3.19

Intro

In this exercise we will study the presence of irrelevant features in the naive Bayes classifier (NBC). We start the exercise with the information that the NBC is a linear classifier. Moreover, the exercise focus on the specific case of binary classification of words. In item a, we will derive the linear classifier expression capable of deciding which class $\{c_1, c_2\}$ is the most probable for the document. In item b, we derive the conditions over β and θ that make a given word w irrelevant. In item c, we analyse the impact of class imbalance in feature relevance. At last, in item d we suggest a additional feature selection step to help ignoring irrelevant words.

Solution

a)

We will write an expression for the log posterior odds ratio in a form that shows the action of the linear classifier:

$$\begin{aligned} &log_2 \frac{p(c=1|x_i)}{p(c=2|x_i)} = log_2 \frac{\frac{p(x_i|c=1)p(c=1)}{p(x_i)}}{\frac{p(x_i|c=2)p(c=2)}{p(x_i)}} = \\ &log_2 \frac{p(x_i|c=1)}{p(x_i|c=2)} = log_2 p(x_i|c=1) - log_2 p(x_i|c=2) = \phi(x_i)^t(\beta_1 - \beta_2) \end{aligned} \tag{1}$$

In (1), we used the fact that $p(c=1) = p(c=2) = 0.5$.

b)

First of all, let's see how the classification of new samples happens. For this, we will assume the following decision criteria: If $\frac{p(c=1|x_i)}{p(c=2|x_i)} > 1$ then document $i \in c_1$. Else, document $i \in c_2$ (if the posteriors are equal, you can choose either class or leave without a selection, this is up to you). The condition for choosing class 1 can be stated as follows:

$$\frac{p(c=1|x_i)}{p(c=2|x_i)} > 1 \iff log_2 \frac{p(c=1|x_i)}{p(c=2|x_i)} > 0 \iff \phi(x_i)^t(\beta_1 - \beta_2) > 0 \tag{2}$$

Therefore, when $\beta_{1,w} > \beta_{2,w}$ the word w is favoring class 1 over class 2. When the opposite holds true, the word w is favoring class 2 over class 1. Thus, for the classifier to ignore the word, we need $\beta_{1,w} = \beta_{2,w}$. This is equivalent to say:

$$\begin{aligned} &\beta_{1,w} = \beta_{2,w} \iff \log_2 \frac{\theta_{1,w}}{1-\theta_{1,w}} = \log_2 \frac{\theta_{2,w}}{1-\theta_{2,w}} \iff \frac{\theta_{1,w}}{1-\theta_{1,w}} = \frac{\theta_{2,w}}{1-\theta_{2,w}} \\ &\iff \theta_{1,w} = \theta_{2,w} \end{aligned} \tag{3}$$

c)

The word w occurs in every document and $n_1 \neq n_2$. Therefore, the posterior mean estimates for the parameters are:

$$\theta_{1,w} = \frac{1 + \sum_{i \in c=1} x_{iw}}{2 + n_1} = \frac{n_1 + 1}{n_1 + 2}$$
$$\theta_{2,w} = \frac{1 + \sum_{i \in c=2} x_{iw}}{2 + n_2} = \frac{n_2 + 1}{n_2 + 2} \tag{4}$$

Let's see when, these two parameters estimates are equal:

$$\frac{n_1 + 1}{n_1 + 2} = \frac{n_2 + 1}{n_2 + 2} \iff n_1 = n_2 \tag{5}$$

Since the classes are imbalanced, $\theta_{1,w} \neq \theta_{2,w}$ and the word w will not be ignored.

d)

Another way to encourage irrelevant features to be ignored is using pre-processing feature selection, like ranking.

Conclusion

In this exercise we studied several aspects of the irrelevant words in a binary NBC. In item a, we discovered that the decision between the classes c_1 and c_2 can be encoded in terms of the linear classifier with the expression $log_2 \frac{p(c=1|x_i)}{p(c=2|x_i)} = \phi(x_i)^t(\beta_1 - \beta_2)$. Therefore, when $\beta_{1,w} > \beta_{2,w} \iff \theta_{1,w} > \theta_{2,w}$, the word w is favoring class 1 over class 2. The opposite happens when $\beta_{1,w} < \beta_{2,w}$. To put it simple, the word w will give points to the most probable class and the class with the most points win.

In item b we deduced the conditions for making a given word w irrelevant in the classification. Based on the decision process above, it is straightfoward to conclude that a word will be irrelevant when the classifier cannot favor either one of the classes. This only happens when $\beta_{1,w} = \beta_{2,w} \iff \theta_{1,w} = \theta_{2,w}$.

In item c we used a reasonable parameter estimator (the posterior mean)to calculate $\theta_{c,w}$. Meanwhile, we also know that the word w is irrelevant because it occurs in every document. We would expect it would be filtered out by our feature selection process. However, as the result have shown us, the word is not filtered. This is due our feature selection approach which only filters words that have the same importance for both classes. Therefore, no matter how big $\theta_{1,w}, \theta_{2,w} (\theta_{1,w} \neq \theta_{2,w})$ are, the words will not be filtered. Note that if we were working only with the MLE, this problem would not have happened.

In item d we suggest the use of ranking as an complement for feature selection. This approach would discard non discriminative words, like the one appearing in item c, without adding too much complexity to the solution. Moreover, ranking does not interfer with the word filtering made by the classifier.