Naive Bayes

y variable is usually our target, or outcome.

x variable is what we observe (our data).

$$p(y|x) = \frac{p(x|y)\,p(y)}{p(x)}$$

$p(y)$ is our prior. This is the belief we have in an event occurring before observing any data (x).

$p(x)$ is our evidence. In NLP, our evidence is a value representing the probability of seeing a particular combination of words together (for example, $p(x_0 = \text{free} | x_1 = \text{money})$ should intuitively be higher than $p(x_0 = \text{tensorflow}, x_1 = \text{ballad})$ because "free" and "money" both tend to occur more frequently in the English language.

and tend to occur more frequently together (co-occurence).

$p(x|y)$ is our likelihood. It represents the likelihood of observation the data (comments, words, documents, etc.) given we know y. In Naive Bayes, we assume conditional independence. This means we can rewrite $p(x|y)$ as $\prod_{i}^{N} p(x_i|y)$ where N is the length of your observed data.

Note: we cannot make the same assumption of independence about the evidence $p(x)$.

Example:

SMS #1:

I ate dinner early HAM

SMS #2:

free money today SPAM

SMS #3

They will go now. HAM

SMS #4

I love to be free and have money HAM

Priors:

$p(y=\text{ham}) = \frac{3}{4}$ $p(y=\text{spam}) = \frac{1}{4}$

Likelihoods:

$$p(x \mid y=\text{ham}) = p(x=\text{free} \mid y=\text{ham}) \times p(x=\text{money} \mid y=\text{ham})$$

$$= \frac{1}{3} \times \frac{1}{3} = \frac{1}{9}$$

$$p(x \mid y=\text{spam}) = p(x=\text{free} \mid y=\text{spam}) \times p(x=\text{money} \mid y=\text{spam})$$

$$= \frac{1}{1} \times \frac{1}{1} = 1$$

Note, if we only care about classifying spam/ham and not the posterior probability, we do not need to actually calculate $p(x)$, our evidence:

$$p(y|x) \propto p(x|y)\,p(y)$$

Then we simply need to determine if $p(y=\text{ham}|x) > p(y=\text{spam}|x)$.

Evidence:

$$p(x) = \prod_{i}^{N} p(x_i) = p(x=\text{free})\,x\,p(x=\text{money})$$

$$p(x) = \sum_{i}^{C} p(x|y_i)\,p(y_i) =$$

$$p(x|y=\text{ham})\,p(y=\text{ham}) + p(x|y=\text{spam})\,p(y=\text{spam})$$

C is the set of distinct classes (in our case, spam or ham).

$$= (1/9)(3/4) + (1/1)(1/4)$$

$$= 1/3 = 0.33$$

Plug back into the Bayes Formula:

$$p(y=\text{ham}|x) = \frac{(1/9)(3/4)}{(1/3)}$$

$$= .25$$

$$p(y=\text{spam}\mid x) = \frac{(1/1)(1/4)}{(1/3)}$$

$$= .75$$

Notice the update from our prior to our posterior:

$p(y=\text{ham}) = 3/4 \rightarrow$

$p(y=\text{ham}\mid x) = 1/4$

$p(y=\text{spam}) = 1/4 \rightarrow$

$p(y=\text{spam}\mid x) = 3/4$

In this case, the posteriors should sum to 1, since the outcomes "spam" or "ham" are mutually exclusive and conditionally exhaustive. If you had made the incorrect assumption of independence for the evidence, your

posterior would be:

$$p(y=\text{ham}|x) = \frac{(1/4)(3/4)}{(3/4)\times(2/4)}$$

$= 0.33$

free appears twice in 4 documents

$$p(y=\text{spam}|x) = \frac{(1/1)(1/4)}{(2/4)\times(2/4)}$$

$= 1$

money appears one time in 4 documents

Notice these sum to >1, which is not logical.