The Naive Bayes Algorithm: Takeaways

Concepts

- When a new message "w_1, w_2, ..., w_n" comes in, the Naive Bayes algorithm classifies it as spam or non-spam based on the results of these two equations:

$$P(Spam|w_1, w_2, ..., w_n) \propto P(Spam) \cdot \prod_{i=1}^{n} P(w_i|Spam)$$
$$P(Spam^C|w_1, w_2, ..., w_n) \propto P(Spam^C) \cdot \prod_{i=1}^{n} P(w_i|Spam^C)$$

- To calculate $P(w_i|Spam)$ and $P(w_i|Spam^C)$, we need to use the additive smoothing technique:

$$P(w_i|Spam) = \frac{N_{w_i|Spam} + \alpha}{N_{Spam} + \alpha \cdot N_{Vocabulary}}$$

$$P(w_i|Spam^C) = \frac{N_{w_i|Spam^C} + \alpha}{N_{Spam^C} + \alpha \cdot N_{Vocabulary}}$$

- Below, we see what some of the terms in equations above mean:

$N_{w_i|Spam}$ = the number of times the word w_i occurs in spam messages
$N_{w_i|Spam^C}$ = the number of times the word w_i occurs in non-spam messages
N_{Spam} = total number of words in spam messages
N_{Spam^C} = total number of words in non-spam messages
$N_{Vocabulary}$ = total number of words in the vocabulary
$\alpha = 1$ (α is a smoothing parameter)

Resources

- A technical intro to a few version of the Naive Bayes algorithm
- An intro to conditional independence