Lecture 9

Multiple word spam filter

Suppose we check for N words: $w_1, w_2, \ldots, w_N$.

Define the 0-1 random variables $X_i = \mathbb{1}\{\text{message has word } w_i\}$

Suppose $X_1 = a_1, X_2 = a_2, \cdots, X_N = a_N$, where a_i are either 0 or 1.

Assume each word appears in a message independent of the other words

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

independence (Naive Bayes)

Example: Consider the earlier example:

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

$P(\text{spam}) = \frac{1500}{1500+3672} = 0.29$

$P(\text{ham}) = 0.71$

Use smoothing, with smoothing parameters $(\alpha, \beta) = (1, 2)$.

(a) Email message has the words w_1, w_3, and w_4, but not the word w_2

$$P(\text{spam} \mid X_1 = 1, X_2 = 0, X_3 = 1, X_4 = 1) = ?$$

(b) Email message has the words w_1, w_3, but not the words w_2 or w_4

$$P(\text{spam} \mid X_1 = 1, X_2 = 0, X_3 = 1, X_4 = 0) = ?$$

$$P(\mathrm{s} \mid X_1 = 1, X_2 = 0, X_3 = 1, X_4 = 1) = \frac{P(X_1 = 1, X_2 = 0, X_3 = 1, X_4 = 1 \mid \mathrm{s})P(\mathrm{s})}{P(X_1 = 1, X_2 = 0, X_3 = 1, X_4 = 1)}$$

$$P(X_1=1, X_2=0, X_3=1, X_4=1 \mid \text{spam}) = P(X_1=1 \mid \text{spam})P(X_2=0 \mid \text{spam}) \times P(X_3=1 \mid \text{spam})P(X_4=1 \mid \text{spam})$$

$$= \frac{16+1}{1500+2} \cdot \left(1 - \frac{621+1}{1502}\right) \cdot \frac{125+1}{1502} \cdot \frac{1}{1502} = 1$$

$$P(X_1=1, X_2=0, X_3=1, X_4=1 \mid \text{ham}) = \frac{154}{3674} \times \left(1 - \frac{1}{3674}\right) \times \frac{32}{3674} \times \frac{1893}{3674} = 2$$

$$P(\text{spam} \mid X_1=1, X_2=0, X_3=1, X_4=1) = \frac{1\,(0.29)}{1\,(0.29) + 2\,(0.71)}$$

$$= 0.0008$$

$\rightarrow$ label "ham"

$$P(\mathrm{s} \mid X_1 = 1, X_2 = 0, X_3 = 1, X_4 = 0) = \frac{P(X_1 = 1, X_2 = 0, X_3 = 1, X_4 = 0 \mid \mathrm{s})P(\mathrm{s})}{P(X_1 = 1, X_2 = 0, X_3 = 1, X_4 = 0)}$$

$$= \frac{\frac{17}{1502} \cdot \left(1 - \frac{622}{1502}\right) \cdot \frac{126}{1502} \cdot \left(1 - \frac{1}{1502}\right)(0.29)}{(\text{above}) + \frac{154}{3674}\left(1 - \frac{1}{3674}\right) \cdot \frac{32}{3674} \cdot \left(1 - \frac{1893}{3674}\right)(0.71)}$$

$$= 0.56$$

$\rightarrow$ label "spam"

[depends on threshold]

Testing the model

We can use the following metrics to test the model:

$$\text{accuracy} = \frac{\text{correct predictions}}{\text{total}}$$

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

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

Example: Consider the email predictions:

	spam	ham
predict spam	101	33
predict ham	38	704

The three evaluation metrics we can use give:

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

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

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

Compact formulation of model

Goal: *pre-compute* parameters for the model, based on training data.

Suppose our spam filter keeps track of N different words. $w_1, w_2, \ldots, w_N$

When a new email message arrives, it is encoded by a vector

$$\vec{a} = [a_1, a_2, \ldots, a_N]$$

$a_1 = 1$ if w_1 in message
$a_1 = 0$ if w_1 is not

with 1's for the words that appear in the message and 0's for those that do not appear.

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

- $\exp\{x\} = e^x$, for an easier way to display the expressions,
- $\sum_{k=1}^{n} c_k = c_1 + c_2 + \cdots + c_n$,
- $\prod_{k=1}^{n} c_k = c_1 \times c_2 \times \cdots \times c_n$,
- exponentials and logs are inverses of each other: $x = e^{\log(x)}$, $\log(e^x) = x$
- property of logs: $\log(a \cdot b) = \log(a) + \log(b)$,
- property of logs: $\log(a^b) = b \log(a)$.
- for a large N set of words, we need to multiply many small probabilities, leading to underflow problems.

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

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 / ham message.

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

if $a_k = 0 \Rightarrow P(X_k = 0 | \text{spam}) = 1 - p_{ks} = (1 - p_{ks})^{1-a_k}$

if $a_k = 1 \Rightarrow P(X_k = 1 \,|\, \text{spam}) = p_{ks} = p_{ks}^{a_k}$

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

$\log a - \log b = \log \frac{a}{b}$

Combining into the logarithm notation:

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

depends on training data.

Let $y_0 = \sum_{k=1}^{N} \log(1 - p_{ks})$ and $y_k = \log \left(\frac{p_{ks}}{1 - p_{ks}} \right)$.

Set $\vec{X} = [X_1, \ldots, X_N]$, $\vec{a} = [a_1, \ldots, a_N]$ and $\vec{y} = [y_1, \ldots, y_N]$:

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

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

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

With y_0, $\vec{y}$, z_0, and $\vec{z}$ *pre-computed*, we classify messages:

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

Example: using the 4 words below, with smoothing $\alpha = 1$ and $\beta = 2$

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

$y_0 = \sum \log(1 - P_{ks})$

$y_k = \log\left(\frac{P_{ks}}{1 - P_{ks}}\right)$

Recall $P(\text{spam}) = .29$ and $P(\text{ham}) = .71$. Then:

$$\begin{aligned}\vec{y} &= \left[\log\left(\frac{\frac{17}{1502}}{1-\frac{17}{1502}}\right), \log\left(\frac{\frac{622}{1502}}{1-\frac{622}{1502}}\right), \log\left(\frac{\frac{126}{1502}}{1-\frac{126}{1502}}\right), \log\left(\frac{\frac{1}{1502}}{1-\frac{1}{1502}}\right)\right] \\ &= [-4.47, -0.35, -2.39, -7.31]\end{aligned}$$

$$\begin{aligned}y_0 &= \log\left(1-\frac{17}{1502}\right) + \log\left(1-\frac{622}{1502}\right) + \log\left(1-\frac{126}{1502}\right) + \log\left(1-\frac{1}{1502}\right) \\ &= -0.63\end{aligned}$$

$$\begin{aligned}\vec{z} &= \left[\log\left(\frac{\frac{154}{3674}}{1-\frac{154}{3674}}\right), \log\left(\frac{\frac{1}{3674}}{1-\frac{1}{3674}}\right), \log\left(\frac{\frac{32}{3674}}{1-\frac{32}{3674}}\right), \log\left(\frac{\frac{1893}{3674}}{1-\frac{1893}{3674}}\right)\right] \\ &= [-3.13, -8.21, -4.73, -0.06]\end{aligned}$$

$$\begin{aligned}z_0 &= \log\left(1-\frac{154}{3674}\right) + \log\left(1-\frac{1}{3674}\right) + \log\left(1-\frac{32}{3674}\right) + \log\left(1-\frac{1893}{3674}\right) \\ &= -0.776\end{aligned}$$

$$P(\vec{X} = [1,0,1,1]\,|\,\text{spam})$$
$$= \exp\{\underbrace{[-4.47, -0.35, -2.39, -7.31]}_{\vec{y}} \cdot \underbrace{[1,0,1,1]}_{\vec{a}} + \underbrace{(-0.63)}_{y_0}\}$$
$$= e^{-14.8}$$
$$= 3.7 \times 10^{-7}$$

$$P(\vec{X} = [1,0,1,1]\,|\,\text{ham})$$
$$= \exp\{\underbrace{[-3.13, -8.21, -4.73, -0.06]}_{\vec{z}} \cdot \underbrace{[1,0,1,1]}_{\vec{a}} + \underbrace{(-0.776)}_{z_0}\}$$
$$= e^{-8.696}$$
$$= 1.67 \times 10^{-4}.$$

$$P(\text{spam}|\vec{X} = [1,0,1,1])$$
$$= \frac{P(\vec{X} = [1,0,1,1]|\text{spam})P(\text{spam})}{P(\vec{X} = [1,0,1,1]|\text{spam})P(\text{spam}) + P(\vec{X} = [1,0,1,1]|\text{ham})P(\text{ham})}$$
$$= \frac{(3.7 \times 10^{-7})(.29)}{(3.7 \times 10^{-7})(.29) + (1.67 \times 10^{-4})(.71)}$$
$$= 0.00090$$

close to earlier computation.

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

close.