CS 476/676

An example of text classification, using Naive Bayes technique.

this example and some other material from Croft's new book "Search Engines: IR in Practice"

document *id*	cheap	buy	banking	dinner	the	*class*
1	0	0	0	0	2	not spam
2	3	0	1	0	1	spam
3	0	0	0	0	1	not spam
4	2	0	3	0	2	spam
5	5	2	0	0	1	spam
6	0	0	1	0	1	not spam
7	0	1	1	0	1	not spam
8	0	0	0	0	1	not spam
9	0	0	0	0	1	not spam
10	1	1	0	1	2	not spam
new						???

Suppose we have some documents, some of which are "spam", the rest "ham". ("ham" is easier to write than "not spam")

Suppose a new document arrives. What's the probability that it's spam? In conditional probability notation, what is $P(\text{"spam"} \mid \text{docnew})$?

Bayes Rule says

(1) $$P(C|D) = \frac{P(D|C)P(C)}{P(D)} = \frac{P(D|C)P(C)}{\sum_{c \in C} P(D|C=c)P(C=c)}$$

Croft. p. 342

We want to know $P(\text{spam}|d)$. If we knew $P(d|\text{spam})$, we'd just apply BR and be done. But we don't know $P(d|\text{spam})$.

Here's where ~~test data,~~ training data comes in. If the terms w_i in a document can be considered independent, then

(2) $$P(d|\text{spam}) \propto \prod_{w \in V} P(w_i|\text{spam})^{tf_{w,d}}$$

So how do we calculate those $P(w_i|\text{spam})$?

Using a maximum likelihood approach,

(3) $P(w_i|c) = \frac{\#\, w,c}{|c|}$

where c is either "spam" 3 documents or "ham" 7 documents.

$|c| = \underbrace{5+7+8}_{\text{\# terms in each docuned in training set}} = 20$

c = spam

$|c| = \underbrace{2+1+2+3+1+1+5}_{\text{\# terms in each docuned in traning set}} = 15$

c = ham

So for exaple $P(\text{the}|\text{'spam'}) = \frac{1+2+1}{20} = .2$

$P(\text{dinner}|\text{'spam'}) = \frac{0}{20} = 0$ uh oh

because if we plug this into eq. (2) the product, hence the probability, is zero.

So eg. (3) isn't satisfactory. Let's use a smoothing function, such as

(4) $$P(w_i|c) = \frac{\#_{w_i,c} + 1}{|c| + |V|}$$ from Croft p. 350

where $|V|$ is the size of the vocabulary, namely 5. Note that in eq (4), $P(w_i|c)$ cannot be zero.

$P(w_i\|c)$	c = spam	c = ham
w_i = cheap	$\frac{(3+2+5)+1}{(5+7+8)+5} = \frac{11}{25} = .44$	$\frac{1+1}{(2+1+2+3+1+1+5)+5} = \frac{2}{20} = .1$
buy	$\frac{2+1}{20+5} = \frac{3}{25} = .12$	$\frac{2+1}{15+5} = .15$
banking	$\frac{4+1}{20+5} = \frac{5}{25} = .2$	$\frac{2+1}{15+5} = .15$
dinner	$\frac{1}{25} = .04$	$\frac{1+1}{20} = .1$
the	$\frac{5}{25} = .2$	$\frac{9+1}{20} = .5$

These "term probabilities" when multiplied, give a "document" probability.

So consider a new document, d_{new} = "cheap buy banking the" (note - no dinner!)

$$P(d_{new} \mid \text{"spam"}) = \prod_{w \in V} P(w \mid \text{"spam"})^{\#w d}$$

$$= \left(\frac{11}{25}\right)^1 \times \left(\frac{3}{25}\right)^1 \times \left(\frac{5}{25}\right)^1 \times \left(\frac{1}{25}\right)^0 \times \left(\frac{5}{25}\right)^1$$

$$= \frac{11 \times 3 \times 5 \times 5}{25 \times 25 \times 25 \times 25} = \frac{33}{25^3} = .002112$$

don't be alarmed

$$P(d_{new} \mid \text{"ham"}) = \left(\frac{2}{20}\right)^1 \times \left(\frac{3}{20}\right)^1 \times \left(\frac{3}{20}\right)^1 \times \left(\frac{2}{20}\right)^0 \times \left(\frac{10}{20}\right)^1 = \frac{180}{(20^4)}$$

$$= .001125$$

Finally! We can apply BR

$$P(\text{"spam"} \mid \text{drew}) = \frac{P(\text{new} \mid \text{"spam"})\, P(\text{"spam"})}{P(\text{new} \mid \text{"spam"})\, P(\text{"spam"}) + P(\text{new} \mid \text{ham})\, P(\text{ham}$$

with three spam examples in the train set, we have $P(\text{spam}) = 0,3$ and $P(\text{ham}) = ,7$

Hence

$$P(\text{"spam"} \mid \text{drew}) = \frac{.002112 * .3}{(,002112 * ,3) + (.001125 * ,7)}$$

$$= \frac{.0006336}{,0006336 + .0007875} = \frac{6336}{6336 + 7875} =$$

$$\frac{6336}{14214} = 0,44$$

what about d_{newer} = "cheap buy cheap banking the"?

$$P(d_{newer} \mid spam) \propto$$

$$\left(\frac{11}{25}\right)^2 \times \left(\frac{3}{25}\right) \times \left(\frac{5}{25}\right) \times \left(\frac{5}{25}\right) = \frac{121 \times 75}{25^4} = .0232$$

$$P(d_{newer} \mid \text{"ham"}) \propto$$

$$\left(\frac{2}{20}\right)^2 \times \left(\frac{3}{20}\right) \times \left(\frac{3}{20}\right) \times \left(\frac{10}{20}\right) = \frac{4 \times 9 \times 10}{20^4}$$

$$= .00306$$

So

$$P(\text{"spam"} \mid d_{newer}) = \frac{.0232 * .3}{(.0232 * .3) + (.00306 * .7)}$$

$$= \frac{.00696}{.00696 + .002142} = \frac{.00696}{.009102} = .76$$