Classification: Introduction

Patrick Breheny

September 13

Introduction

- We now turn our attention to *discriminant analysis*, a method for analyzing data in which the outcome variable is categorical
- In this section, we suppose that the outcome Y takes on values in a discrete set $\mathcal{G}$; for example:
  - Type of leukemia: $\mathcal{G} = \{\text{ALL}, \text{AML}, \text{CLL}, \text{CML}\}$
  - Type of brain cell: $\mathcal{G} = \{\text{Immature neuron}, \text{Mature neuron}, \text{Glial cell}\}$
- The task of developing a predictor $G(\mathbf{x})$, which predicts a value of $\mathcal{G}$ depending on the explanatory variables $\mathbf{x}$, is often called *classification*

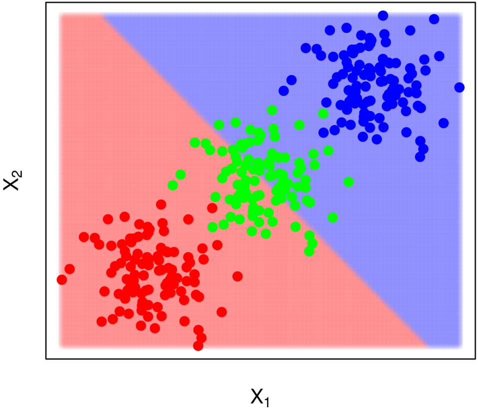
Notation and terminology

- Consider indexing the elements of $\mathcal{G}$ with $1, \ldots, K$, where K is the number of classes, and let G_i denote the index of Y_i
- We will be discussing methods which model a *discriminant function* $\delta_k(\mathbf{x})$ for each class, and then classify $\mathbf{x}$ according to the class with the largest discriminant function
- The *decision boundary* between classes k and l is the set of points for which $\delta_k(\mathbf{x}) = \delta_l(\mathbf{x})$
- Ideally, such a method will also allow us to calculate $\Pr(G = k|\mathbf{x})$, as this is the most useful and interpretable discriminant function

Linear regression of an indicator matrix

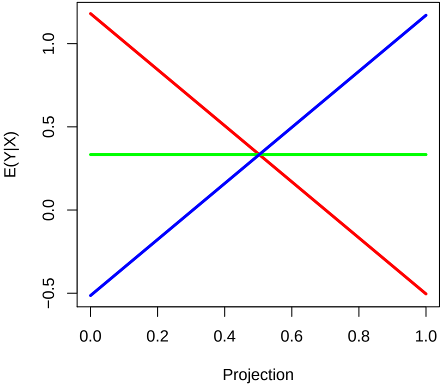
- We begin by considering a very simplistic model: linear regression applied to indicator variables
- Let us rewrite the outcome as a K-dimensional vector of indicator variables, each indicating whether or not $G_i = k$
- We could then fit K linear regression models, one for each indicator
- Our discriminant function would then be $\delta_k(\mathbf{x}) = \mathbf{x}^T\widehat{\boldsymbol{\beta}}_k$ – *i.e.*, we classify according to the conditional expectation of Y given $\mathbf{x}$

How well does this work?



What went wrong?

The horizontal axis measures the position along the line joining the means of the three classes:



Masking

- The green class is *masked* by the other two – obviously, an undesirable effect
- This is an extreme example, but masking (complete or partial) can easily occur when K is large
- For this reason, linear regression is not a very good classifier –other methods, such as discriminant analysis and logistic/multinomial regression, do not suffer from the masking problem and perform much better