Homework 2
-Vibhuti Mahajan (vm2486)

ASSIGNMENT-2

- Vibhuti Mahajan (vm2486)

1. $$\hat{\pi}, \hat{\theta}_y^{(1)}, \hat{\theta}_y^{(2)} = \underset{\pi, \theta_y^{(1)}, \theta_y^{(2)}}{\arg\max} \sum_{i=1}^{n} \ln p(y_i|\pi) + \sum_{i=1}^{n} \ln p(x_{i1}|\theta_{y_i}^{(1)}) + \sum_{i=1}^{n} \ln p(x_{i2}|\theta_{y_i}^{(2)})$$

MLE estimates:

(a) $$\hat{\pi} = \underset{\pi}{\arg\max}\ \ell \Rightarrow \frac{d\ell}{d\pi} = \frac{d}{d\pi}\left(\sum_{i=1}^{n} \ln\left(\pi^{y_i}(1-\pi)^{1-y_i}\right)\right) = 0$$

$$= \sum_{i=1}^{n} \frac{d}{d\pi}\left(y_i \ln\pi + (1-y_i)(1-\pi)\right) = 0$$

$$= \sum_{i=1}^{n}\left(\frac{y_i}{\pi} - \frac{1-y_i}{1-\pi}\right) = 0$$

$$\Rightarrow \hat{\pi} = \frac{\sum_{i=1}^{n} y_i}{n} = \bar{y}$$

To find ML estimates for x given $Y = y$:

(b) $$\hat{\theta}_0^{(1)} = \underset{\theta_0^{(1)}}{\arg\max}\ \ell = \frac{d\ell}{d\theta_0^{(1)}} = \frac{d}{d\theta_0^{1}}\left(\sum_{i=1}^{n} \ln\left(\theta_{y_i}^{(1)\,x_{i1}}(1-\theta_{y_i}^{(1)})^{(1-x_{i1})}\right)\right)$$

$$= \frac{d}{d\theta_0^{1}}\left(\sum_{i=1}^{n} 1(y_i=0) \ln\left(\theta_{y_i}^{1\,x_{i1}}(1-\theta_{y_i}^{1})^{(1-x_{i1})}\right)\right) = 0$$

This is the conditional distribution of $(x_i|y=0)$

Note :- $$P(x_i|Y=0) = \frac{P(X_i = x_i \cap Y=0)}{P(Y=0)}$$

$$= \sum_{y_i=0}\left(\frac{x_{i1}}{\theta_{y_0}^{(1)}} - \frac{1-x_{i1}}{1-\theta_{y_0}^{(1)}}\right) = 0$$

$$\Rightarrow \theta_{y_0}^{(1)} = \frac{\sum_{y_i=0} x_{iy_0}}{\sum_{y_i=0} 1} \quad \frac{\sum_{i=1, y_i=0}^{n} x_{i1}}{\sum_{i=1, y_i=0}^{n} n}$$

Similarly, $$\hat{\theta}_1^{(1)} = \frac{\sum_{i=1, y_i=1}^{n} x_{i1}}{\sum_{i=1, y_i=1}^{n} n}$$

(c) $\hat{\theta}_0^{(2)} = \underset{\theta_0^{(2)}}{\arg\max}\, l = \frac{d}{d\theta_0^{(2)}} \sum_{i=1}^{n} 1(y_i=0) \ln\left(\theta_{y_i}^{(2)} x_{i2}^{-(\theta_{y_i}^{(2)}+1)}\right) = 0$

$$= \sum_{i=1, y_i=0}^{n} \frac{d}{d\theta_0^{(2)}} \left(\ln \theta_{y_i}^{(2)} \cancel{x_{i2}} - (\theta_{y_i}^{(2)}+1) \ln x_i^{(2)}\right) = 0$$

$$\Rightarrow \sum_{i=1, y_i=0}^{n} \left(\frac{1}{\theta_{y_i}^{(2)}} - \ln x_{i2}\right) = 0$$

$$\Rightarrow \hat{\theta}_0^{(2)} = \frac{\sum_{i=1, y_i=0}^{n} 1}{\sum_{i=1, y_i=0}^{n} \ln x_{i2}}$$

Similarly, $\hat{\theta}_1^{(2)} = \sum_{i=1, y_i=1}^{n} \frac{1}{\sum_{i=1, y_i=1}^{n} \ln x_{i2}}$

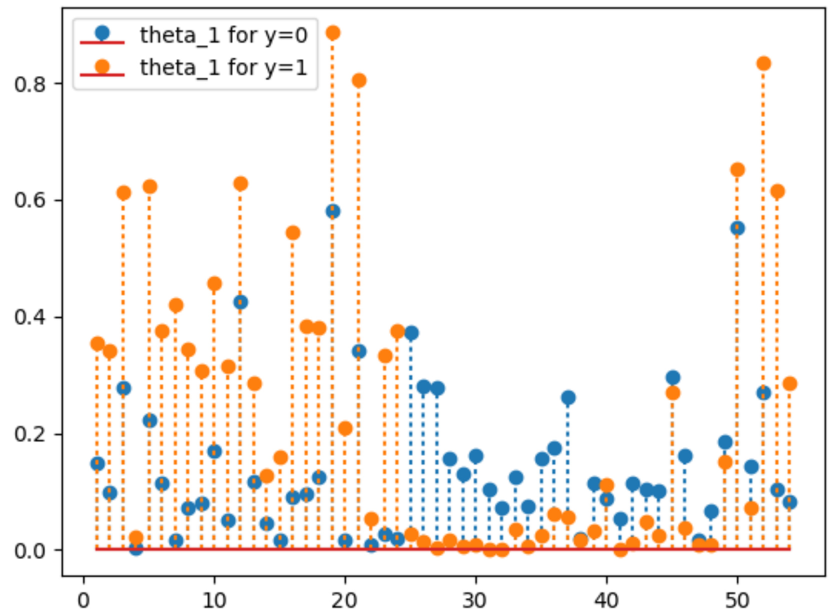
Problem 2:

a) Naïve Bayes Classifier:

Confusion Matrix

	Predicted y=0	Predicted y=1
Actual y=0	54	2
Actual y=1	5	32

Accuracy: 92.47311 %



b) Feature 16 is the frequency of occurrence of the word 'free' and feature 52 is the frequency of occurrence of '!'. Occurrence of either of these words makes the mail more probable to be classified as a spam, which does make sense since these words are more likely to be found in non-genuine mails.

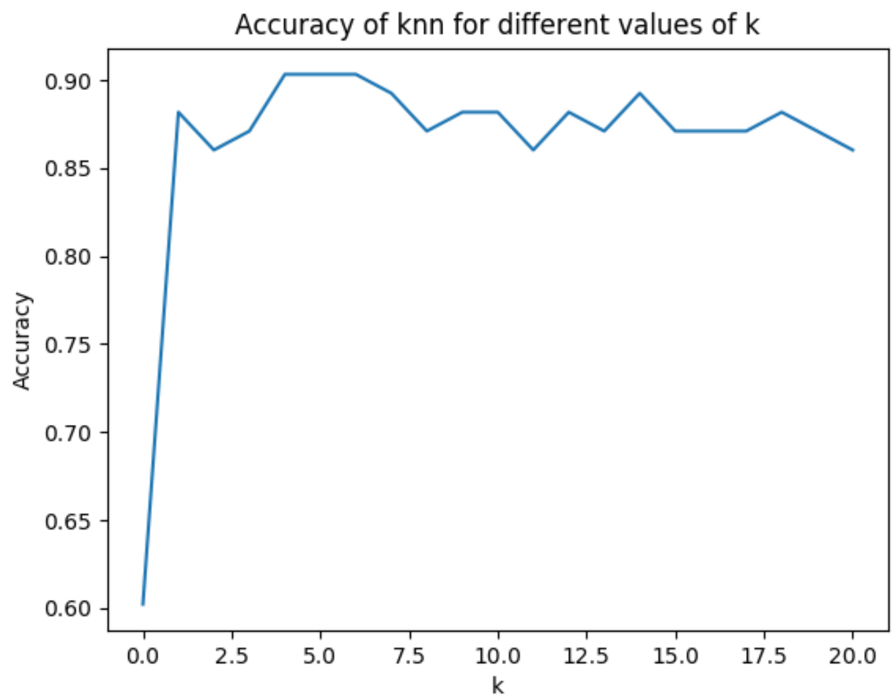
If the word 'free' occurs in the document, then it is 6 times more likely to be classified as a spam mail.

If the word '!' occurs in the document, then it is 3 times more likely to be classified as a spam mail

c) KNN Classifier: Accuracy values for different values of k.

K	Accuracy
1.	0.88172043
2.	0.86021505
3.	0.87096774
4.	0.90322581
5.	0.90322581
6.	0.90322581
7.	0.89247312
8.	0.87096774

9.	0.88172043
10.	0.88172043
11.	0.86021505
12.	0.88172043
13.	0.87096774
14.	0.89247312
15.	0.87096774
16.	0.87096774
17.	0.87096774
18.	0.88172043
19.	0.87096774
20.	0.86021505



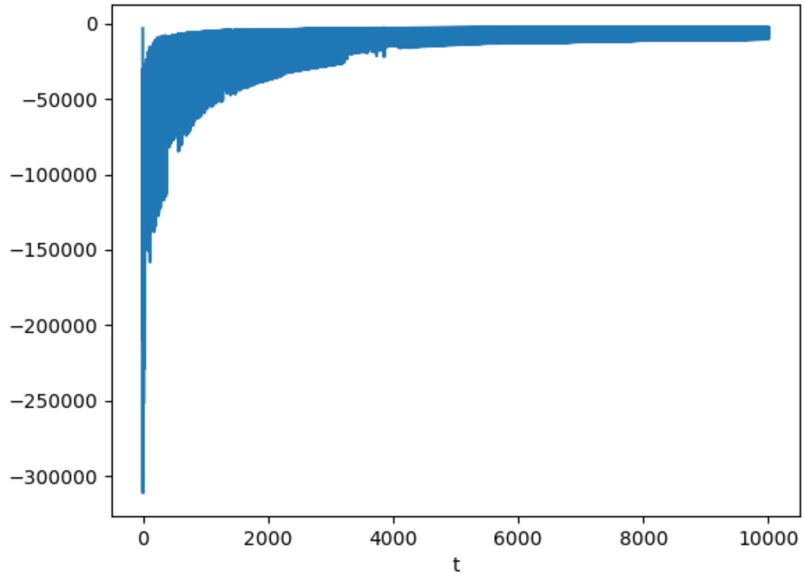
Value of accuracy is maximum for k = 4, 5, 6 = 90.32% . hence I would probably use k=4 since computation complexity is less for a smaller k.

d) Logistic Regression Classifier:

	Predicted y=-1	Predicted y=1
Actual y=-1	54	2
Actual y=1	27	10

Accuracy: 68.82%

There are a lot of spam mails, which are being classified as not-spam. This is probably due to the skewness in the data where we have more examples of non-spam mails as compared to non-spam mails



Plot of Objective function v/s iteration number for the gradient descent

e) <u>Newton's Method Classifier</u>:

	Predicted y=-1	Predicted y=1
Actual y=-1	54	2
Actual y=1	6	31

Accuracy: 91.398%

Plot of Objective function v/s iteration number for the gradient descent for Newton's Method.