Notes:
The Elements of Statistical Learning

Daniel Saunders

August 23, 2018

Notes

1. Some *emphasis* is from the book, some is added.
2. Abbreviationsare used liberally and must sometimes be inferred from context.

1 Introduction

Statistical learning plays a key role in many areas of science, namely statistics, data mining, and artificial intelligence, and intersects with engineering and other disciplines.

The book is about learning from data. Typically, we have a quantitative or categorical outcome measurement that we want to predict based on a set of *features*. We have a *training* set of data, in which observe both outcome and feature for a set of objects. Using this data, we build a prediction model (*learner*) which enables us to predict outcomes for unseen objects.

The above describes the *supervised learning* problem, called so because of the presence of the outcome measurement to guide the learning process. In the *unsupervised learning* problem, no outcome measurements are available, so we must instead describe how the data are organized or clustered.

2 Overview of Supervised Learning

2.1 Introduction

For each of the examples in Chapter 1, there is a set of variables known as the *inputs* (measured or preset), which have influence over one or more *outputs*. For each example, the goal is to use the inputs to predict the outputs. This is known as *supervised learning*.

In the statistics / pattern recognition literature, the inputs are often called the *predictors*, *independent variables*, or *features*, whereas the outputs are called the *responses* or *dependent variables*.

2.2 Variables Types and Terminology

Outputs variables may vary in nature; some *quantitative* measurements are larger than others, and close measurements are close in nature. On the other hand, *qualitative* measurements assume values in a finite set, without explicit ordering, and sometimes are descriptive labels rather than numbers to denote the classes. Qualitative variables are sometimes referred to as *categorical* variables, *discrete* variables, or *factors*.

The distinction in output type has led to a naming convention for prediction tasks: *regression* when we predict quantitative outputs, and *classification* when we predict qualitative outputs. Both can be viewed as tasks in function approximation.

Inputs can also vary in measurement type, with some qualitative and some quantitative variables. Some methods are better suited to one type or the other, or both.

A third variable type is *ordered categorical* (e.g., small, medium, or large), where there is an ordering, but no metric notion is appropriate.

Qualitative variables are typically represented numerically by codes (sometimes referred to as *targets*). Binary variables can be represented simply by 0 and 1, or -1 and 1. With more than two categories, a commonly used coding is via *dummy variables* (*one-hot encoding*), where a K-level qualitative variables is represented by a vector of K bits, only one of which is "on" at a time.

2.3 Two Simple Approaches to Prediction: Least Squares and Nearest Neighbors