Spam filtering using Bayes model

Assignment #2

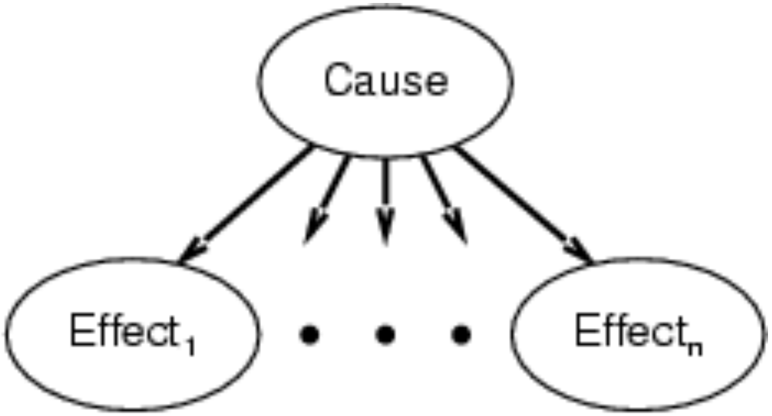
What we learned (or will learn) in Class

We do not need to know this P…

$P(\text{Cause} \mid \text{Effect}_1, \ldots, \text{Effect}_n) =$
$P(\text{Cause}, \text{Effect}_1, \ldots, \text{Effect}_n) / P(\text{Effect}_1, \ldots, \text{Effect}_n)$

We will just compute this P and find the 'Cause' which shows the highest P

$$\mathbf{P}(Cause, Effect_1, \ldots, Effect_n) = \mathbf{P}(Cause) \prod_i \mathbf{P}(Effect_i \mid Cause) .$$



- This probability distribution is called a **naive Bayes** model —"naive" because it is often used (as a simplifying assumption) in cases where the "effect" variables are not actually conditionally independent given the cause variable. (The naive Bayes model is sometimes called a **Bayesian classifier**)

- **We will use this Bayes model (which assumes variables are conditionally independent given a certain variable) for spam email filtering.**
- **We assume words in the email are conditionally independent given Spam (or Ham).**

Spam Message Filtering

- **Spam email example**
  - **You lose weight you sleep…..**
- **Compute P(Spam | 'you', 'lose', 'weight', 'you'…) and P(Ham | 'you', 'lose', 'weight', 'you'…)**

If P(Spam| …) > P(Ham| …), then the message is classified as a Spam message, otherwise it is a Ham message.

- **Using** $$P(Y, W_1 \ldots W_n) = P(Y) \prod_i P(W_i | Y)$$

you will compute P(Spam, 'you', 'lose', 'weight', 'you'…)

= P(Spam) x P('you'|Spam) x P('lose'|Spam)…

- **Because each P value is so small … after multiplication, you will have a very small number such as 0.00000000000000000000000…001232.**
- **To make the computation easier, we take logarithm on each side.**

log P(Spam, ‘you’, ‘lose’, ‘weight’, ‘you’…)=

log (P(Spam) x P(‘you’|Spam) x P(‘lose’|Spam) x P(‘weight’|Spam) x P(‘you’|Spam))x…

- **Using log a*b = log a + log b ,**

log P(Spam, ‘you’, ‘lose’, ‘weight’, ‘you’…)=

log (P(Spam) x P(‘you’|Spam) x P(‘lose’|Spam)…)=

log P(Spam) + log P(‘you’|Spam) + log P(‘lose’|Spam)…

Conclusion: To make the computation easier, we will take logarithm on each P value and do only addition (no multiplication).

Example: Spam Filtering

(Below is just example. You have to compute the real numbers from the training data)

Example)

$P(Y)$

```
ham : 0.66
spam: 0.33
```

$P(W|\text{spam})$

```
lose :  0.0156
to   :  0.0153
and  :  0.0115
of   :  0.0095
you  :  0.0093
a    :  0.0086
with:   0.0080
from:   0.0075
...
```

$P(W|\text{ham})$

```
the  :  0.0210
to   :  0.0133
lose :  0.0019
2002:   0.0110
with:   0.0108
from:   0.0107
you  :  0.0105
a    :  0.0100
...
```

(# of occurrences of W in all Spam emails) /
(# of occurrences of all words in all Spam emails)

(A) Prob. of spam=> log P(Spam) + log P('you'|Spam) + log P('lose'|Spam)… = log 0.33 + log 0.0093 + log 0.0156 + ….
(B) Prob. of Ham => log P(Ham) + log P('you'|Ham) + log P('lose'|Ham)… = log 0.66 + log 0.0105 + log 0.0019+ …
Because A>B, the message will be classified as spam.

Caution (Smoothing)

- $P(Spam, W_1, W_2, W_3 \ldots . W_n) =$

$P(Spam) \times P(W_1|Spam) \times P(W_2|spam) \times ..x\ P(W_n|spam)$

If $P(W_k|Spam)= 0$, then the whole P value will be zero after multiplication, regardless of the possibility of other words as a spam message. So, we will not allow zero value for $P(W_k|Spam)$ or $P(W_k|Ham)$.

- If P(w|Spam) or P(w|Ham) is 0, then add 1 to numerator.
  - e.g.) The number of occurrence of all words in the spam email messages is 1,900,323, but the number of occurrence of the word ‘spectacular’ is zero in spam. i.e., p(‘spectacular’| Spam) = 0/1,900,323=0. Then, change p(‘spectacular’| Spam) to 1/1,900,323.