4.1 An Overview of Classification

Zongyi Liu

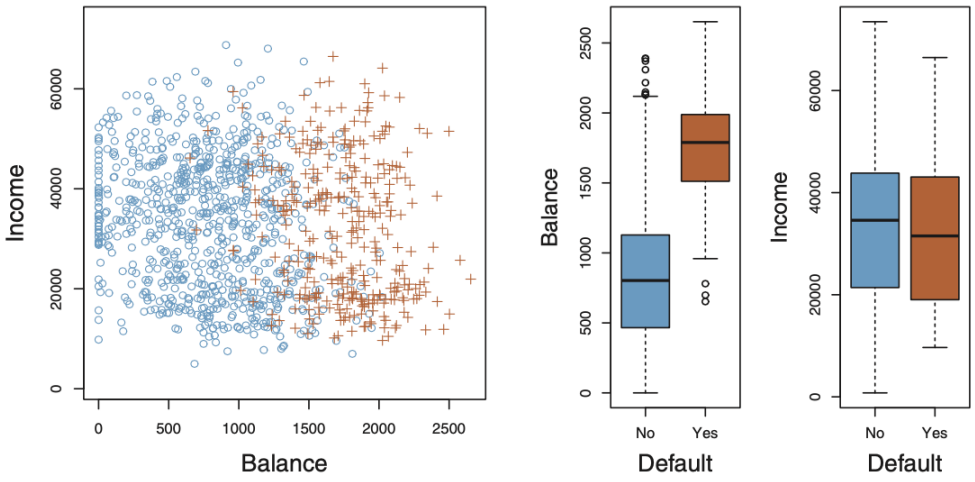
2023-05-18

4.1 An Overview of Classification

In previous chapters, we consider variables to be quantitative, but in many situations, they are qualitative. In this chapter we will discuss the mostly used three classifiers: logistic regression, linear discriminant analysis, and K-nearest neighbors.

Classification problems occur often, perhaps even more so than regression problems. Some examples include:

1. A person arrives at the emergency room with a set of symptoms that could possibly be attributed to one of three medical conditions. Which of the three conditions does the individual have?
2. An online banking service must be able to determine whether or not a transaction being performed on the site is fraudulent, on the basis of the user's IP address, past transaction history, and so forth.
3. On the basis of DNA sequence data for a number of patients with and without a given disease, a biologist would like to figure out which DNA mutations are deleterious (disease-causing) and which are not.



- Left: The annual incomes and monthly credit card balances of a number of individuals. The individuals who defaulted on their credit card payments are shown in orange, and those who did not are shown in blue.
- Center: Boxplots of `balance` as a function of `default` status.
- Right: Boxplots of `income` as a function of `default` status.