Imbalanced Classification

Starting From Fall 2025

•

By Jiajun Zhang

Contents

Chapter 1

An Overview of Classical Classification Techniques

1.1 Introduction to the classification problem

We begin our literature by considering regression models. Let us consider a set of observations $\mathcal{D}_n := \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \cdots, (\boldsymbol{x}_n, y_n)\}$ where $\boldsymbol{x}_i \in \mathbb{R}^p$ is the *covariate vector* (*we also denote as observations*) of dimension p and $y_i \in \mathbb{R}$ is the *response*. The objective of regression analysis is to find a function $m(\boldsymbol{x})$ such that it describes the relationship between the covariate vector and its response so we are able to make predictions on new observations. In classical regression models, the response y_i is assumed to be a continuous random variable with joint distribution $\mathbb{P}_{\boldsymbol{x},y}$ along with the covariate vectors and the objective is to estimate the conditional distribution $\mathbb{P}_{y|\boldsymbol{x}}$. If we seek for linear models then the regression function for observation $\boldsymbol{x}_i \in \mathcal{D}_n$ takes the form

$$y_i = \boldsymbol{x}^\top \boldsymbol{\beta} + \epsilon_i \tag{1.1.1}$$

where $\boldsymbol{\beta} \in \mathbb{R}^p$ is the *parameter*, ϵ_i is the *noise* and the *ordinary least squares estimator* $\widehat{\boldsymbol{\beta}}_n$ of $\boldsymbol{\beta}$ can be obtained. While in the sense of classification the response is assumed to be *categorical* which is represented by different classes $\mathcal{C}_1, \mathcal{C}_2, \cdots, \mathcal{C}_k$. For example, those classes can be genders; blood types; type of disease, and so on. In the most common scenario, all classes are taken to be disjoint so each $\boldsymbol{x}_i$ is assigned to one and only one class. The choice of the classes depends on the interest of researchers. For example, if we wish to study the relationship between the number of hours students spend for preparing exams and their grade (pass or fail), we may consider two classes to be

$$\mathcal{C}_1 := \{\text{The student's grade is above } 50\} \quad \text{and} \quad \mathcal{C}_2 := \{\text{The student's grade is below } 50\}.$$

Just like regression, the goal of classification is to set up a relation between $\boldsymbol{x}_i$ and $\mathcal{C}_i$ based on observed data $\mathcal{D}_n$, by dividing the space into separate *decision regions* $\mathcal{R}_1, \cdots, \mathcal{R}_k$. Each $\mathcal{R}_i$ contains all the observation of class $\mathcal{C}_i$ and we would be interested in finding the *decision boundary* between each two distinct classes. Once we have new observations, we will assign $\boldsymbol{x}$ to one of those k classes based on several quantities such as *similarity*, *distance* and so on.

1.2 Linear Discriminant Functions

A linear discriminant function takes an input observation $\boldsymbol{x}$ and assigns it to one of K classes. For simplicity let us consider the case when $K = 2$, the linear discriminant function takes the form

$$y(\boldsymbol{x}) = \boldsymbol{x}_i^\top \boldsymbol{\beta} + w_0 \tag{1.2.1}$$

where $\boldsymbol{\beta}$ is called a *weight vector* and w_0 is a *noise* term (or *bias*). The graph of $y(\boldsymbol{x})$ is simply a linear function in the xy-plane, and we define the *decision boundary* to be the line $y(\boldsymbol{x}) = 0$. Depends on our interest, say we have two classes $\mathcal{C}_1, \mathcal{C}_2$, and we will assign a new observation $\boldsymbol{x}$ to $\mathcal{C}_2$ if $y(\boldsymbol{x}) < 0$, and assign $\boldsymbol{x}$ to $\mathcal{C}_1$ if $y(\boldsymbol{x}) > 0$, as the figure shown below:

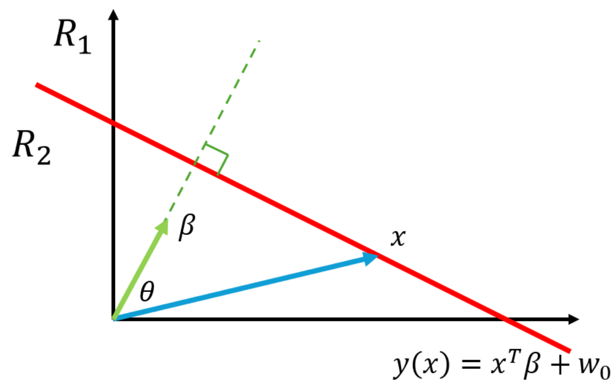

Figure 1.1: Illustration of the geometry of a linear discriminant function when $k = 2$

Let us consider two points $\boldsymbol{x}_A, \boldsymbol{x}_B$ such that they both lie on the decision surface, meaning $y(\boldsymbol{x}_A) = y(\boldsymbol{x}_B) = 0$, and we have $\boldsymbol{x}_A^\top \boldsymbol{\beta} = \boldsymbol{x}_B^\top \boldsymbol{\beta}$, which further implies $(\boldsymbol{x}_A - \boldsymbol{x}_B)^\top \boldsymbol{\beta} = \boldsymbol{0}$, i.e the parameter vector $\boldsymbol{\beta}$ is orthogonal to the discriminant function $y(\boldsymbol{x})$ (as shown in figure 1.1). Next, we consider the distance d from origin to the discriminant function $y(\boldsymbol{x})$. Denote θ $(0 \leq \theta \leq \pi)$ by the angle between $\boldsymbol{x}$ and $\boldsymbol{\beta}$ where $\boldsymbol{x}$ is an arbitrary vector on $y(\boldsymbol{x})$ shown in figure 1.1, then the projection theorem yields that

$$\boldsymbol{x}^\top \boldsymbol{\beta} = ||\boldsymbol{x}|| \cdot ||\boldsymbol{\beta}|| \cos\theta \tag{1.2.2}$$

where by definition we see that $||\boldsymbol{x}|| \cos\theta = d$, so combine with the fact that $y(\boldsymbol{x}) = 0$ we have

$$\frac{\boldsymbol{x}^\top \boldsymbol{\beta}}{||\boldsymbol{\beta}||} = -\frac{w_0}{||\boldsymbol{\beta}||} := d \tag{1.2.3}$$

Bibliography

[1] Bishop M. Christopher. *Deep Learning, Fundations and Concepts*.