Assignment 3

Classification of results is very important in medical domain. We also need to know how well the decision boundary and the results work. Thus, we need to find a test to determine the accuracy of classification. There are two methods to decide that

1) Confusion Matrix
2) Sensitivity & Specificity

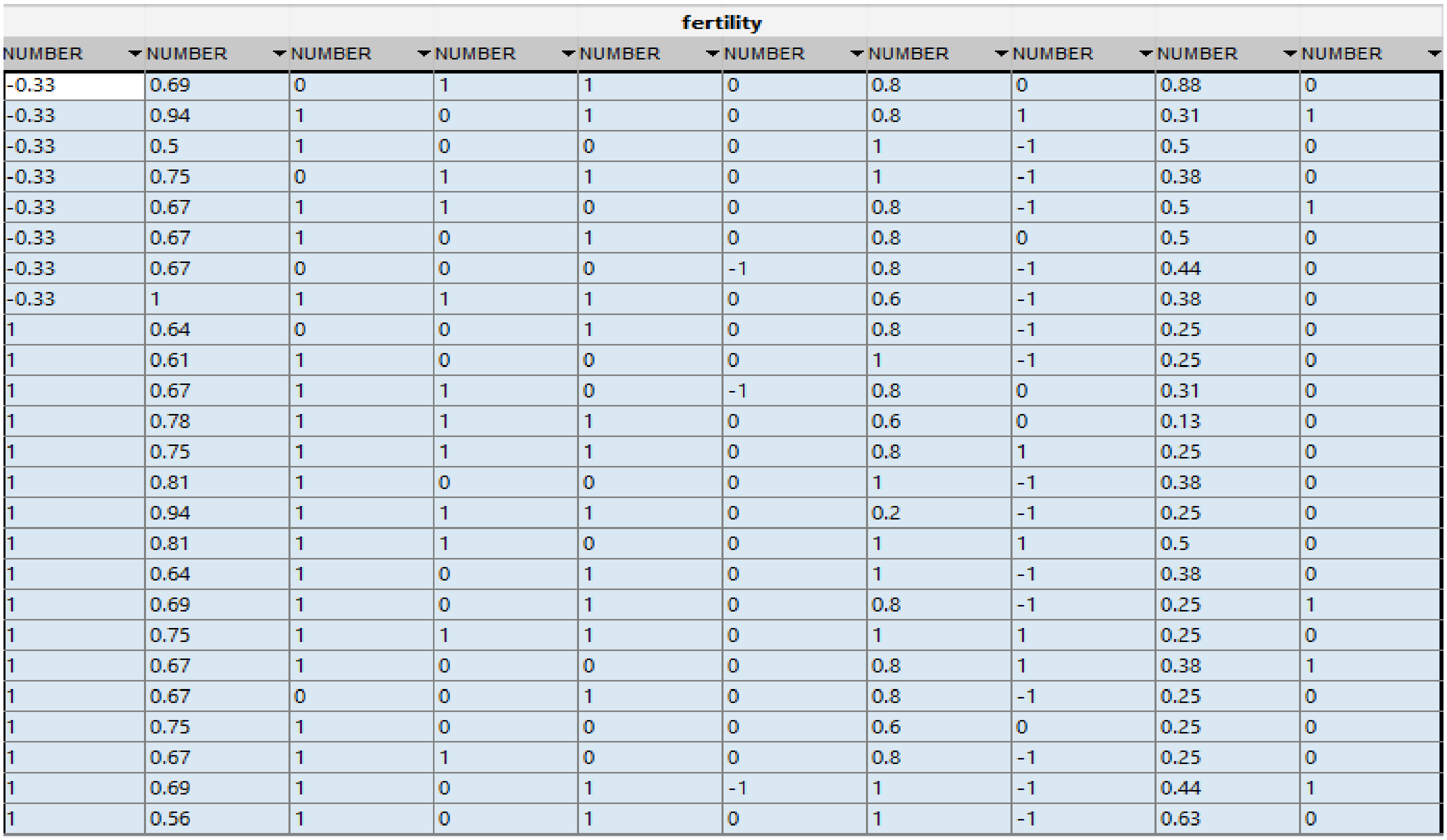
Latter is usually used to classify just two classes which is usually the case in medical domain. Its either the test is positive or negative. Hence, it becomes a good method to determine the accuracy of the tests based on two parameters sensitivity and specificity.

1. Sensitivity – measures the proportion of positives that are correctly identified as such (e.g. the percentage of sick people who are correctly identified as having the condition).
2. Specificity- measures the proportion of negatives that are correctly identified as such (e.g. the percentage of healthy people who are correctly identified as not having the condition).

In this assignment, we were allocated three tasks-:

1. Find the beta coefficients for the given features and outcomes.
2. Compare the beta column vector with the correlation vector obtained between outcomes and features
3. Vary the decision boundary from 0 to 1 and obtain the ROC.

Data – We have obtained data from UCI repository - https://archive.ics.uci.edu/ml/datasets/Fertility



fertility									
NUMBER	NUMBER	NUMBER	NUMBER	NUMBER	NUMBER	NUMBER	NUMBER	NUMBER	NUMBER
-0.33	0.69	0	1	1	0	0.8	0	0.88	0
-0.33	0.94	1	0	1	0	0.8	1	0.31	1
-0.33	0.5	1	0	0	0	1	-1	0.5	0
-0.33	0.75	0	1	1	0	1	-1	0.38	0
-0.33	0.67	1	1	0	0	0.8	-1	0.5	1
-0.33	0.67	1	0	1	0	0.8	0	0.5	0
-0.33	0.67	0	0	0	-1	0.8	-1	0.44	0
-0.33	1	1	1	1	0	0.6	-1	0.38	0
1	0.64	0	0	1	0	0.8	-1	0.25	0
1	0.61	1	0	0	0	1	-1	0.25	0
1	0.67	1	1	0	-1	0.8	0	0.31	0
1	0.78	1	1	1	0	0.6	0	0.13	0
1	0.75	1	1	1	0	0.8	1	0.25	0
1	0.81	1	0	0	0	1	-1	0.38	0
1	0.94	1	1	1	0	0.2	-1	0.25	0
1	0.81	1	1	0	0	1	1	0.5	0
1	0.64	1	0	1	0	1	-1	0.38	0
1	0.69	1	0	1	0	0.8	-1	0.25	1
1	0.75	1	1	1	0	1	1	0.25	0
1	0.67	1	0	0	0	0.8	1	0.38	1
1	0.67	0	0	1	0	0.8	-1	0.25	0
1	0.75	1	0	0	0	0.6	0	0.25	0
1	0.67	1	1	0	0	0.8	-1	0.25	0
1	0.69	1	0	1	-1	1	-1	0.44	1
1	0.56	1	0	1	0	1	-1	0.63	0

Fig 1: Fertility data – UCI repository

Task I

In this task, we had to generate the beta vector using the WEINER-HOPF equation. Firstly, the data is named fertility from UCI repository. In this task, we did the following steps:

1. Imported data and save features into X(100x9). Also added bias of ones at the start of X to take beta0 into consideration.

2. Imported output classes from data (100x1).
3. Generated beta from the equation by simple matrix multiplication (10x1) with the equation below

$$\hat{\beta} = (X^T X)^{-1} X^T Y.$$

Results

1. Beta column vector was of length 10 including the beta0.
2. Values in beta were ranging from -0.2 to 0.4

Expectations

Expected values to be closer to 0 and 1 as the output class in the dataset was closer to 0 and 1 but that wasn't the case.

Task II

In this task, we were supposed to compare the correlation vector between outcome and features with the beta vector. In this task, I performed the following tasks:-

1. Took the same variables X and Y as in task I.
2. Concatenated both the vectors with Y at the beginning.
3. Found the correlation using corr function.
4. Extracted the first column from the correlation matrix which is the correlation between outcomes and features
5. Compared the correlation vector with beta and plotted the graph.
6. Tried three configurations of X. column standardized, fully standardized and just X.

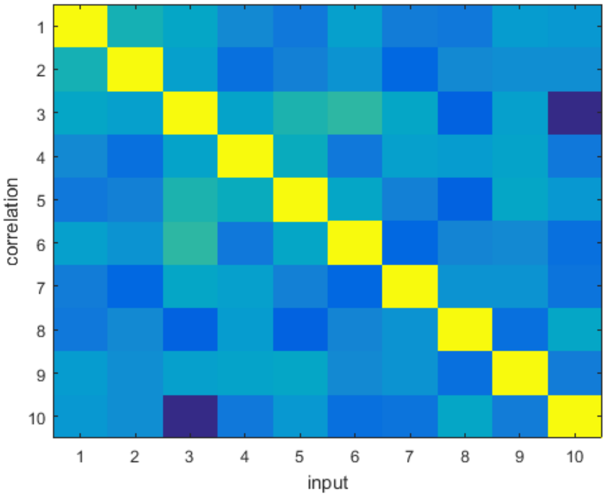


Fig 2: Correlation

Results:

1. Found correlation matrix to show appreciable similarities with beta.

Expectations

It was expected that both the vectors will show similarities. Ideally, as we want Ycap to be same as Y, both beta and correlation vector will be same as well. As Ycap = X*beta & Y = X*tau, tau is correlation vector. In the plots below, we have plotted tau and beta and compared the values of two against the feature.

1. In first plot we can see that the graph is similar but the output could have been better. This is because tau is standardized while beta is not. Change in values can add on weight to one feature while reduce the other features weight.
2. Second plot is the standardized plot where feature values are subtracted and divided by mean. Here, we get values between 0 and 1 but accuracy could have been better.
3. In third case, we have just normalized the columns. It helps but not much.
4. In the fourth case, every element is properly normalized according to the total mean. Thus, we could see significant similarities between two plots.

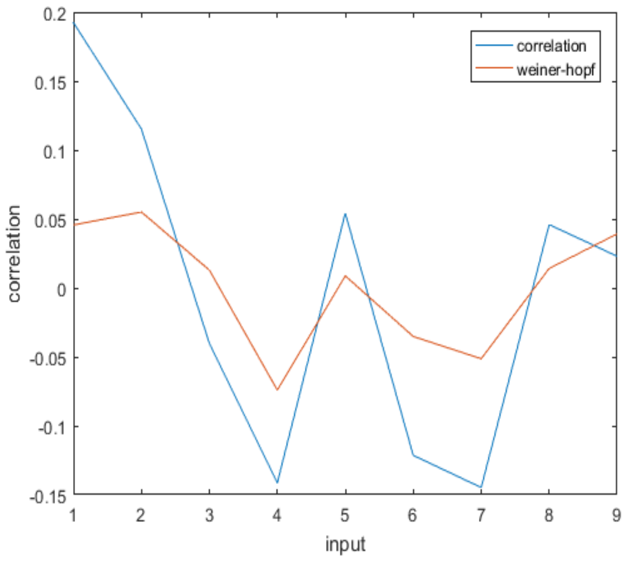


Fig 3: Normal output

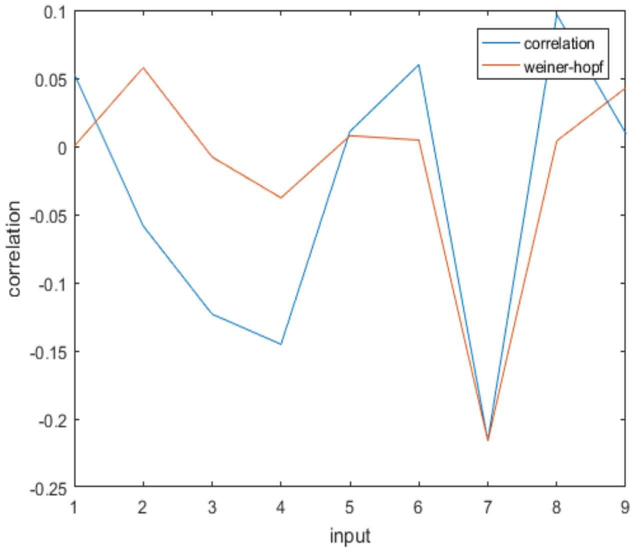


Fig 4: Standardized output

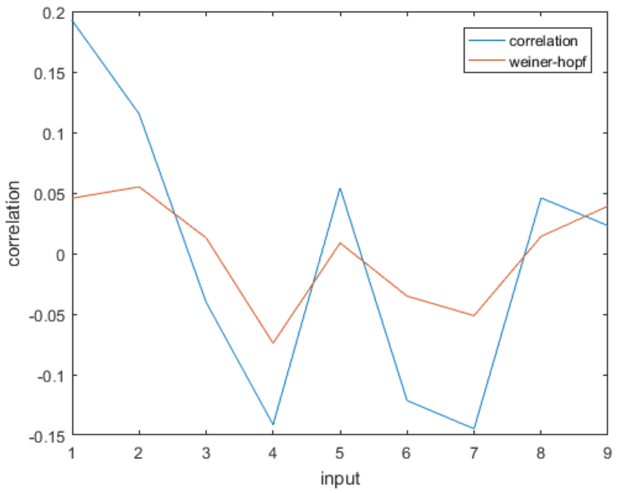


Fig 5: Standardized output

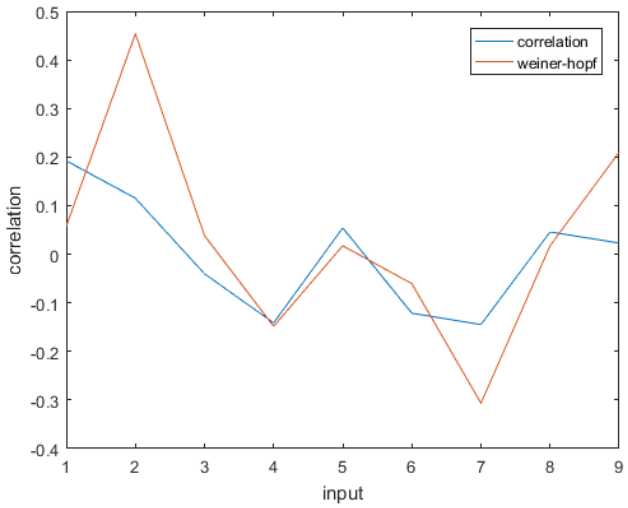


Fig 6: Standardized by column

Task III

In this task, we were supposed to move the decision boundary between 0 and 1 and plot the ROC. To obtain the result, I took following steps :-

1. Found out Yf from X and beta and made a scatter plot.

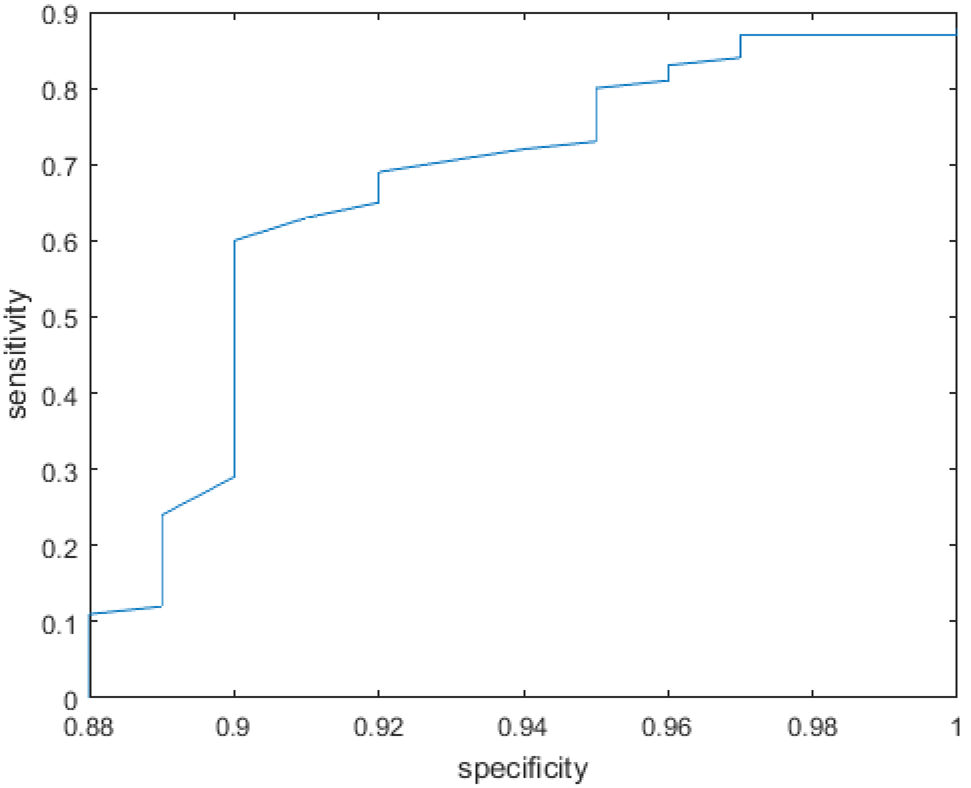
2. Initialized a variable xb to iterate from 0 to 1 in steps of 0.01 for having large number of samples.
3. Initialized two loops to go through boundaries and samples and calculated TN, TP, FP and FN.
4. According to the formula, calculated sensitivity and specificity for each iteration in outer loop
5. Then plotted sensitivity vs 1-specificity.

Results

1. For this data, we didn't achieve 100% sensitivity and the max we achieved was 88%. We achieved 100% specificity for some cases.

Expectation

It was expected that we won't achieve 100% sensitivity as we were not able to achieve a good match between correlation and beta. Thus, the values of Ycap and Y were not very close which could add on to the inaccurate classifier. This might have happened due to under-sampled 1s in the data. We had a lot of class 0 instances but very less class 1 instances. Thus, all the values are closer to 0s and it's difficult to decide a good decision boundary for this kind of problem. If the data had a large number of samples and also equal instances of both the classes then, probably the results would have been better. ROC is plotted below.


sensitivity
specificity