Info

Analysis:

The Naïve Bayes classifier function from sklearn.naive_bayes was used to classify both the datasets. There are 3 different types ofclassifier functions

- Gaussian Naïve Bayes,
- Bernoulli's Naïve Bayes and
- Multinomial Naïve Bayes

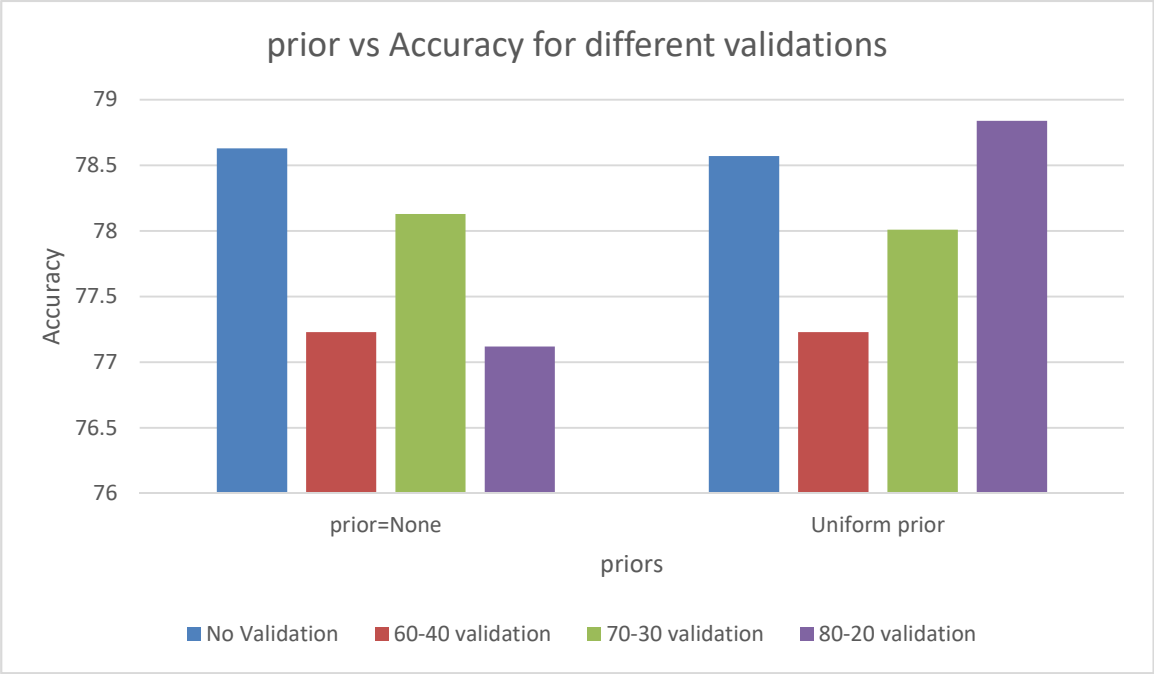
They differ mainly by the assumptions they make regarding the distribution of $P(x_i|y)$. We tried different combinations of the parameters for Naïve Bayes Classifiers and measured the accuracy for each combination. The analysis is given below.

1) **Prior:**

We tried 2 cases.

- Prior = none
- Prior = Guassian.

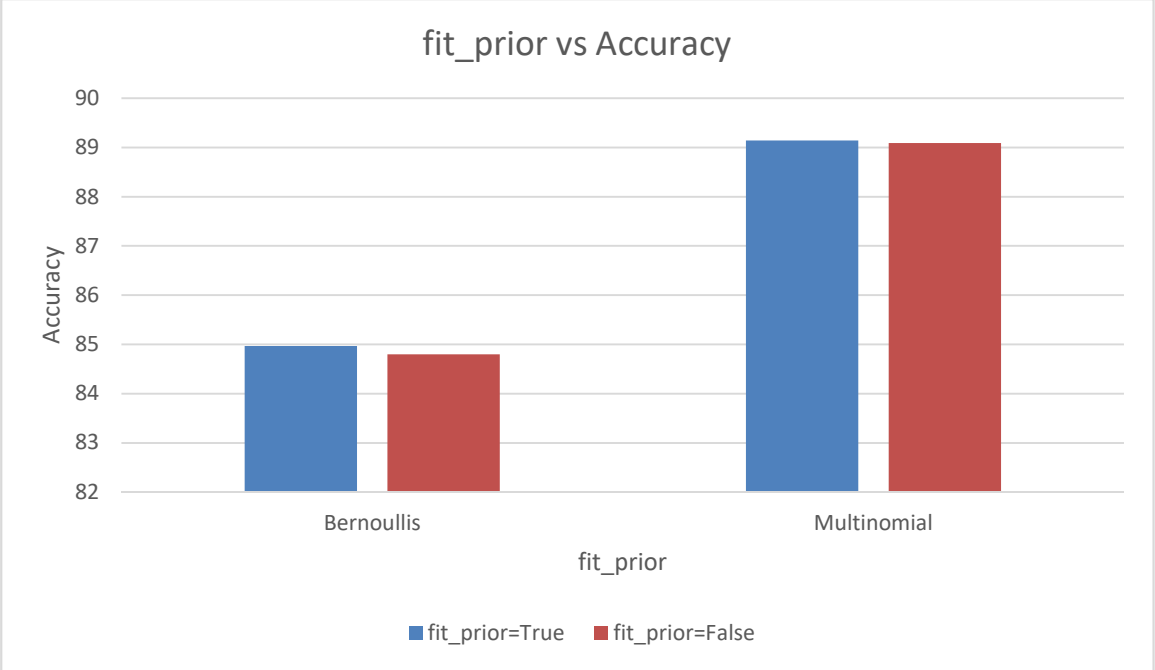
The results are given below.



The main reason for accuracy remaining almost the same for different priors could be that the uniform and non-uniform priors could almost be the same for the digit recognition dataset.

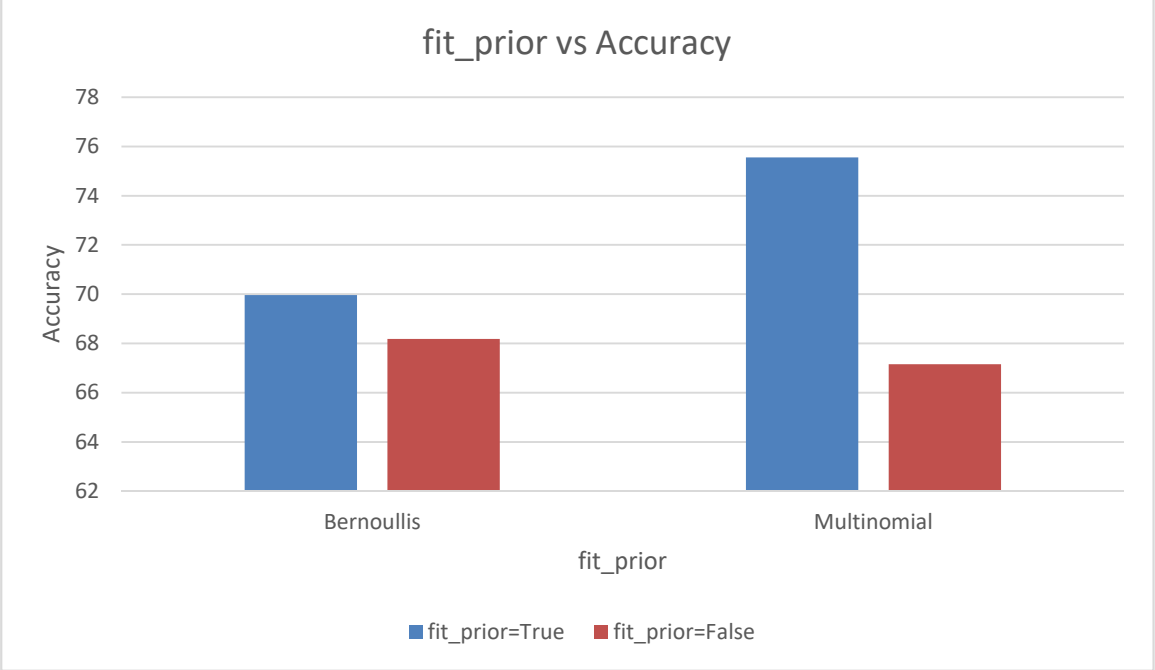
2) **fit_prior:**

It tells the classifier whether to learn the class prior probabilities or not. The accuracy when this parameter is set to true and false is given below.



The accuracy remained same when the fit_prior was set to true and false.

Influence of **fit_prior** on Amazon dataset



When there is an imbalance in class distribution, instead of using uniform prior probabilities it is better to calculate prior probabilities based on their frequency of occurrence as using uniform probabilities tend to wrongly predict new samples.

Digit Recognition Dataset:

We used the GaussianNB, MultinomialNB and BernoulliNB classifier function from sklearn library to classify the data. The accuracy results with different splits are given in the NaiveBayes_Results.pdf file.

Amazon dataset:

We performed sentiment classification on it. In Sentiment classification a review given in text format, is classified as a positive review or a negative review.

Easy Parts:

The easy part of this assignment is implementation of Digit Recognition. It took less time and analysis was quicker. So, it became easy for taking the decision of parameters in both the algorithms.

Challenging part of the Assignment:

The Amazon Baby Review dataset was challenging. Understanding what was given and what was asked from the dataset was the most difficult part. Once we got to know what to do, we tried different methods to predict the ratings from reviews. The naïve method was to searchfor words with positive and negative meanings and predict the rating. Then we used Natural Language Processing Toolkit to analyse the data and predict the rating from the reviews.

Results:

The Naïve Bayes classifier gave accuracy of 89.14% on the digit recognition dataset. It gave an accuracy of 75.56% on the amazon review dataset. A detailed report on all the results obtained is given in the NaiveBayes_Results.pdf file.