Laplacian Smoothing

We usually compute the probability of a word given a class as follows:

$$P\left(\text{w}_i \mid \text{class}\right) = \frac{\text{freq}\left(\text{w}_i, \text{class}\right)}{N_{\text{class}}} \quad \text{class} \in \{\text{Positive, Negative}\}$$

However, if a word does not appear in the training, then it automatically gets a probability of 0, to fix this we add smoothing as follows

$$P\left(\text{w}_i \mid \text{class}\right) = \frac{\text{freq}(\text{w}_i, \text{class})+1}{(N_{\text{class}}+V)}$$

Note that we added a 1 in the numerator, and since there are V words to normalize, we add V in the denominator.

N_{class}: frequency of all words in class

V: number of unique words in vocabulary

Mark as completed

👍 **Like**   👎 **Dislike**   ⚑ **Report an issue**