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: For each given size, find the best subset of variables that minimizes the residual sum of squares. Strategies for choosing the size will be discussed later. Not applicable for a large number of variables.
- Forward- and backward-stepwise selection: Searching through all possible subsets quickly becomes infeasible. Instead, sequentially add or remove variables.
- Forward-stagewise regression: A more constrained version of forward-stepwise, with benefits in high dimensions.

- Ridge regression: The idea is to make the selection process continuous by shrinking the coefficient values, which can further reduce variance. The penalty for coefficient sizes is quadratic (L_2). Variables with small variance are shrunk the most.
- The lasso (least absolute shrinkage and selection operator): Similar to ridge regression, but the penalty is in absolute coefficient value (L_1). This can make some coefficients to be exactly zero.
- Elastic-net penalty: A convex combination of the ridge and the lasso penalty.
- Least angle regression: Similar to forward-stepwise regression with the difference, that entering variables are not fit completely but only until it no longer has the highest correlation with the current residual. A slight modification provides an efficient way of computing the entire lasso path.
- Principal components regression: Use transformed variables with large sample variance.
- Partial least squares: Also constructs a set of linear combinations of the inputs for regression, but unlike principal components regression it uses the dependent variable for construction.

4 Linear Methods for Classification

Linear here means linear decision boundaries. One option is linear regression for indicator response matrix. Classes can be masked by others here , which can be avoided by adding polynomial terms. Another option is linear Discriminant Analysis. If error covariance is not constant, this yields quadratic discriminant functions. The difference in covariances can be regularized.