**Logistic Regression**

SCM 516: Final Project

Team – 9

*Submitted by – Aman Kumar(akuma185), Sejal Kulkarni(sskulk18), Eshwar Chandra(echandr1), Xueli Zhang(xzhan480)*

Abstract

We worked on a two-class classification problem for Logistic regression, we worked on a data set for Breast Cancer Prediction which we obtained from UCSIS website. We used Stat Tools to analyze the dataset. Here we had to predict whether a person has cancer or not (Malign or Benign) based on parameters like mean radius, mean texture, mean perimeter, etc. Our aim was to classify the Malignant cases correctly and reduce the number of Malignant cases being classified as Benign by the algorithm.

Exploratory Data Analysis

Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image. Number of instances are 569, number of attributes present in the dataset are 32, there are two types of diagnosis (M= Malignant, B = Benign). The data breakdown is Benign: 357 (65.5%) and Malignant: 241 (34.5%).

The mean, standard error, and "worst" or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features. All feature values are recoded with four significant digits. Missing attribute values: none

Analysis, Output and Interpretation

We have divided the features into 3 parts for doing the analysis, we first ran logistic regression by selecting the mean measurements, second iteration we did with standard error values, and third iteration we took the worst measurement values. We will discuss about each iteration and findings in the following sections of our report.

**Iteration 1: Using Mean Measurement Values**

Target = 7.377 + 2.046(MeanRadius) - 0.384(Meantexture) + 0.071(MeanPerimeter) - 0.039(MeanArea) + 76.412(MeanSmoothness) + 1.474(MeanCompactness) - 8.463(MeanConcavity) -66.837(MeanConcavePoints) - 16.281(MeanSymmetry) + 68.188(MeanFractalDimension).

|  |  |  |  |
| --- | --- | --- | --- |
| Classification Matrix | Malignant | Benign | Percent Correct |
| Malignant | 347 | 10 | 97.2% |
| Benign | 19 | 193 | 91.04% |

If we use only the mean values for the prediction whether the person being diagnosed has breast cancer or not we get a total of 94.9% correct predictions. Out of 357 Malignant cases 347 have been identified correctly, since we are concerned about malignant cases predicted properly, we will look at the malignant cases that have been classified as benign which in this case are 19. So, to try reducing that we changed our predictor features in the next iteration.

**Iteration 2: Using Standard Error Values**

Target = 3.392 + 28.243(RadiusError) + 0.3344(TextureError) - 0.076(PerimeterError) - 0.393(AreaError) - 4.520(SmoothnessError) - 57.066(CompactnessError) - 3.623(ConcavityError) - 39.661(ConcavePointsError) + 6.781(SymmetryError) + 317.289(FractalDimensionError)

|  |  |  |  |
| --- | --- | --- | --- |
| Classification Matrix | Malignant | Benign | Percent Correct |
| Malignant | 342 | 15 | 95.80% |
| Benign | 42 | 170 | 80.19% |

For this iteration we tried taking the standard error in the values for the prediction of whether the person being diagnosed has breast cancer or not, and we got a total of 89.98% correct predictions, lower when compared to the correct predictions we got in our first iteration. Here, out of 357 Malignant cases 342 have been identified correctly, but we have 42 cases which are classified as Benign which originally were Malignant. Since the number of false negatives in this case is too much we cannot take this logistic regression model as significant compared to our first iteration.

**Iteration 3: Using Worst Measurement Values**

Target = 29.005 + 0.534(WorstRadius) - 0.282(WorstTexture) - 0.013(WorstPerimeter) - 0.018(WorstArea) - 53.932(WorstSmoothness) + 8.316(WorstCompactness) - 4.581(WorstConcavity) - 37.537(WorstConcavePoints) - 9.621(WorstSymmetry) + 7.867(WorstFractalDimension)

|  |  |  |  |
| --- | --- | --- | --- |
| Classification Matrix | Malignant | Benign | Percent Correct |
| Malignant | 353 | 4 | 97.2% |
| Benign | 7 | 205 | 91.04% |

In this iteration we have used the worst values of the three values for each measurement for the prediction and when we built the logistic regression model taking this data to determine our target it gave a 98.07% accuracy. Out of 357 Malignant cases 353 were identified correctly and the false negatives i.e. the Malignant being classified as the benign dropped to a count of 7.

**Conclusion**

After three iterations with different set of features we can say that the best set of predictors out of our iterations is when we predict the situation of cancer based on “worst measurement” measures, which not only had a better accuracy rate but also the Malignant cases being classified as Benign were reduced significantly compared to other iterations. But, the true performance of the model will be checked on how it performs on the data where the actual condition is not known.

**Sources**:

<https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)>

<https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_breast_cancer.html>