The Elements of Statistical Learning

Hastie, Tibshirani, and Friedman (2009). The Elements of Statistical Learning. Second Edition. Springer.

1 Introduction

Statistical learning plays a key role in science, finance, industry, and many more areas. This book is about learning from data: supervised learning (presence of outcome variable for learning, the focus of this book) and unsupervised learning (outcome variable is unobserved).

Running examples:

- Classification of spam emails
- Explaining prostate specific antigen from clinical measurements via regression
- Classification of handwritten digits
- Clustering of DNA microarray data for cancer diagnostic

2 Overview of Supervised Learning

Two simple but powerful prediction methods are least squares linear models and k-nearest neighbors. The former makes huge assumptions about structure and yields stable but possibly inaccurate predictions (low variance, high bias), the latter makes very mild assumptions with often accurate but unstable predictions (at least if k is low, leading to low bias and high variance).

Local methods like k-nearest neighbors suffer from the curse of dimensionality: in high dimensions, samples only sparsely populate the input space and are close to an edge (extrapolation might be required). By imposing restrictions on the model class (e.g., linear models), this can be avoided. Many models have been proposed that lie in the spectrum between rigid model assumptions and flexibility, they will be presented in the book.

3 Linear Methods for Regression

Linear regression models are simple and often adequate and interpretable. The Gauss-Markov theorem states that the least squares estimates have the smallest variance among all linear unbiased estimates. However, it might be a good idea to trade a little bit bias for a large reduction in variance. This is possible with different variable subset selection and shrinkage methods:

- Best-subset selection:
- Forward- and backward-stepwise selection:
- Forward-stagewise regression:
- Ridge regression:
- The lasso:
- Least angle regression:
- Principal components regression:
- Partial least squares: