $\mathcal{D} = w_7, w_3, w_2, w_3, w_4$ when the multinomial document model is used. What if the Bernoulli document model is employed instead?

2. The Bernoulli document model can also suffer from the zero probability problem discussed in Section 7.5. How would you apply add-one smoothing in this case?