Report 4 on Naive Bayesian Classification Approach in Healthcare Applications

Authors:
Sanskar Gupta 2019ME20922
Daksha Kurhade 2019ME20894

Aim

In this report we have decided upon a number of things after report 2 and report 3. In report2, we extracted relevant data for this project from Kaggle. (the reference is given below). The data was not available with the research paper originally. Then once we had extracted the data, the next aim was to process it so that it could be used further. The processing of data further involved a number of steps. Some of them included things like computing averages or calculating attributes from a very large gene which was there in the data. We also had to extract smaller sets of data from this very large set of data we had so that we could do curve fitting/ Naive Bayes Classification on those smaller sets of data which were more relevant to this project. After data processing the next aim is curve fitting of data/ Naive Bayes classification in which we are supposed to find the relation between available data and whether the cancer type is ALL or AML. In report 2, we had tried to complete the data processing part but we were not able to come to the results we had desired. However, in report 3, We were successfully able to mine data relevant to our report from the large data set. In report 3, we used MS Excel algorithms to find out the average, standard deviation and thus compute S2N and and T value for all the genes. Based on that, we were able to find out the relevant gene set based on which we had to further classify the patients into ALL or AML cancer types. Till report 3, we were complete with the mining part. Now, in this final report we present you the classification along with the previous results we obtained.

Data Set

The data set was found out on a platform known as 'Kaggle'. The data set was that of gene expression which could be easily used to classify Cancer types based on the molecular genes present in the body.
The data set basically had three files. In the first file we had patient ID and the corresponding cancer type whereas in the second file and in the third file we had data of particular genes of 72 patients which were supervised for each patient we had data for 7129 genes present in the body. This is a very large data set and hence it had to be preprocessed so that relevant attributes could be taken out and curve fitting could be done on them.

Earlier in report 2, we tried to import the file into python and do the pre-processing. But, we could not succeed and the results obtained were not up to the expectation levels. Therefore, we dedicated ourselves to starting afresh using EXCEL. So, that proper results could be obtained.

Below are the steps undertaken to mine the data, to pre-process it to obtain relevant data which could be used for classification.

1. First of all we merged the two files to get one file with data of all 72 patients off 7129 genes. We stored that data in the Excel workbook named 'Initialdata.xlsx'
2. We then removed the "Gene Description" column first, and changed the "Gene Accession Number" to "ID"
3. We then removed all the columns with the heading of 'CALL'. These columns had redundant data which would not be used further.
4. We then added a row by the name of 'Class'. In this row, we stored the cancer types of all the patients- whether it was ALL or AML.
5. We then calculated various parameters of sum, number of entries, average, standard deviation for both ALL and AML types.
6. Basically, all the gene entries were not useful to us, Entries less than 20 and greater than 16000 were of no use. This was mentioned in the research paper. We calculated the sum of all the gene entries (which were greater than 20 and less than 16000) for a particular gene ID for all the patients which were ALL. This was

stored in Sum(ALL). The number of such entries which were added was stored in Number(ALL). By dividing these two, we computed Average(ALL). If the number(ALL) was zero, then average was assigned zero. We also calculated the standard deviation of the relevant entries (>=20 and <=16000) and stored them in Std. Deviation (ALL). For the case when Number was zero or one, the standard deviation shows #Div/0!. It should not be considered as an error. The entire process was done for AML also. The excel file 'Initialdata.xlsx' shows all this.

7. Then we calculated two values- S2N value and T Value. The formula for the both is given below.

$$S2N\ Value = (Avg1 - Avg2)/(Std1 + Std2)$$

$$T\ Value = (Abg1 - Avg2)/(((Std1 * Std1)/N1) + ((Std2 * Std2)/N2))$$

where
Avg1=Average(ALL), Avg2=Average(AML),
Std1=Standard deviation (ALL), Std2=Standard deviation (AML),
N1= Number (ALL), N2= Number (AML)

8. These S2N and T-value were calculated for all the genes. When N1 or N2 were zero or one, the values were taken to be zero.
9. Then , the genes were sorted on the basis of S2N value. The data is now stored in 'finaldata.xlsx'. Sheet 1 stores the data sorted according to S2N value. The top 50 S2N value genes were taken out and stored in the Sheet 2.
10. In the 'finaldata.xlsx', in sheet 3, Data was sorted according to the T-value. The top 50 T-value genes were taken out and stored in sheet 4.
11. In sheet 5, We have stored the genes which were common to sheet2 and sheet 4. Basically the genes which are stored in sheet 4 and sheet 5 are relevant genes which can be used to classify the cancer type of patients into ALL or AML. We have reduced the number of genes from 7129 to 19 relevant genes. We would further reduce the number of genes in our classification algorithm ahead in our next report. For data visualisation, which we have today in this report we have taken out two sets of genes and drawn graphs corresponding to them to realise the ALL and AML Cancer types on graphs/scatter plots in different colours.

Basically, whatever we did above was microarray data cleaning in which we try to mine data. We tried to clean data and obtain relevant data from a very-very large data set.

After data pre-processing, we took two genes at a time and Drew 3 graphs for different sets of genes and clustered them into ALL or AML cancer types. The main aim of this activity was data visualisation and trying to visualise the kind of results which we would get up on classification. We have attached the code in our files. The readme files are present. We also have attached the graphs for reference.

Moving on to classification and to find out the further results

We would now curve fit this data. We would basically try to find out the relationship between the type of cancer and the gene attributes which are present in the data. We will find out the probabilities, do relevant computation and finally do naive bayes classification. After curve fitting and finding the relation between the type of cancer and the data set we would test this data again and try to find out any discrepancy in our model. So by doing this curve fitting we basically have divided data into two classes based on gene attributes. Using this model if gene attributes are given to us we will be able to calculate S2N value and T value from those attributes then we can run those attributes on a data set and will predict what type of cancer the patient has whether it's ALL or AML class.

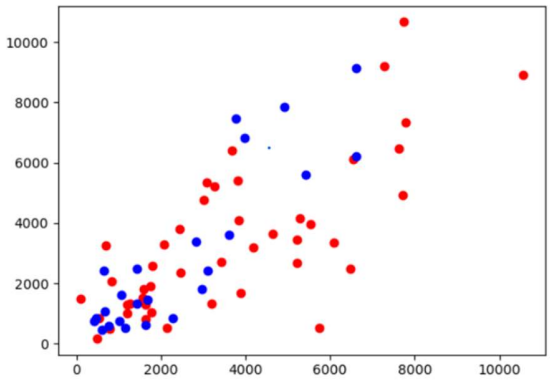
The procedure which we followed to move ahead after data mining is as follows:

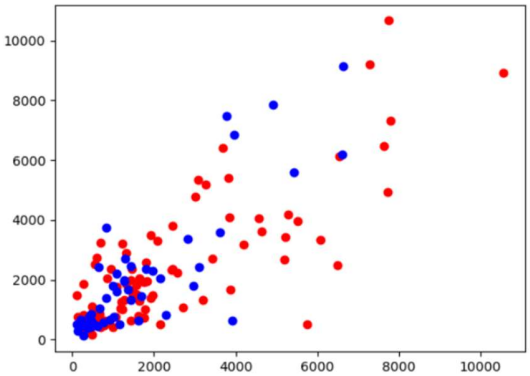
1) We were left with 19 genes in sheet 5 of our finaldata.xlsx. Now these are just like the parameters on which the cancer type depends. We had made scatter plots showing the ALL and AML distribution using 6 genes, the genes which were the most representative of the population. We decided to pick 3 genes out of these and make our model based on these gene sets. It would also be easy for us to compute the posterior and to classify later the patients.

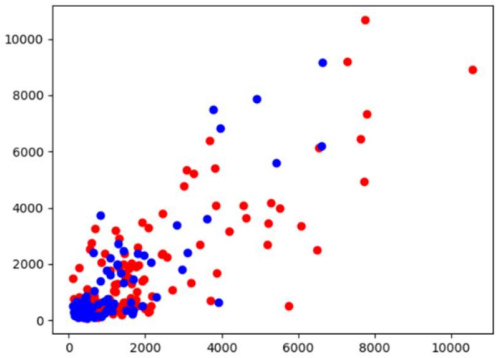
2) We chose 3 genes. For each gene, we had gene values of 72 patients. We at first bifurcated these to obtain the gene values corresponding to the ALL class and the AML class. There were 47 ALL patients and 25 AML patients. We basically had six sub datasets, one for ALL, one for AML, each for three genes. Then we did curve fitting on to this data. According to the central limit theorem, these data sets should correspond to graphs of normal distribution. We curve fitted the datasets and plotted the graphs of the normal distribution. We drew three graphs. In each graph, we had two subgraphs superimposed on each other. We had each graph for each gene. So, for a gene, we drew a graph, in which we had two subgraphs which represented the normal distributions corresponding to the ALL and the AML class. In short, we drew two normal distributions, one for ALL class, one for AML class, we superimposed them together into one graph. This was for one gene. We had to do this for all of them and thus made three graphs which further had two subgraphs each.
3) The main motive of this graph was to compute the likelihood so that we could compute the posterior. What we did for computing the likelihood, is that we took a particular patient gene data (the patient which we had to test for), we took the gene values of the patient as input to the three graphs we had made. The corresponding output from the graph is what is called the likelihood. We obtained one likelihood value for each gene for each ALL and AML class. This led to a total 6 likelihood values, 3 for ALL and 3 for AML. Then we multiplied the three likelihoods of the ALL to the prior for the ALL function. Multiplying the prior to likelihoods gave me the posterior. So, we calculated the posterior ALL and the posterior AML. Now, whichever posterior was more, our patient belonged to that class.
4) We had a problem which was underflow. There was a chance that likelihood values were very very low. So, we took the logarithm of the entire equation to obtain the log of posterior as the sum of the log of the prior and the log of the likelihoods. This helped us to easily compare the posterior values.
5) We also tested our model on three patients to compare the results.
6) The graphs have been attached. Also, all the code contains the training of the model by computing the likelihoods and curve fitting. It also contains the testing which has been carried out on the patients.

Graphs drawn

- Scatter Plots

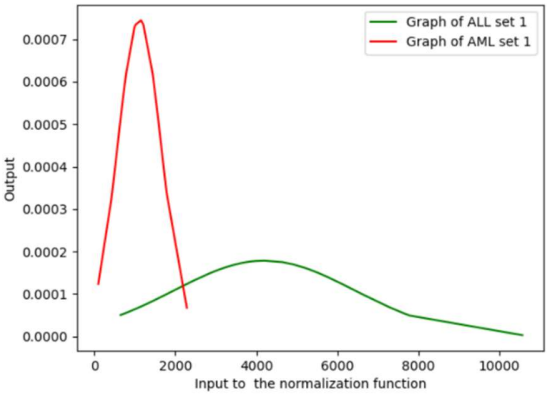




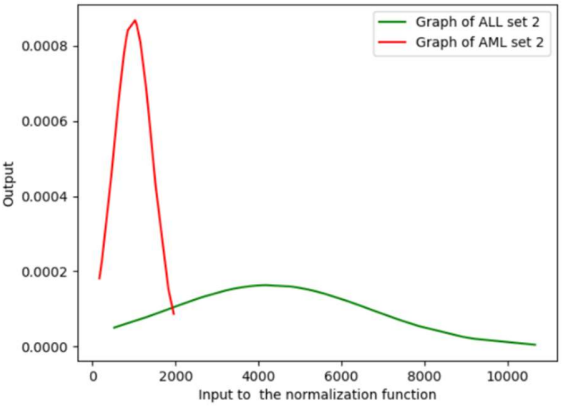


- Normal Distributions

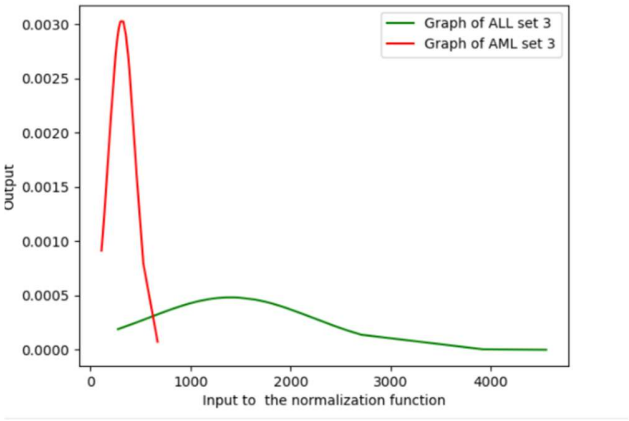
1. Graph of Set-1



2. Graph of Set-2



3. Graph Of Set-3



Code snippet showing result passed for all the patients

```
Test PASSED for case1, we desired for AML and got AML!
The Final Posterior of ALL in Case1: 1.3635323128761285e-12
The Final Posterior of AML in Case1: 1.5547025298647543e-11

Test PASSED for case2!, we desired for ALL and got ALL!
The Final Posterior of ALL in Case2: 4.431848361998773e-12
The Final Posterior of AML in Case2: 5.060864421217519e-19

Test PASSED for case3!, we desired for AML and got AML!
The Final Posterior of ALL in Case3: 6.533141833052343e-13
The Final Posterior of AML in Case3: 1.709050485901564e-10

Process finished with exit code 0
```

Conclusion and Discussion

Thus we have completed our project. We have scatter plotted the data to visualise the two classes. We also have drawn normal distribution function graphs for the three genes we have used. We ran our model on two patients- two of AML and one of ALL. We achieved 100% efficiency in our model. The patients got classified into their respective models. The posterior computed for those patients gave their results accurately.

We would also like to say that this classification is naive (since we have used naive bayes classification). In this all the gene parameters are assumed independent of each other. They may not be in reality. But the scope of the project says that we will focus only on naive bayes classification which we have done. We also have achieved 100% efficiency in our model. Which is a good model.

Statement of contribution: Two members of our team, Daksha Kurhade and Sanskar Gupta have worked together to compile everything in this report, and for coding both of them have collaborated together and have done the entire work together. Discussions, extra work done while collaboration on MS teams, the code, the excel sheets the report, and all effort in this report has been done together. The PPT and video also have been made together by both of us.

As informed in the previous report as well, the third member is no longer continuing with us. Daksha and Sanskar have worked together equally in this report and also in the previous reports as well.

Statement of contribution: Two members of our team, Daksha Kurhade and Sanskar Gupta have worked together to compile everything in this report, and for coding both of them have collaborated together and have done the entire work together. Discussions, extra work done while collaboration on MS teams, the code, the excel sheets the report, and all effort in this report has been done together. The PPT and video also have been made together by both of us.

As informed in the previous report as well, the third member is no longer continuing with us. Daksha and Sanskar have worked together equally in this report and also in the previous reports as well.

References

Dataset downloaded from here
https://www.kaggle.com/crawford/gene-expression?select=data_set_ALL_AML_train.csv