Classification

Introduction

Least squares classification addresses the task of fitting a model to data where the outcome, or dependent variable y, takes on a limited set of values. This outcome is often referred to as a label or categorical variable and is typically a scalar number based on an n-vector x. In its simplest form, this problem involves only two possible values, such as TRUE (1) or FALSE (0), which is referred to as a Boolean classification problem.

Key Concepts

Least squares classification aims to find a linear decision boundary that separates different classes by minimising the sum of squared differences between predicted and actual labels. This boundary aims to reduce classification errors by fitting a hyperplane that best divides the data points.

Methodology

The linear model is represented as:

$$y = \beta_0 + \beta_1 x$$

where β_0 is the intercept, β_1 is the slope, and $\boldsymbol{\beta} = (\beta_0, \beta_1)$ represents the coefficients. The error for each data point is:

$$e_i = \hat{\beta}_0 + \hat{\beta}_1 x_i - y_i$$

The goal is to minimise the sum of squared errors (SSE):

$$\sum_{i=1}^{n} e_i^2 = \|\mathbf{A}\hat{\boldsymbol{\beta}} - \mathbf{y}\|^2$$

The solution to this is obtained by the Moore-Penrose pseudoinverse:

$$\hat{\boldsymbol{\beta}} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{y} = \mathbf{A}^\dagger\mathbf{y}$$

This provides the weight vector $\hat{\boldsymbol{\beta}}$, which is used to classify new data based on the sign of $\mathbf{A}\hat{\boldsymbol{\beta}}$. Data points are classified as:

$$\hat{y} = \begin{cases} 1 & \text{if } \mathbf{A}\hat{\boldsymbol{\beta}} > 0 \\ 0 & \text{otherwise} \end{cases}$$

Thus, based on the sign of $\hat{y}$, each data point is classified into one of the binary categories.

Applications

Least squares classification has a wide range of use cases, such as:

- **Spam Detection**: Classifying emails as spam or not based on features such as word count, use of exclamation points, or the presence of all-caps words.
- **Fraud Detection**: Analysing transaction patterns to detect potentially fraudulent activities by building models from historical data.
- **Disease Detection**: Predicting the likelihood of a diagnosable illness based on patient data, including medical history, test results, and symptoms.

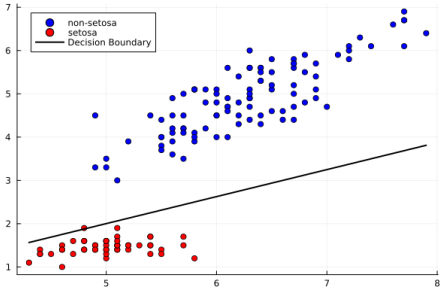

Figure 1: Example Classification