Math 345 - Notes
Naive Bayes II
October 3, 2018

Naive Bayes (cont.)

Multiple word spam filter

Suppose we check for N words: $w_1, w_2, \ldots, w_N$. We define the 0-1 random variables $X_i = \mathbb{1}\{\text{message has word } w_i\}$, that is, X_i equals 1 if w_i is in the message, and it equals 0 otherwise. Suppose $X_1 = a_1, X_2 = a_2, \cdots, X_N = a_N$, where a_i are either 0 or 1. We assume each word appears in a message independent of the other words on the list, that is:

$$P(X_1 = a_1, X_2 = a_2, \cdots, X_N = a_N \,|\, \text{spam}) = P(X_1 = a_1 \,|\, \text{spam})P(X_2 = a_2 \,|\, \text{spam}) \cdots P(X_N = a_N \,|\, \text{spam})$$

Similarly, we use independence on the set of *ham* messages. Note that independence is not a very reasonable assumption, since we know certain *spam* words like to appear together, such as *prince, rich, fortune, Nigeria* etc. This is why the model is called the **Naive** Bayes model. However, it is quite efficient.

Example: Consider the example from last time:

	spam	*ham*
	1500	3672
meeting	16	153
pharmacy	621	0
money	125	31
Digipen	0	1892

Using the four words above ($N = 4$), let $w_1 =$ *meeting*, $w_2 =$ *pharmacy*, $w_3 =$ *money*, and $w_4 =$ *DigiPen*. Suppose the email message has the words w_1, w_3, and w_3, but not the word w_2. Should we classify it as spam?
Let us use smoothing, with smoothing parameters $(\alpha, \beta) = (1, 2)$, and use "s" and shorthand for spam and "h" for ham:

$$
\begin{aligned}
& P(\text{spam} \,|\, X_1 = 1, X_2 = 0, X_3 = 1, X_4 = 1) \\
= \; & \frac{P(X_1 = 1, X_2 = 0, X_3 = 1, X_4 = 1 \,|\, \text{s})P(\text{s})}{P(X_1 = 1, X_2 = 0, X_3 = 1, X_4 = 1 \,|\, \text{s})P(\text{s}) + P(X_1 = 1, X_2 = 0, X_3 = 1, X_4 = 1 \,|\, \text{h})P(\text{h})} \\
= \; & \frac{P(X_1 = 1|\text{s})P(X_2 = 0|\text{s})P(X_3 = 1|\text{s})P(X_4 = 1|\text{s})P(\text{s})}{P(X_1 = 1|\text{s})P(X_2 = 0|\text{s})P(X_3 = 1|\text{s})P(X_4 = 1|\text{s})P(\text{s}) + P(X_1 = 1|\text{h})P(X_2 = 0|\text{h})P(X_3 = 1|\text{h})P(X_4 = 1|\text{h})P(\text{h})} \\
= \; & \frac{(17/1502)(1 - 622/1502)(126/1502)(1/1502)(0.29)}{(17/1502)(1 - 622/1502)(126/1502)(1/1502)(0.29) + (154/3674)(1 - 1/3674)(32/3674)(1893/3674)(0.71)} \\
= \; & 0.00080
\end{aligned}
$$

We classify this email message as **ham**.

Exercise: try various combinations for the four words. For example,

$$P(\text{spam} \,|\, X_1 = 1, X_2 = 0, X_3 = 1, X_4 = 0) = 0.56.$$

Testing the model

We can use the following metrics to test the model:

$$\text{accuracy} = \frac{\text{correct predictions}}{\text{total}} = \frac{\text{spam predicted spam, ham predicted ham}}{\text{total messages}}$$

$$\text{precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}} = \frac{\text{spam predicted spam}}{\text{predicted spam}}$$

$$\text{recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}} = \frac{\text{spam predicted spam}}{\text{total spam}}$$

Consider the example below:

	spam	ham
predict spam	101	33
predict ham	38	704

The three evaluation metrics we can use give:

(a) $\text{accuracy} = \dfrac{101 + 704}{101 + 33 + 38 + 704} = .9189$

(b) $\text{precision} = \dfrac{101}{101 + 33} = .7537$

(c) $\text{recall} = \dfrac{101}{101 + 38} = .7266$

Compact formulation of model

We continue our discussion of Naive Bayes, by *pre-computing* some of the parameters for the model, based on the training data. Suppose our spam filter keeps track of N different words. When a new email message arrives, it is encoded by a vector $\vec{a} = [a_1, a_2, \ldots, a_N]$ with 1's for the words that appear in the message and 0's for those that do not appear. For example, if the word w_k appears in the message, then $a_k = 1$.

Remark: We will use the following facts and notation in our derivations:

- the notation $\exp\{x\} = e^x$, for an easier way to display the expressions,
- the summation notation $\displaystyle\sum_{k=1}^{n} c_k = c_1 + c_2 + \cdots + c_n$,
- the product notation $\displaystyle\prod_{k=1}^{n} c_k = c_1 \times c_2 \times \cdots \times c_n$,
- the property that exponentials and logs are inverses of each other: $x = e^{\log(x)}$,
- the property of logs: $\log(a \cdot b) = \log(a) + \log(b)$,
- the property of logs: $\log(a^b) = b \log(a)$.

Remark that for a large N set of words to be tested, we need to multiply many small probabilities, so we might run into underflow problems. To avoid it, we can work with logarithms instead:

$$\begin{aligned} P(X_1 = a_1, \cdots, X_N = a_N \,|\, \text{spam}) &= \prod_{k=1}^{N} P(X_k = a_k \,|\, \text{spam}) \\ &= \exp\{\log(P(X_1 = a_1 \,|\, \text{spam}) \times \cdots \times P(X_N = a_N \,|\, \text{spam}))\} \\ &= \exp\{\log(P(X_1 = a_1 \,|\, \text{spam})) + \cdots + \log(P(X_N = a_N \,|\, \text{spam}))\} \\ &= \exp\left\{\sum_{k=1}^{N} \log(P(X_k = a_k \,|\, \text{spam}))\right\}, \end{aligned}$$

For a more compact way to write these probabilities, we let

$$p_{ks} = P(X_k = 1|\text{spam}), \qquad p_{kh} = P(X_k = 1|\text{ham})$$

be the probabilities that w_k appears as a spam message, or a ham message respectively. Thus,

$$P(X_k = a_k|\text{spam}) = P(X_k = 1|\text{spam})^{a_k}[1 - P(X_k = 1|\text{spam})]^{1-a_k} = p_{ks}^{a_k}(1 - p_{ks})^{1-a_k}.$$

$$P(X_k = a_k|\text{ham}) = P(X_k = 1|\text{ham})^{a_k}[1 - P(X_k = 1|\text{ham})]^{1-a_k} = p_{kh}^{a_k}(1 - p_{kh})^{1-a_k}.$$

Combining into the logarithm notation:

$$\begin{aligned} P(X_1 = a_1, \cdots, X_N = a_N \,|\, \text{spam}) &= \exp\left\{\sum_{k=1}^{N} \log\left[p_{ks}^{a_k}(1 - p_{ks})^{1-a_k}\right]\right\} \\ &= \exp\left\{\sum_{k=1}^{N} [a_k \log(p_{ks}) + (1 - a_k)\log(1 - p_{ks})]\right\} \\ &= \exp\left\{\sum_{k=1}^{N} \left[a_k \log\left(\frac{p_{ks}}{1 - p_{ks}}\right)\right] + \sum_{k=1}^{N} \log(1 - p_{ks})\right\}. \end{aligned}$$

Note that the second sum depends only on the training data, and not on the message to be tested. Let $y_0 = \sum_{k=1}^{N} \log(1 - p_{ks})$ and $y_k = \log\left(\frac{p_{ks}}{1 - p_{ks}}\right)$. Using vector notation, we set $\vec{X} = [X_1, \ldots, X_N]$, $\vec{a} = [a_1, \ldots, a_N]$ and $\vec{y} = [y_1, \ldots, y_N]$:

$$P(\vec{X} = \vec{a} \,|\, \text{spam}) = \exp\{\vec{a} \cdot \vec{y} + y_0\}.$$

Similarly, for the set *ham*,

$$P(\vec{X} = \vec{a} \,|\, \text{ham}) = \exp\{\vec{a} \cdot \vec{z} + z_0\},$$

where $z_0 = \sum_{k=1}^{N} \log(1 - p_{kh})$ and $\vec{z} = [z_1, z_2, \ldots, z_N]$ with $z_k = \log\left(\frac{p_{kh}}{1 - p_{kh}}\right)$.

Now that we have y_0, $\vec{y}$, z_0, and $\vec{z}$ pre-computed for our model, we simply use dot products to find probabilities and classify messages:

$$P(\text{spam}|\vec{X} = \vec{a}) = \frac{\exp\{\vec{y} \cdot \vec{a} + y_0\}P(\text{spam})}{\exp\{\vec{y} \cdot \vec{a} + y_0\}P(\text{spam}) + \exp\{\vec{z} \cdot \vec{a} + z_0\}P(\text{ham})}$$

Example: using the 4 words in the example from the previous section, we have

	spam	ham
	1500	3672
meeting	16	153
pharmacy	621	0
money	125	31
DigiPen	0	1892

Using smoothing with $\alpha = 1$ and $\beta = 2$, and $P(\text{spam}) = .29$ and $P(\text{ham}) = .71$. The pre-computed parameters are:

$$\begin{aligned}
\vec{y} &= \left[\log\left(\frac{\frac{17}{1502}}{1-\frac{17}{1502}}\right), \log\left(\frac{\frac{622}{1502}}{1-\frac{622}{1502}}\right), \log\left(\frac{\frac{126}{1502}}{1-\frac{126}{1502}}\right), \log\left(\frac{\frac{1}{1502}}{1-\frac{1}{1502}}\right)\right] = [-4.47, -0.35, -2.39, -7.31] \\
y_0 &= \log\left(1-\frac{17}{1502}\right) + \log\left(1-\frac{622}{1502}\right) + \log\left(1-\frac{126}{1502}\right) + \log\left(1-\frac{1}{1502}\right) = -0.63 \\
\vec{z} &= \left[\log\left(\frac{\frac{154}{3674}}{1-\frac{154}{3674}}\right), \log\left(\frac{\frac{1}{3674}}{1-\frac{1}{3674}}\right), \log\left(\frac{\frac{32}{3674}}{1-\frac{32}{3674}}\right), \log\left(\frac{\frac{1893}{3674}}{1-\frac{1893}{3674}}\right)\right] = [-3.13, -8.21, -4.73, -0.06] \\
z_0 &= \log\left(1-\frac{154}{3674}\right) + \log\left(1-\frac{1}{3674}\right) + \log\left(1-\frac{32}{3674}\right) + \log\left(1-\frac{1893}{3674}\right) = -0.776
\end{aligned}$$

Then, for $\vec{X} = [1,0,1,1]$, we compute the probability of the message being spam.

$$P(\vec{X} = [1,0,1,1] \,|\, \text{spam}) = \exp\{[-4.47, -0.35, -2.39, -7.31] \cdot [1,0,1,1] + (-0.63)\} = e^{-14.8} = 3.7 \times 10^{-7}.$$

$$P(\vec{X} = [1,0,1,1] \,|\, \text{ham}) = \exp\{[-3.13, -8.21, -4.73, -0.06] \cdot [1,0,1,1] + (-0.776)\} = e^{-8.696} = 1.67 \times 10^{-4}.$$

$$\begin{aligned}
P(\text{spam}|\vec{X} = [1,0,1,1]) &= \frac{P(\vec{X} = [1,0,1,1]|\text{spam})P(\text{spam})}{P(\vec{X} = [1,0,1,1]|\text{spam})P(\text{spam}) + P(\vec{X} = [1,0,1,1]|\text{ham})P(\text{ham})} \\
&= \frac{(3.7 \times 10^{-7})(.29)}{(3.7 \times 10^{-7})(.29) + (1.67 \times 10^{-4})(.71)} \\
&= .00090
\end{aligned}$$

Classifying new messages is now easy:

$$P(\text{spam}|\vec{X} = [1,0,1,0]) = \frac{\exp\{\vec{y} \cdot [1,0,1,0] + y_0\}(.29)}{\exp\{\vec{y} \cdot [1,0,1,0] + y_0\}(.29) + \exp\{\vec{z} \cdot [1,0,1,0] + z_0\}(.71)} = .5623$$

$$P(\text{spam}|\vec{X} = [0,1,0,0]) = \frac{\exp\{\vec{y} \cdot [0,1,0,0] + y_0\}(.29)}{\exp\{\vec{y} \cdot [0,1,0,0] + y_0\}(.29) + \exp\{\vec{z} \cdot [0,1,0,0] + z_0\}(.71)} = .999184$$