ISCO630E

*Analysis Report on **Assignment 5***

Name - Sashank Mishra
Roll Number - IIT2016515

QUESTION 1

Dataset : Spam-Ham Data
5572 rows, 2 columns

	label	message
0	0	Go until jurong point, crazy.. Available only ...
1	0	Ok lar... Joking wif u oni...
2	1	Free entry in 2 a wkly comp to win FA Cup fina...
3	0	U dun say so early hor... U c already then say...
4	0	Nah I don't think he goes to usf, he lives aro...

Data Preprocessing :

- The number of unique entries is calculated, which turns out to be 5169. Therefore, the dataset has repeated entries, which are removed.
- The special characters, like '<' '>' are removed, only those characters are selected which are either alphanumeric or blank space.
- The whole sentence is converted into lowercase.
- Data is splitted into train set and test set, with test set having 30% of the total data.

Calculations:

1. Calculate P(spam) and P(ham).
2. Get count of all the words present in spam and ham emails differently.
3. Calculate P(wi|spam) and P(wi|ham), for all the words.
4. Classify according to the naive bayes formula

Analysis :

Accuracy achieved - 93.29%

We now obtain the Confusion Matrix on the test dataset. The confusion matrix is then visualized as follows:

Other parameters calculated:
Precision - 100%
Recall - 44.68%
F_1 Score - 0.61764

QUESTION 2

Dataset : Hooghly River Satellite Images, on 4 different image bands:
R-band,G-band,B-band,I-band
All images have dimensions 512*512, with a GIF format

Data Generation and Processing :

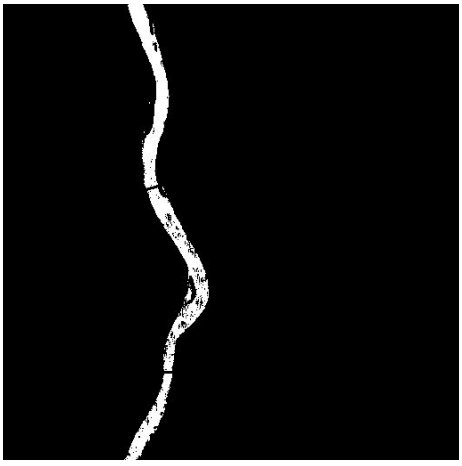
- We need to extract the river and the non river points for training. For this, we take one of the image, whiten the small part of the river, and blacken the small part of non-river. Then we extract the respective coordinates programmatically. Through this process, we get 1276 river points, and 6352 non-river points.
- We select coordinates of 50 river points and 100 non-river points for training. We extract the label from all the 4 images on that respective coordinate.
- Finally, we have river pints of shape (50,4) and non river points of shape (100,4).
- Also, we create a combined image from all the labels

Calculations:

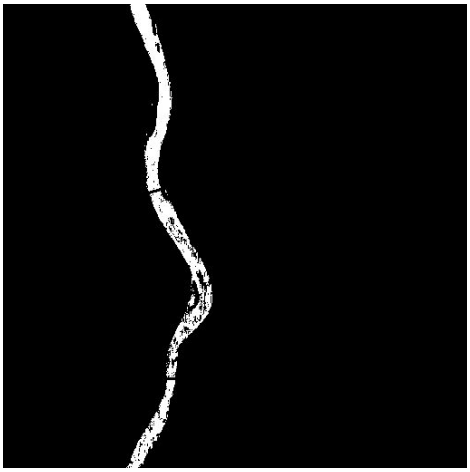
1. Calulate mean and covariance for both the classes
2. In the original combined image, for every pixel, calculate its deviation and transposed deviation from mean of river class. Then multiply it with the covariance of the river class.
3. Calculate density function (multivariate normal distribution) p1 of river class.
4. Perform step 2 and 3 for non-river class as well, and calculate p2.
5. For every pixel location, apply baye's rule and create a binary image so as to identify the river in the original image.

Result:

For P1 = 0.7, P2 = 0.3



For P1 = 0.5, P2 = 0.5



For P1 = 0.3, P2 = 0.7