Classifying Celestial Bodies

Navid Mohseni, Sylvester Mensah, Ethan Corr

For our project, our research question is “From a set of scientific measurements, can we classify a celestial body as a star, galaxy, or quasar object?”. We aim to use predictors like ultraviolet filtered intensity, infrared filtered intensity, and redshift of light waves (from this Kaggle datatset from CERN) to build 2 classification models and compare their predictions on a holdout set. The two classification models we will build are (1) a multinomial logistic classifier and (2) a multiclass support vector machine.

Methods

Method 1: We will use a multinomial logistic regression model to estimate the log odds of the data point belonging to each class.

$$\eta_{ik} = \log\left(\frac{\pi_{ik}}{\pi_{iK}}\right) = \beta_{k0} + \beta_{k1}x_1 + \cdots + \beta_{kp}x_p = \mathbf{x}_i'\boldsymbol{\beta}_k$$

Method 2: We will use a multi-class Support Vector Machine to create 3 decision boundaries for the 3 classes, one decision boundary for each combination of 2 classes.

$$\begin{aligned} \max_{\boldsymbol{\beta},\beta_0,\epsilon_1,\ldots,\epsilon_n,M} \quad & M \\ \text{s.t.} \quad & \|\boldsymbol{\beta}\| = 1, \\ & y_i(\mathbf{x}'\boldsymbol{\beta} + \beta_0) \geq M(1-\epsilon_i), \\ & \epsilon_i \geq 0, \sum_{i=1}^{n} \epsilon_i \leq B, \ i = 1,\ldots,n, \end{aligned}$$

Data Exploration

The dataset contains 100k data points with 6 numerical predictors. The predictors are all characteristics of the light emitted from a celestial body. They are:

- Ultraviolet filter (in the photometric system)
- Green filter (in the photometric system)
- Red filter (in the photometric system)
- Near Infrared filter (in the photometric system)
- Infrared filter (in the photometric system)
- Redshift (based on increase in wavelength)

Sampling 1k points from each class, we performed Anderson-Darling tests on all 6 numerical predictors. For all 3 classes, none of the variables come from a normal distribution. Because of this, we avoid using Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA) due to the incompatibility of MVN distribution for each class.