20 Newsgroup Dataset

1. The 20news group dataset is a collection of organized into 20 different newsgroups. First, we use multinomial Naive Bayes to model class conditional densities.
2. Consider representing an text as the set {x1, x2, . . . , xn} of distinct words contained in the text.
3. We are interested in computing $p(C_k|x_1, x_2, ..., x_n)$ for Ck classes
4. Let *x=(x1,x2,…,xn)*. Using Bayes theorem we get

$$p(C_k|x) = \frac{p(C_k)p(x|C_k)}{p(x)} \propto p(C_k)p(x|C_k) = p(C_k, x_1, x_2, ..., x_n)$$

5. In Naive Bayes, we assume all features are iid

$$= p(x_1|C_k) \cdot p(x_2|C_k) ... \cdot p(x_n|C_k) \cdot p(C_k)$$
$$= p(C_k) \prod_{i=1}^{n} p(x_i|C_k)$$

6. We calculate the posterior probability for each class of a given text and assign the class label which has a higher posterior probability.

$$y = arg\ max\ p(y = C_k) \prod p(x|y = C_k)$$

7. We calculate prior probability for each class

$$p(C_k) = p(y = C_k) = \frac{\sum_{i=1}^{N} I(y_i = C_k)}{N}$$

8.We calculate class conditional densities

p(x/y=Ck) = frequency of xi appeared in class k/(frequency of all words of class k)

So we can calculate posterior probabilities and assign the class labels.

Number of most frequent words	Error
385	51.59
872	39.12
1868	31.87
2866	29.14
3865	27.5
4863	26.36
5862	25.98
6861	25.24
9860	24.05
19859	21.63
29856	21.29

Dirichlet

Constructing a bayes classifier for 20 news dataset using Dirichlet class conditional densities involves the following steps:

1. select the top 1000 most frequent words among the entire training data after removing the stop words.

2. create a matrix consisting of word indices as columns and respective word counts as values which are indexed
by document index for each class separately.

3. perform laplace smoothing for all the matrices(xi+a/ N + a*d)

4. assuming the data comes from Dirichlet distribution, estimate the parameters(alpha(i)'s) for each class.

5. logarithm of the Dirichlet PDF is calculated, since calculating the coefficient of PDF involved overflow in python.

6. calculate the priors as a proportion of the documents in each class.

7. To predict the class of test data(after laplace smoothing), log(PDF(test_data))+ log(prior(i)) is calculated for all class
using the MLE estimates.

8. class with the highest of the above value is used as the predicted class.

results:

Hyperparameters:
features, smoothing parameter(a), test accuracy

1000	0.003	24.14%
1000	0.03	22.96%
1000	0.0003	28.60%
1000	0.0009	26.10%
2000	0.003	21.16%
2000	0.03	21.05%
2000	0.0003	23.51%
2000	0.0009	22.25%

observations:

increasing the number of features from 1000 to 2000 decreased accuracy, this might be due to the reason that the new features are not important features and their addition decreased the importance of the first 1000 features.

The accuracy is sensitive towards the smoothing factor, we have a good accuracy around 0.0003 value.

SMSSPAMCOLLECTION

It is a collection of 5572 messages and its label(spam and not spam).
We used **multinomial Naive Bayes** to model class conditional densities. We used 70% of data for training and 30% for data for testing. We are using the most frequent 2500 words. We got an accuracy of 94%.
We also used **Dirichlet distribution** for modeling class conditional densities. We got an accuracy of 77%.