Ionosphere Structure Detection

Vaibhavi Lokegaonkar

December 2021

1 Determining whether there exists a structure in the ionosphere or not (Classification Problem)

The dataset given to us contains parameters to check whether there's some structure in the ionosphere or not. We have been given two labels: 'g' denoting that there's a structure and 'b' denoting that there's no structure.

1.1 Data Exploratory Analysis

For the univariate and multivariate guassian classifier, we need to select columns from the ones given in the dataset for classification. We can take those columns with a very high correlation to the label column as shown in the correlation matrix below:

Correlation Matrix

The columns 0, 2, 4, 6 have a higher correlation (= 0.5) with the column which contains the classification.

1.2 Data Preprocessing

1.2.1 Missing Values

There are a total of about 15 to 20 missing values in the dataset. The total dataset size is 351. We would not lose a lot if we delete the missing valued rows, since they might be even 10. Hence, we delete all rows containing missing values.

1.3 Train test split

We shuffle the dataset and allocate the first 80% of the data as training and the rest of the data as testing data.

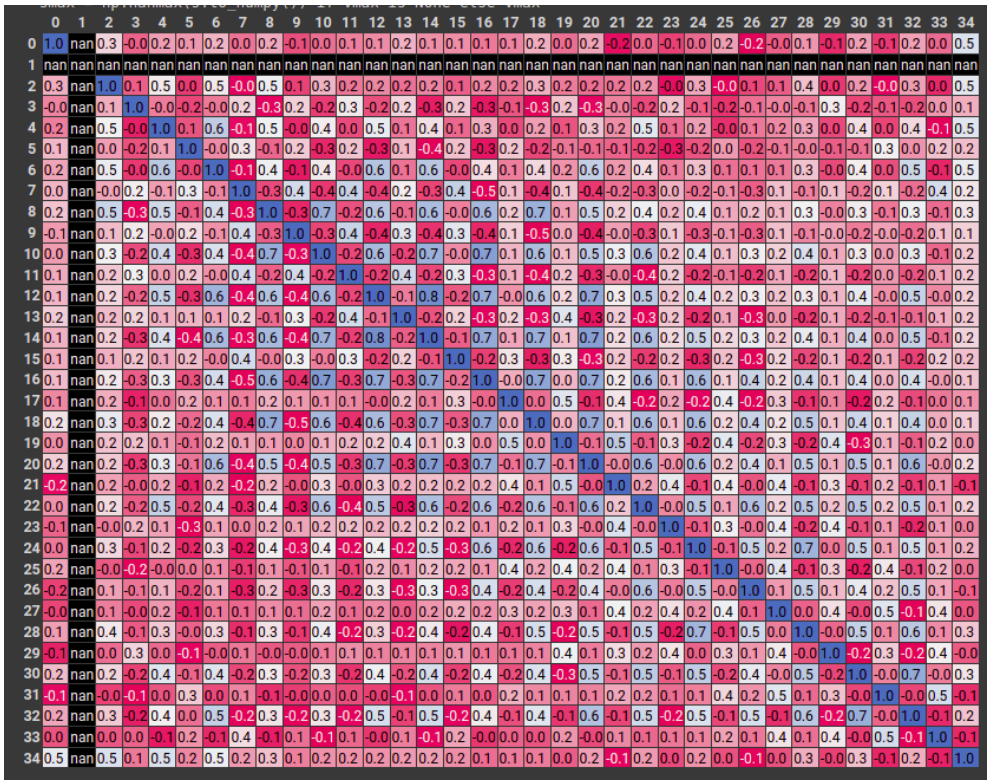

Figure 1: Correlation Matrix for Ionosphere Dataset

1.4 Learning Algorithm

Logistic Regression

We used the following function as the gradient of the loss function in Gradient Descent, obtained after performing maximum likelihood estimation on the probability of a class given the features. This is equal to the $\sigma(w^T x)$.

$$\sum_{n=1}^{331} x_i(\sigma(w^T x) - y_i)$$

We further apply gradient descent and newton's method to converge on a weight matrix.

We calculated the following expression for hessian$[i][j]$:

$$\sum_{n=1}^{331} x_{ki} x_{kj}$$

Univariate Gaussian Classifier

In addition, we have implemented the Univariate Gaussian Classification Method. We observed all the possible accuracies that it gave by using columns 0, 2, 4, 6. In some cases

we also got a zero division error. This occured only if the standard deviation that we got from the data was zero. This may happen, since we are randomly shuffling the data and then splitting it. The accuracies that we obtained are shown in the notebook.

1.5 Multivariate Gaussian Classifier

We can use multivariate Gaussian classifier for classifying the data. We could use columns: 0, 2, 4, 6 as they showed the highest dependency (correlation) on the classification columns. I further added more columns to the dataset for this model, to increase the accuracy. They had correlation in the range of 0.2 - 0.3.

1.6 Results: Accuracy on train and test sets

While implementing the above methods, we came across matrices x which were not invertible. In such cases, a small multiple λ of the identity matrix of the same size as that of the matrix is deducted, so as to get a invertible matrix. Note that, λ is in the order of 10^-5 or 10^-6. Due to it's small value the subtraction has a very little effect on the accuracy.

Optimization Method	**Train set accuracy**	**Test set accuracy**
Univariate Gaussian	0.8143	0.8030
Logistic Regression Gradient Descent	0.8787	0.8939
Logistic Regression Newton's Method	0.8939	0.8484

Table 1: Accuracy of the test and train set