Soft Computing Assignment

GDA implementation using Raw Data and Normal distributed data-set

Md Afzal Ansari(MIT2019072)

1 DATA-SET

The data-set consist of two features of a microchip. Based on these two features either the chip is accepted or rejected.

Dimensions : 118 X 2

```
            x1        x2  y
0    0.051267  0.699560  1
1   -0.092742  0.684940  1
2   -0.213710  0.692250  1
3   -0.375000  0.502190  1
4   -0.513250  0.465640  1
..        ...       ... ..
113 -0.720620  0.538740  0
114 -0.593890  0.494880  0
115 -0.484450  0.999270  0
116 -0.006336  0.999270  0
117  0.632650 -0.030612  0

[118 rows x 3 columns]
```

The labels in the dataset is given in the form of 0 and 1.
Here, 0 means rejected
1 means accepted.

2 Hypothesis

When we have a classification problem in which the input features are continuous random variable, we can use GDA, it's a generative learning algorithm in which we assume p(x—y) is distributed according to a multivariate normal distribution and p(y) is distributed according to Bernoulli. So the model is given by:

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

3 Cost Function

We need to define the log likelihood function L and then by maximising L with respect to model parameters, find the maximum likelihood parameters.

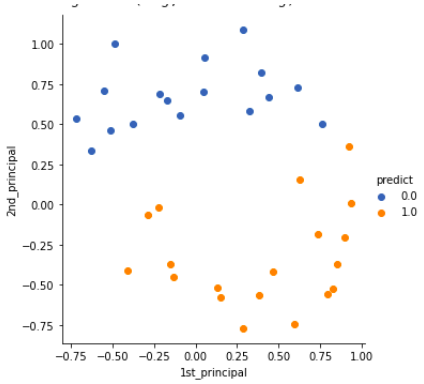
$$\ell(\phi, \mu_0, \mu_1, \sum) = \log \prod_{i=1}^{m} p(x^{(i)}, y^{(i)}; \phi, \mu_0, \mu_1, \sum)$$
$$= \log \prod_{i=1}^{m} p(x^{(i)}|y^{(i)}; \phi, \mu_0, \mu_1, \sum) p(y^{(i)}; \phi)$$

4 Comparision

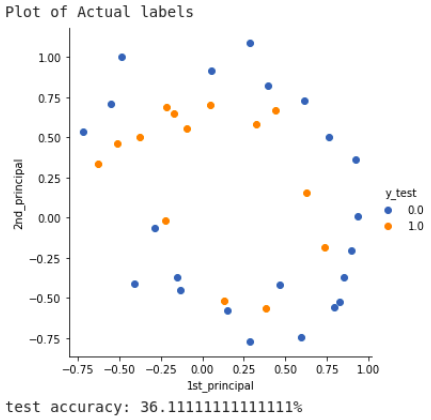
Implementation of GDA from scratch using Normal distributed data using Box-Muller transformation

The scatter plot obtained for the test data is used to compare the result.

- The scatter plot for the test data and predicted label is shown below:



- The scatter plot for the test data and actual label is shown below:



- The Accuracy obtained for the classification of the test data is 36.11

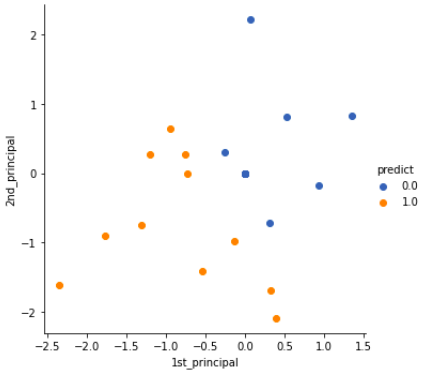
- The classification of the data is poorer than the classification result obtained on Normal distributed data.The Accuracy fallen to about 10 percent on using raw data-set.

- On multiple runs the accuracy increased up to 52 percent in correctly classifying the microchips.

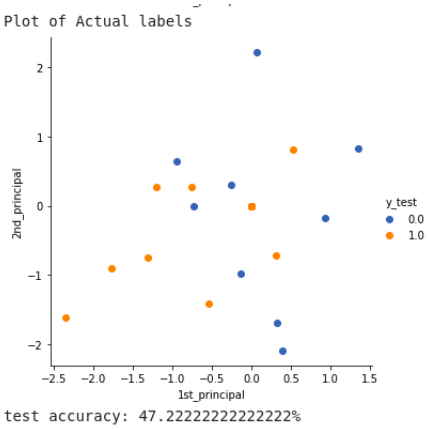
Implementation of GDA from scratch using Raw data

The scatter plot obtained for the test data is used to compare the result.

- The scatter plot for the test data and predicted label is shown below:



- The scatter plot for the test data and actual label is shown below:



- The Accuracy obtained for the classification of the test data is 47.22

- The classification of the data is better than the classification result obtained on raw data.The Accuracy increased to about 10 percent on using Normal distributed data-set.

- On multiple runs the accuracy increased up to 66 percent in correctly classifying the microchips.