ASSIGNMENT 5 - IIT2016039

Problem Statement 1:

Using Naive Bayesian classifier predict where a given mail is spam or not. Use the data set provided for this purpose. (structured data set)

- First we read the data set and separate it in into training and testing and also inputs and outputs.

- We now have to preprocess the data. In preprocessing we followed the following steps :
    - First we removed all letters into lower case
    - We removed stop words
    - We removed punctuations
    - And finally we performed stemming

- Now we have to vectorize the data. In it we simply assign a number to each word and calculate frequency in each test case

- Then we sum the frequencies of each word in spam and ham

- The assuming that the occurrence of each word is independent and we calculate the probability of being spam and ham

- Then we assign the class that has more probability

By doing that we got and accuracy of 97.8

Confusion Matrix = [[942 12] [24 137]] = [[TP FP] [FN FP]]

Where T = True, F = False, P = Positive, N = Negative.

Problem Statement 2:

Using Naive Bayesian classifier predict river non river using Satellite data set of Hooghly river (unstructured data set)

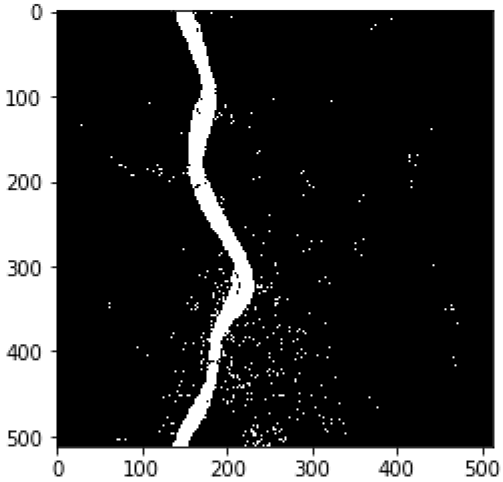
- Here we have the points first we have to find the river and non river points, For this we run a program which will give points on clicking the image. So by that we first take 50 river pixels and then 100 non river pixels.

- Then we have to find covariance and variance

- Using those we have to apply Gaussian Distribution and find the river and non river points in the whole image

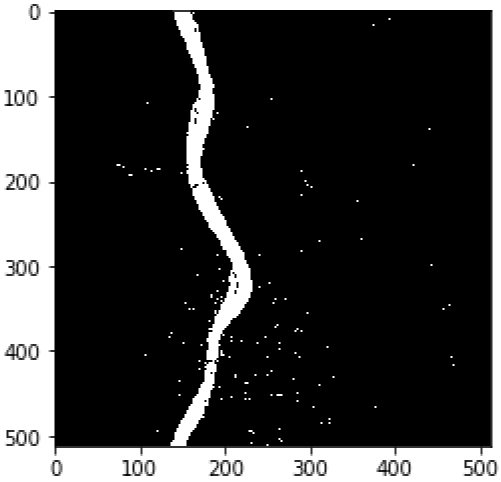
Here we are mentioned to classify the pixels in three different cases:
Case 1:

Probability of pixel to be river = 0.3 and non river is 0.7 then the output is as follows:



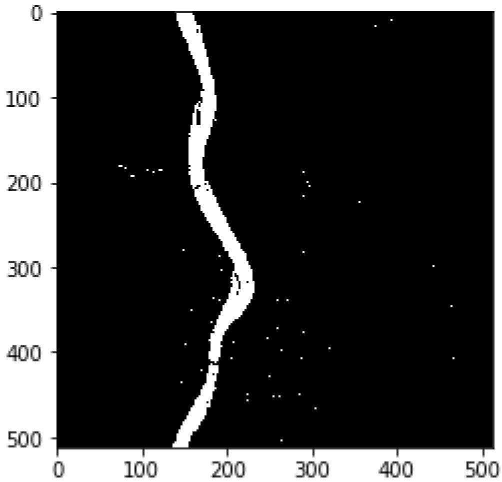
Case 2:

Probability of pixel to be river = 0.5 and non river is 0.5 then the output is as follows:



Case 3:

Probability of pixel to be river = 0.7 and non river is 0.3 then the output is as follows:


0
100
200
300
400
500
0
100
200
300
400
500