GDA - IRM2016502

Gaussian discriminant analysis (GDA) is a generative model for classification where the distribution of each class is modeled as a multivariate Gaussian. In this approach we try to model p(x|y) and p(y) as oppose to p(y|x) we did earlier, it’s called Generative Learning Algorithms. Once we learn the model p(y) and p(x|y) using training set, we use Bayes Rule to derive the p(y|x) as

Here below are the parameters needed to be calculated

$$\phi = \frac{1}{m}\sum_{i=1}^{m} 1\{y^{(i)} = 1\}$$

$$\mu_0 = \frac{\sum_{i=1}^{m} 1\{y^{(i)} = 0\}x^{(i)}}{\sum_{i=1}^{m} 1\{y^{(i)} = 0\}}$$

$$\mu_1 = \frac{\sum_{i=1}^{m} 1\{y^{(i)} = 1\}x^{(i)}}{\sum_{i=1}^{m} 1\{y^{(i)} = 1\}}$$

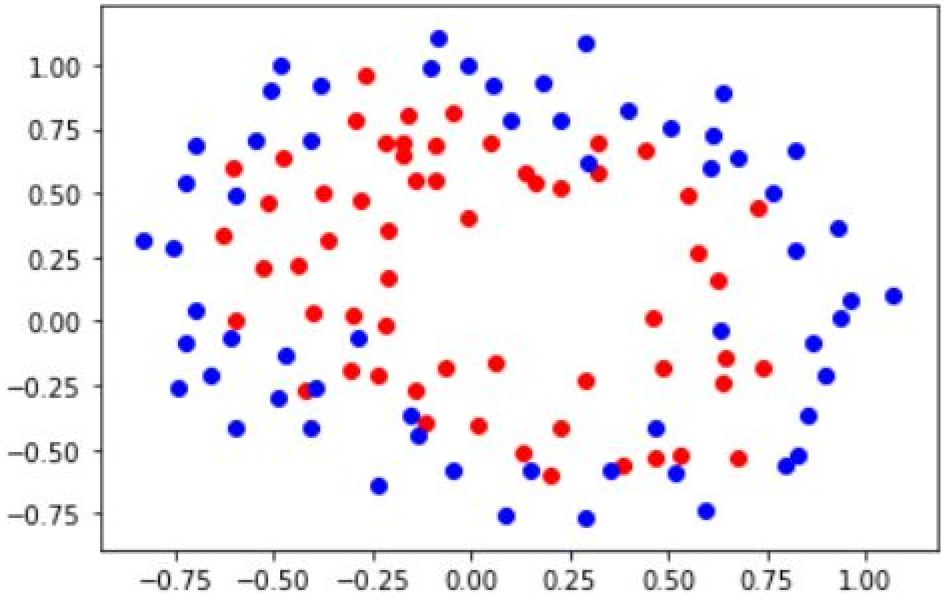
$$\sum = \frac{1}{m}\sum_{i=1}^{m} (x^{(i)} - \mu_k)(x^{(i)} - \mu_k)^T \quad \text{where } k = 1\{y^{(i)} = 1\}$$

So the model can be defined as

$$p(y) = \phi^y(1-\phi)^{(1-y)}$$
$$p(x\,|\,y=0) = \frac{1}{(2\pi)^{n/2}|\sum|^{\frac{1}{2}}} exp(-\frac{1}{2}(x-\mu_0)^T \sum^{-1} (x-\mu_0))$$
$$p(x\,|\,y=1) = \frac{1}{(2\pi)^{n/2}|\sum|^{\frac{1}{2}}} exp(-\frac{1}{2}(x-\mu_1)^T \sum^{-1} (x-\mu_1))$$

Here the parameters of the model are ϕ, μ_0, μ_1 and $\sum$.
And n is the dimension of the density function
Note : While there are two separate mean vectors μ_0 and μ_1(each for one class), but the covariance matrix $\sum$ is common for all the classes.

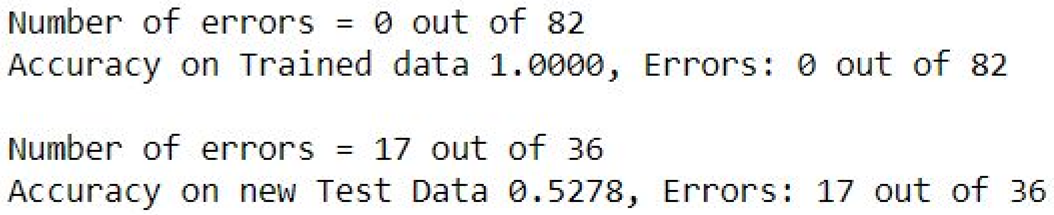
Using microchip data, split 70% for training and 30% for testing.
Here is the scatter plot of the raw data.



Microchip data from the drive shared.

We can observe that the data is not gaussian within their classes. Hence we cannot expect GDA to work well on this data.

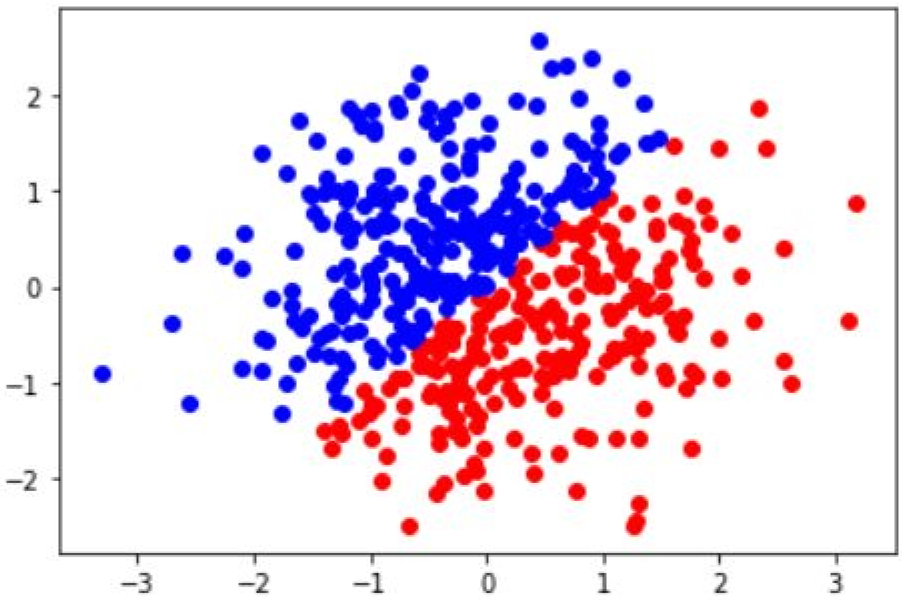
Below are the Test results.



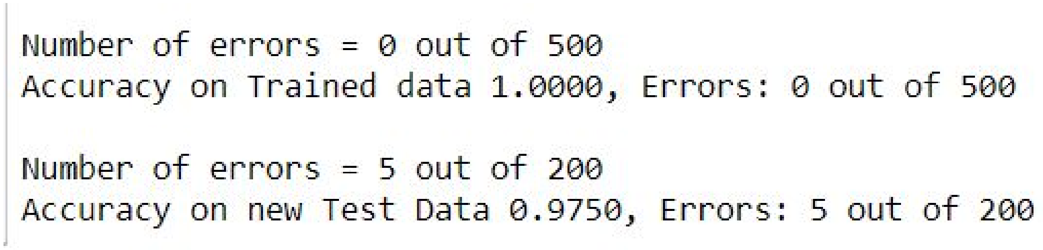
```
Number of errors = 0 out of 82
Accuracy on Trained data 1.0000, Errors: 0 out of 82

Number of errors = 17 out of 36
Accuracy on new Test Data 0.5278, Errors: 17 out of 36
```

After trying with the given raw data I've generated sufficient Gaussian data for both the features and classes and then applied GDA.

The below is the scatter plot of randomly generated data which follows Gaussian distribution.



And here are the results of training and testing.



```
Number of errors = 0 out of 500
Accuracy on Trained data 1.0000, Errors: 0 out of 500

Number of errors = 5 out of 200
Accuracy on new Test Data 0.9750, Errors: 5 out of 200
```

From Above two results we could observe that the data must be distributed according to gaussian distribution so as to achieve good results.