Mid-Semester

Name: P. V. Sriram

Roll No.: 1801CS37

Vocabulary

In both the Multi-Nomial and Multi-Variate classes, we first extract the unique words in the Train dataset, after performing the preprocessing (Removing the stop words, punctuations). These unique words would be the part of the vocabulary. Additionally I have also included a “cutoff frequency” to consider only words occuring above a threshold.

But due to the relatively small size of dataset, we keep the cutoff frequency as zero, which results in the following dictionary.

'recent'
'years'
'researchers'
'computer'
'vision'
'proposed'
'many'
'deep'
'learning'
'methods'
'various'
'tasks'
'facial'
'recognition'
'made'
'enormous'
'leap'
'using'
'techniques'
'systems'

'benefit'
'hierarchical'
'architecture'
'learn'
'discriminative'
'face'
'representation'
'widely'
'used'

The vocabulary consists of 29 words.

Model Parameters

Both the models consider two parameters while training:

1) **Alpha** :
   a) The smoothing parameter.
   b) The grace frequency we allot to the words which do not at all occur in any class inorder to avoid the probability collapse.
   c) In this case, we notice that the frequency range lies in between 10e-2 to 10e-4.
   d) Therefore we consider 0.001 additionally in numerator for best performance.
2) **Cutoff Frequency** :
   a) Threshold parameter.
   b) In a dataset, to avoid memory overloading, we need to keep the feature size low.
   c) This could be achieved by adding only those words which occur more frequently in a document.
   d) But due to the low size of dataset, we do not opt for such a threshold, hence 0.

Comparision

In a general scenario of larger datasets, it is observed that Multi-Nomial Naive bayes is more accurate tehan the Bernoulli variant. This is because of the fact that the former is more precise in gathering properties of the constituent words. However the former is also slightly more computational then the later due to the fact that we are required to store more data and are required to look for all occurences of a single word. However the later just needs the existence of a word.

However, in this particular case. Due to the relatively small size of the dataset (3 sentences) we can observe that occurences and frequencies are largely overlapping, owing to the fact that there are only few words repeating themselves in a sentence. Therefore there is no much difference in efficiency and accuracy.