Basics of Linear Classification

Rishabh Iyer

March 19, 2024

1 Introduction

Linear classification is a foundational concept in machine learning, and one of the simplest class of models. A good understanding of linear classification will enable the readers to understand more concept classes of models which often build upon these simple linear classification models (e.g., polynomial classification, kernel machines, and neural networks).

Linear classification refers to the process of categorizing data points into two or more classes based on a linear combination of input features. Formally, given a set of input features $x \in \mathbb{R}^n$ and a target variable y that can take on values from a finite set of classes, the goal of linear classification is to find a decision boundary—a hyperplane—that separates the classes as well as possible.

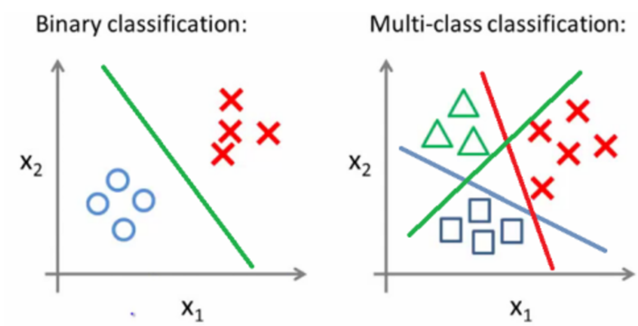

Figure 1: Illustratig Binary vs Multiclass Classification

Binary vs Multiclass Classification: In the case of Binary classification, we have two classes. If the number of classes is greater than 2, we have Multiclass classification. Figure 1 shows the difference between binary classification (two classes) and multiclass classification.

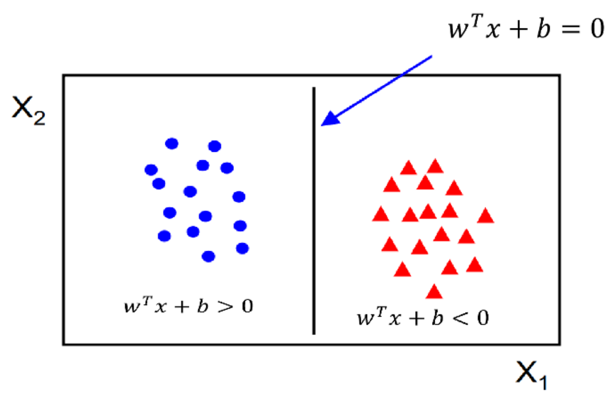

Figure 2: Illustratig a Linear Decision Boundary and how it separates the region into two parts: positives and negatives.

1.1 Applications of Linear Classification

Despite its simplicity, Linear Classification algorithms have been successfully applied in various domains, including:

- **Text Classification:** Categorizing documents into different topics based on their content.
- **Spam Detection:** Classifying emails or messages as spam or not spam based on their attributes.
- **Image Recognition and Classification:** Linear Classifiers can handle the high dimensionality of image data effectively, making them suitable for recognizing patterns or objects within images, particularly when used with good feature extractor models or embeddings.
- **Bioinformatics:** Linear Classifiers are used for protein classification, cancer classification, and gene expression analysis, where they help in identifying patterns in complex biological data.

These are some of the applications where the decision boundary between classes can be linear. Perceptrons, Linear Support Vector Machines (SVMs), Logistic Regression, etc. are some of the examples of Linear Classification Models that we will study in future blog articles in detail.

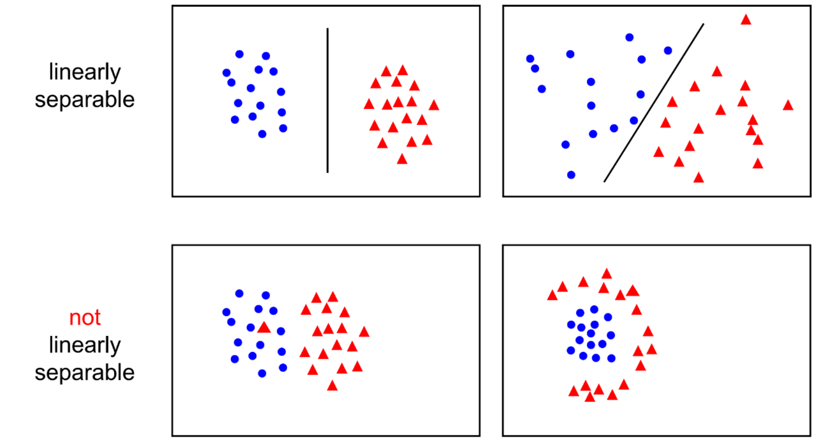

Figure 3: Illustrating Linear Separable dataset (top row) and non-linearly separable datasets (bottom row)

2 Linear Decision Boundary and Linear Separability

2.1 Linear Decision Boundary

A linear classifier makes predictions based on a linear combination of input features. The decision boundary of a linear classifier is defined by the equation:

$$w^T x + b = 0,$$

where w is the weight vector, x is the input feature vector, and b is the bias term. This equation represents a hyperplane in the feature space that separates the classes.

The space is divided into two regions by the hyperplane:

- $w^T x + b < 0$: The region where the predicted class is negative.
- $w^T x + b > 0$: The region where the predicted class is positive.

Figure 2 illustrates this concept. The blue dots are the positives and correspond to $w^T x + b > 0$ while the red triangles are the negatives and correspond to $w^T x + b < 0$. The separating hyperplane has the equation $w^T x + b = 0$.

2.2 Linear Separability

A dataset is said to be *linearly separable* if there exists at least one hyperplane that can perfectly separate the data points into their respective classes without any misclassification. In other words, all data points of one class lie on one side of the hyperplane, and all data points of the other class lie on the opposite side. Figure 3 shows the difference between linearly separable and not linearly separable. The top row is linearly separable, since there exists atleast one straight line (drawn) which separates the two classes. For the bottom row datasets, no such straight line exists. Notice that for the bottom row right plot, the dataset is *separable* using a non-linear classifier but is not linearly separable.

2.3 Characteristics of Loss Functions for Linear Classification

For classification tasks with linear models, the choice of loss function is crucial for guiding the learning algorithm towards an effective decision boundary. Suitable loss functions share several characteristics:

- **Margin Maximization:** The loss function should encourage a decision boundary that not only separates the classes but also maximizes the margin between the classes and the boundary. This principle underlies the hinge loss used in Support Vector Machines (SVMs).

- **Penalty for Misclassification:** The loss should penalize predictions that are on the wrong side of the decision boundary, with the penalty typically increasing with the distance from the boundary.

- **Differentiability:** While not strictly necessary (as seen with the hinge loss), differentiability of the loss function is desirable for the ease of optimization using gradient-based methods. When the loss function is not differentiable (e.g., the 0-1 loss), surrogate loss functions or subgradient methods can be employed.

- **Convexity:** Convex loss functions ensure that the optimization problem has a single global minimum, making the learning process more stable and predictable. This is a key property for loss functions like logistic loss and squared loss.

In the next few blog articles on Perceptrons and Support Vector Machines, and we will design loss functions which attempt to satisfy the above desiderata.

3 Conclusion

Understanding the linear decision boundary, the concept of linear separability, and the characteristics of effective loss functions are foundational for designing and implementing linear classifiers. These concepts not only guide the selection of models and optimization strategies but also influence the interpretability and performance of the resulting classifiers.