Estimating the Probability of a Star Being a Hypervelocity Star Using Bayesian Classification

1 Introduction

Classifying stars based on their observed properties is fundamental in astrophysics, providing insights into stellar evolution and dynamics. Hypervelocity stars (HVS) are stars moving at velocities exceeding the typical escape velocity of the galaxy. Accurately identifying HVS requires careful statistical analysis, especially when considering measurement uncertainties inherent in astronomical observations.

This report presents a Bayesian classification approach to estimate the probability that a star with given observed features and associated measurement uncertainties is a hypervelocity star. We detail the statistical methodology, including the incorporation of measurement uncertainties into the model, and provide a Python implementation of the calculation.

2 Data and Problem Description

2.1 Features, Labels, and Measurement Uncertainties

- **Features:**
  - X: The $bp - rp$ color index of the star.
  - Y: The absolute magnitude of the star (G_{mag}).
- **Measurement Uncertainties:**
  - σ_x: Uncertainty in the $bp - rp$ color index. Typically defined by the uncertainty in the distance that propagates to the reddening, since we do not know if the star is behind or in front of the extinction layer
  - σ_y: Uncertainty in the absolute magnitude. This is typically large given by the implied distance uncertainty, which in itself is derived from the proper motion errors.
- **Labels:**
  - is_hvs: A boolean indicator where:
    - True: The star is a hypervelocity star.
    - False: The star is not a hypervelocity star.

Given a new star with observed features $(x_{\text{obs}}, y_{\text{obs}})$ and measurement uncertainties (σ_x, σ_y), our goal is to compute the probability that it belongs to each class (is_hvs = True or False). This involves updating our beliefs about the class membership based on both the observed data and the measurement uncertainties.

3 Statistical Methodology with Measurement Uncertainties

To estimate the probabilities, we use Bayesian classification, which combines prior knowledge with observed data to update the probability of a hypothesis. In our case, the hypothesis is whether the star is an HVS. The incorporation of measurement uncertainties is crucial, as it allows us to account for errors in the observed features, leading to more accurate and reliable probability estimates.

3.0.1 Step 1: Calculating Prior Probabilities

The prior probability represents our initial belief about the likelihood of each class before considering the observed data. It reflects the class distribution in the dataset and provides a baseline for our Bayesian inference.

For each class c (*True* or *False*):

$$P(\text{Class} = c) = \frac{N_c}{N_{\text{total}}}$$

where:

- N_c is the number of stars in class c.
- N_{total} is the total number of stars in the dataset.

By calculating the priors, we quantify the proportion of stars in each class, which influences the posterior probabilities.

3.0.2 Step 2: Computing Mean Vectors and Covariance Matrices

The mean vector and covariance matrix characterize the distribution of features within each class, essential for modeling the class-conditional probability density functions (PDFs). They encapsulate the central tendency and variability of the features.

Mean Vector ($\boldsymbol{\mu}_c$) For class c:

$$\boldsymbol{\mu}_c = \frac{1}{N_c} \sum_{i=1}^{N_c} \mathbf{x}_i^{(c)}$$

where $\mathbf{x}_i^{(c)} = \begin{bmatrix} X_i^{(c)} \\ Y_i^{(c)} \end{bmatrix}$ is the feature vector of the i-th star in class c.

The mean vector represents the average position of the stars in feature space for each class, indicating where the data is centered.

Covariance Matrix (Σ_c)

$$\Sigma_c = \frac{1}{N_c - 1} \sum_{i=1}^{N_c} \left(\mathbf{x}_i^{(c)} - \boldsymbol{\mu}_c \right) \left(\mathbf{x}_i^{(c)} - \boldsymbol{\mu}_c \right)^\top$$

The covariance matrix captures the variability and correlation of the features within each class. The diagonal elements represent the variance of each feature, while the off-diagonal elements represent the covariance between features.

3.0.3 Step 3: Adjusting Covariance Matrices

Covariance matrices must be positive definite to ensure that the multivariate normal distribution is properly defined and invertible. Issues such as small sample sizes or multicollinearity can lead to singular or near-singular matrices, causing computational problems.

Adjustment Method We add a small constant ϵ to the diagonal elements:

$$\Sigma_c' = \Sigma_c + \epsilon I$$

where:

- I is the identity matrix.
- ϵ is a small positive constant (e.g., 1×10^{-6}).

This adjustment ensures all eigenvalues of Σ_c' are positive, making the matrix invertible and suitable for use in the multivariate normal distribution.

3.0.4 Step 4: Incorporating Measurement Uncertainties

Measurement uncertainties affect the precision of our observed features. Ignoring them can lead to overconfidence and potentially misleading probability estimates. By incorporating uncertainties, we acknowledge the limitations of our data and adjust the model accordingly. The modeling of each data point will be as a 2D Gaussian distribution with the corresponding standard deviations in x, y given by the σ_x and σ_y uncertainties.

Modeling Measurement Errors We model the measurement errors as a bivariate normal distribution centered at the true values:

$$P_{\text{error}}(\mathbf{x}_{\text{obs}} | \mathbf{x}) = \mathcal{N}(\mathbf{x}_{\text{obs}}; \mathbf{x}, D)$$

where:

- $\mathbf{x}_{\text{obs}} = \begin{bmatrix} x_{\text{obs}} \\ y_{\text{obs}} \end{bmatrix}$ is the observed feature vector.
- $\mathbf{x} = \begin{bmatrix} x \\ y \end{bmatrix}$ is the true feature vector.
- $D = \begin{bmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_y^2 \end{bmatrix}$ is the measurement error covariance matrix.

Convolving Distributions The observed data likelihood given the class is obtained by integrating over the true features:

$$P(\mathbf{x}_{\text{obs}}|\text{Class} = c) = \int P_{\text{error}}(\mathbf{x}_{\text{obs}}|\mathbf{x})P(\mathbf{x}|\text{Class} = c)\, d\mathbf{x}$$

This convolution accounts for the uncertainty in the observed data by combining the measurement error distribution with the class-conditional distribution. It effectively integrates over all possible true values, weighted by how likely they are given the class and the measurement errors.

Result of Convolution The convolution of two Gaussian distributions is another Gaussian:

$$P(\mathbf{x}_{\text{obs}}|\text{Class} = c) = \mathcal{N}(\mathbf{x}_{\text{obs}}; \boldsymbol{\mu}_c, \Sigma_c' + D)$$

Thus, we adjust the covariance matrices:

$$\Sigma_c'' = \Sigma_c' + D$$

The adjusted covariance matrix Σ_c'' incorporates both the intrinsic variability of the class and the measurement uncertainties, providing a more realistic assessment of the data's spread.

3.0.5 Step 5: Calculating Likelihoods with Adjusted Covariances

The likelihood represents the probability of observing the data given the class. With the adjusted covariance matrices, we can compute likelihoods that account for both the natural variation within the class and the measurement uncertainties.

Likelihood Function For class c:

$$P(\mathbf{x}_{\text{obs}}|\text{Class} = c) = \frac{1}{2\pi|\Sigma_c''|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}_{\text{obs}} - \boldsymbol{\mu}_c)^\top(\Sigma_c'')^{-1}(\mathbf{x}_{\text{obs}} - \boldsymbol{\mu}_c)\right)$$

Components:

- $|\Sigma_c''|$: Determinant of the adjusted covariance matrix, indicating the volume of the distribution.
- $(\Sigma_c'')^{-1}$: Inverse of the adjusted covariance matrix, used to compute the Mahalanobis distance.
- The exponent represents the squared Mahalanobis distance, measuring how many standard deviations the observed data is from the class mean, considering the combined covariance.

3.0.6 Step 6: Applying Bayes' Theorem

Bayes' theorem allows us to update our prior beliefs with the observed data to compute the posterior probabilities. It combines the prior probability and the likelihood to give a revised probability of the hypothesis.

Bayes' Theorem

$$P(\text{Class} = c|\mathbf{x}_{\text{obs}}) = \frac{P(\mathbf{x}_{\text{obs}}|\text{Class} = c) \cdot P(\text{Class} = c)}{P(\mathbf{x}_{\text{obs}})}$$

where:

$$P(\mathbf{x}_{\text{obs}}) = \sum_{\text{all classes } c} P(\mathbf{x}_{\text{obs}}|\text{Class} = c) \cdot P(\text{Class} = c)$$

Interpretation:

- **Numerator**: The product of the likelihood and prior probability for class c, representing the unnormalized posterior.
- **Denominator**: The total probability of observing the data, serving as a normalization constant to ensure that the posterior probabilities sum to 1.

3.0.7 Step 7: Handling Edge Cases

Computational stability is important. If the total probability $P(\mathbf{x}_{\text{obs}})$ is zero due to extremely low likelihoods (e.g., the data point is an extreme outlier in all classes), we need to prevent division by zero and handle the uncertainty appropriately.

If $P(\mathbf{x}_{\text{obs}}) = 0$, we assign equal posterior probabilities:

$$P(\text{Class} = \text{True}|\mathbf{x}_{\text{obs}}) = P(\text{Class} = \text{False}|\mathbf{x}_{\text{obs}}) = 0.5$$

This reflects maximum uncertainty, acknowledging that the model provides no information to favor one class over the other in this scenario. It's a conservative approach to handle the lack of evidence.

4 Conclusion

Incorporating measurement uncertainties into the Bayesian classification framework provides a more accurate estimation of the probability that a star is a hypervelocity star. By adjusting the covariance matrices to include measurement errors, we account for the uncertainties inherent in astronomical observations. This leads to more reliable probability estimates and better-informed classifications.

Understanding and implementing this methodology is essential for astrophysical analyses where precision is crucial. The detailed steps and rationales provided ensure transparency in the calculations and facilitate further enhancements or adaptations to different datasets or classification tasks.

5 References

- Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.
  - Chapter 2: Probability Distributions
    - Section 2.3: The Gaussian Distribution

        * Section 2.3.4: Marginal and Conditional Gaussian Distributions
    - Chapter 4: Linear Models for Classification
        * Section 4.2: Probabilistic Generative Models
        * Section 4.2.2: Discriminant Functions for the Normal Density
- SciPy Documentation: Multivariate Normal Distribution
- NumPy Documentation: Covariance Matrix Calculation
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. Springer.
    - Chapter 2: Overview of Supervised Learning
    - Chapter 4: Linear Methods for Classification