Dhaval Patel

COMSC 415

Prof. Sonya Cates

2/28/2019

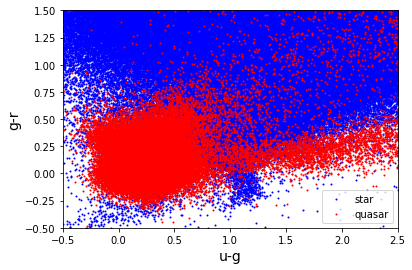
Programming Project 1: Classification of Stars and Quasars

**Objective**: For this project the objective is to build a classifier to label celestial objects as stars or quasars based on the 4 features u-g, g-r, r-i, i-z. These features are photometric data (colors of light).

**Part 1: Analyzing the data set**

We are using the SDSS (Sloan Digital Sky Survey) data set. The data set includes a total of 705,290 labeled objects with 4 features each. The labels are ‘star’ or ‘quasar’. The training data set consists of 505,290 of these objects with their labels and the test data set is the remaining 200,000 objects from the original data set.

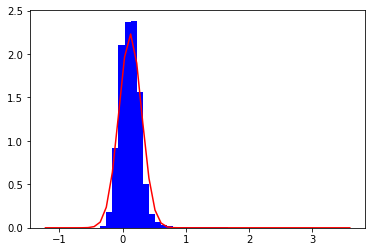
To start, the training data has plenty of objects to train the classifier we choose. Visualizing the data, we can see a strong overlap between the 4 features. For example, As the figure below shows, the data points between the u-g and g-r features are highly scattered. These issues might make it difficult to choose a good model to fit the data.



Next, the correlation coefficient was discovered between features. The table below shows a matrix. There is a high amount of correlation which means that assuming independence for these features for the Naïve Bayes might not be a good idea. The only low numbers are the correlation between the u-g/r-i and u-g/i-z features. This is probably because of the high amount of overlap seen in the data.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Features** | **u-g** | **g-r** | **r-i** | **i-z** |
| u-g | 1 | 0.354 | 0.155 | 0.174 |
| g-r | 0.354 | 1 | 0.737 | 0.687 |
| r-i | 0.155 | 0.737 | 1 | 0.81 |
| i-z | 0.174 | 0.687 | 0.81 | 1 |

Finally, bar graphs of features were examined to determine the distributions. All the distributions looked gaussian. For example, the r-i feature distribution of quasars class below shows a clear picture.



Actual Stars in Training Data: 430,827

Actual Quasars in Training Data: 74,463

An important thing to note here is that there are significantly more stars than quasars in the data and this causes the data to be unbalanced and therefore, accuracy is not the most reliable source for interpretation of the results.

**Part 2: Classification using Naïve Bayes**

Naïve Bayes was used to classify the data and the results are below. This model is not the best to use for this problem. This is because assuming independence is not ideal for this data. As stated above, the high amount of overlap makes the Gaussian Naïve Bayes too simple of a model. As seen in the results, the accuracy was only .617 and more importantly the number of false positives is way too high. Precision is also only .14.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Classes** | **Precision** | **Recall** | **F1-Score** | **Support** |
| 0 (star) | 0.99 | 0.59 | 0.74 | 186721 |
| 1 (quasar) | 0.14 | 0.95 | 0.25 | 13279 |
|  |  |  |  |  |
|  | Accuracy = .61724 | | | |

|  |  |  |
| --- | --- | --- |
| **Classes** | **Predicted Star (0)** | **Predicted Quasar (1)** |
| Actual Star (0) | 110858 True Positives | 75863 False Positives |
| Actual Quasar (1) | 689 False Negatives | 12590 True Negatives |

The goal at this point is to improve accuracy but also reduce the number of false negatives and false positives as accuracy alone is not the best interpretation.

Part 3: Increasing complexity by using Gaussian Mixtures

As Naïve Bayes is a simple model that cannot fit the data being used here, a Gaussian Mixture to increase the complexity of the model is used. Gaussian mixture allows the model to fit to more than just 1 gaussian distributions. After replicating the Naïve Bayes results, the gaussian mixture components(n) and covariance types are changed to better fit the model to the data.

The methodology used here is to first try different covariance matrices with a specific n and then increase the n, examining the results each time. The training data was split into a new training set of 350,000 and the remaining 155,290 objects in a validation set. The ratio of training to test/validation remains like the original data (roughly 70% for training).

In testing the different types of covariance matrices with n = 1, there was a significant improvement of using “tied’ and “full” covariance over the “diag”. When moving to a higher value of n, the data was considered carefully. There are 4 features but 5 distinct color profiles. The most likely guess was to use 5 gaussian components because this could significantly reduce the overlap between the features.

* n=1, diag (Validation)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Classes** | **Precision** | **Recall** | **F1-Score** | **Support** |
| 0 (star) | 0.98 | 0.77 | 0.86 | 132496 |
| 1 (quasar) | 0.4 | 0.91 | 0.56 | 22794 |
|  |  |  |  |  |
|  | Accuracy = .7878 | | | |

|  |  |  |
| --- | --- | --- |
| **Classes** | **Predicted Star (0)** | **Predicted Quasar (1)** |
| Actual Star (0) | 101597 True Positives | 30899 False Positives |
| Actual Quasar (1) | 2046 False Negatives | 20748 True Negatives |

* n=1, tied (Validation)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Classes** | **Precision** | **Recall** | **F1-Score** | **Support** |
| 0 (star) | 0.98 | 0.85 | 0.91 | 132670 |
| 1 (quasar) | 0.5 | 0.88 | 0.64 | 22620 |
|  |  |  |  |  |
|  | Accuracy = .8531 | | | |

|  |  |  |
| --- | --- | --- |
| **Classes** | **Predicted Star (0)** | **Predicted Quasar (1)** |
| Actual Star (0) | 112575 True Positives | 20095 False Positives |
| Actual Quasar (1) | 2708 False Negatives | 19912 True Negatives |

* n=1, full (Validation)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Classes** | **Precision** | **Recall** | **F1-Score** | **Support** |
| 0 (star) | 0.98 | 0.85 | 0.91 | 132512 |
| 1 (quasar) | 0.51 | 0.88 | 0.64 | 22778 |
|  |  |  |  |  |
|  | Accuracy = .8578 | | | |

|  |  |  |
| --- | --- | --- |
| **Classes** | **Predicted Star (0)** | **Predicted Quasar (1)** |
| Actual Star (0) | 113217 True Positives | 19295 False Positives |
| Actual Quasar (1) | 2786 False Negatives | 19992 True Negatives |

* n=5, diag (Validation)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Classes** | **Precision** | **Recall** | **F1-Score** | **Support** |
| 0 (star) | 0.99 | 0.99 | 0.99 | 132477 |
| 1 (quasar) | 0.91 | 0.92 | 0.92 | 22813 |
|  |  |  |  |  |
|  | Accuracy = .9755 | | | |

|  |  |  |
| --- | --- | --- |
| **Classes** | **Predicted Star (0)** | **Predicted Quasar (1)** |
| Actual Star (0) | 130512 True Positives | 1965 False Positives |
| Actual Quasar (1) | 1836 False Negatives | 20977 True Negatives |

As the results point out above, this method proved to be incredibly fruitful. The accuracy, precision, recall all went up a significant amount and the false positives/false negatives were reduced greatly. Next, the covariance matrices were changed. The “tied” covariance reduced the false positives but increased the false negatives while the “full” covariance reduced the false negatives but increased the false positives. These were minor changes. “diag” covariance was in between the other two with respect to the false positives/false negatives. For this reason, the n=5, diag model was chosen to be the final model.

* n=5, diag (Final Test results)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Classes** | **Precision** | **Recall** | **F1-Score** | **Support** |
| 0 (star) | 1 | 0.97 | 0.98 | 186721 |
| 1 (quasar) | 0.66 | 0.93 | 0.77 | 13279 |
|  |  |  |  |  |
|  | Accuracy = .9631 | | | |

|  |  |  |
| --- | --- | --- |
| **Classes** | **Predicted Star (0)** | **Predicted Quasar (1)** |
| Actual Star (0) | 180234 True Positives | 6487 False Positives |
| Actual Quasar (1) | 902 False Negatives | 12377 True Negatives |

The final test results were great. The accuracy is above 95% but the important things are false positives/false negatives numbers. The false negatives number is lower than expected and that is good because with the unbalanced data, the ratio of false classification of quasars to actual number of quasars needs to be lower. The precision score for class 1 took a hit but this is also largely because of the unbalanced data. All in all, the model chosen seems to be a great choice based on these results.