Math 345 - Notes
Naive Bayes I
October 1, 2018

Naive Bayes

We discuss an algorithm used in supervised learning to classify unlabeled data. It is most commonly used to build a simple spam filter, where a message is classified as spam or ham (not spam). However, one can use Naive Bayes to classify data into more than two categories (think of using the generalized Bayes' Formula).

One word spam filter

Suppose we wish to classify a message as *spam* or *ham*, given that a specific word has appeared in the message. We compute

$$P(\text{spam}|\text{word}) = \frac{P(\text{word}|\text{spam})P(\text{spam})}{P(\text{word}|\text{spam})P(\text{spam}) + P(\text{word}|\text{ham})P(\text{ham})}.$$

Note: this comes directly from Bayes' formula, hence the name of the model Naive **Bayes.**

All these probabilities come from the available data that has been classified. The probabilities are easy to compute by means of simple counting: Suppose there are N total messages in the training set, out of which N_s is the number of spam messages and $N - N_s$ is the number of ham messages.

$$P(\text{spam}) = \frac{N_s}{N}, \qquad P(\text{ham}) = \frac{N - N_s}{N},$$

$$P(\text{word}|\text{spam}) = \frac{\#\text{ spam messages containing } word}{N_s},$$

$$P(\text{word}|\text{ham}) = \frac{\#\text{ ham messages containing } word}{N - N_s}.$$

We set up the rule:

If $P(spam|word) > 0.5$, then label as spam.

Example: Suppose the training set has 1500 spam messages and 3672 ham messages.
(a) The word *meeting* appears in 16 spam messages and in 153 ham messages.

(b) The word *pharmacy* appears in 621 spam messages and in 0 ham messages.

(c) The word *money* appears in 125 spam messages and in 31 ham messages.

(d) The word *DigiPen* appears in 0 spam messages and in 1892 ham messages.

Then for this example, $P(\text{spam}) = \dfrac{1500}{1500 + 3672} = 0.29$ and $P(\text{ham}) = 0.71$ and the associated probabilities are:

(a) $P(\text{spam}|meeting) = \dfrac{(16/1500)(0.29)}{(16/1500)(0.29) + (153/3672)(0.71)} = 0.094$, so label as **ham**.

(b) $P(\text{spam}|pharmacy) = \dfrac{(621/1500)(0.29)}{(621/1500)(0.29) + (0/3672)(0.71)} = 1$, so label as **spam**.

(c) $P(\text{spam}|\textit{money}) = \dfrac{(125/1500)(0.29)}{(125/1500)(0.29) + (31/3672)(0.71)} = 0.801$, so label as **spam**.

(d) $P(\text{spam}|\textit{DigiPen}) = \dfrac{(0/1500)(0.29)}{(0/1500)(0.29) + (1892/3672)(0.71)} = 0$, so label as **ham**.

Remarks:

1. One can use a probability threshold higher than 0.5 for classifying as spam!
2. For this training data, there is zero probability of getting a ham message that uses the word *pharmacy*, or a spam message that uses the work *DigiPen*. However, that should not be the case.
   - A ham message using *pharmacy* could be "I have to stop by the pharmacy before coming home."
   - A spam message using *DigiPen* could be "DigiPen students, awesome opportunity to get rich!"

   To account for these possibilities, we use smoothing, so that no word has a 1 or 0 probability of appearing in the set of spam or ham, as we show below.

Smoothing

Let the pair (α, β) be a pair of smoothing parameters. They should be small, but one can also try to optimize the model for a good pair, in the testing stage of the model.

$$P(\text{word}|\text{spam}) = \frac{\alpha + (\#\text{ spam messages containing } \textit{word})}{\beta + N_s},$$

$$P(\text{word}|\text{ham}) = \frac{\alpha + (\#\text{ ham messages containing } \textit{word})}{\beta + N - N_s}.$$

We will use $\alpha = 1$ and $\beta = 2$. Then the four examples above with smoothing parameters lead to

(a) $P(\text{spam}|\textit{meeting}) = \dfrac{(17/1502)(0.29)}{(17/1502)(0.29) + (154/3674)(0.71)} = 0.099$, so label as **ham**.

(b) $P(\text{spam}|\textit{pharmacy}) = \dfrac{(622/1502)(0.29)}{(622/1502)(0.29) + (1/3674)(0.71)} = 0.998$, so label as **spam**.

(c) $P(\text{spam}|\textit{money}) = \dfrac{(126/1502)(0.29)}{(126/1502)(0.29) + (32/3674)(0.71)} = 0.797$, so label as **spam**.

(d) $P(\text{spam}|\textit{DigiPen}) = \dfrac{(1/1502)(0.29)}{(1/1502)(0.29) + (1893/3674)(0.71)} = 0.0005$, so label as **ham**.