The Elements of Statistical Learning - Overview of Supervised Learning

Diego ROJAS

September 17, 2022

1 Introduction

Supervised learning aims at predicting the value of outputs based on the inputs.

Input is synonym to *predictors*, *independent variables* and *features*.

Output is synonym to *responses* and *dependent variables*.

2 Variable Types and Terminology

Variables can be *qualitative* (also called *factors*, *discrete* or *categorical*) or *quantitative*.

Regression is when we predict quantitative outputs. *Classification* is when we predict qualitative outputs.

Variables can also be *ordered categorical* such as *small*, *medium*, *large*.

Qualitative variables are often encoded into numeric values which are referred to as *targets*.

Inputs are denoted by X, quantitative outputs by Y, qualitative outputs by G.

The ith observed value from the vector X is denoted as x_i (where x_i is a scalar or a vector).

The learning task is hence defined as follows: given the value of X, make a good prediction of the output Y, denoted $\hat{Y}$.

3 Two Simple Approaches to Prediction

3.1 Linear Models and Least Squares

The linear model is defined as follows:

$$\hat{Y} = \hat{\beta}_0 + \sum_{j=1}^{p} X_j \hat{\beta}_j$$

Where $\hat{\beta}_0$ is the *intercept* or *bias*. If we include $\hat{\beta}_0$ in $\hat{\beta}$ and include 1 in X the model can be expressed as an inner product:

$$\hat{Y} = X^\top \hat{\beta}$$

One way of fitting this model over a set of training data $\boldsymbol{X}^{N \times p}$ of size N consists in minimising the RSS (*residual sum of squares*) loss function defined as:

$$RSS(\beta) = (\boldsymbol{y} - \boldsymbol{X}\beta)^\top (\boldsymbol{y} - \boldsymbol{X}\beta)$$

Where $\boldsymbol{y}$ is a vector of the outputs in the training set. We can differentiate w.r.t β and obtain:

$$\boldsymbol{X}^\top (\boldsymbol{y} - \boldsymbol{X}\beta) = 0$$

Then, if $\boldsymbol{X}^\top \boldsymbol{X}$ is nonsingular, the unique solution is given by:

$$\hat{\beta} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$$

Therefore $\hat{y}_i$ is given by $x_i^\top \hat{\beta}$.

3.2 Nearest Neighbour Methods

The nearest neighbour methods, look at the observations in the training set $\mathcal{T}$ which are closest to x to form $\hat{Y}$. Specifically, the k nearest neighbours.

$$\hat{Y}(x) = \frac{1}{k} \sum_{x_i \in N_k(x)} y_i$$

3.3 From Least Squares to Nearest Neighbours

4 Statistical Decision Theory

5 Local Methods in High Dimension

6 Statistical Models, Supervised Learning and Function Approximation

6.1 A Statistical Model for the Joint Distribution $Pr(X, Y)$

6.2 Supervised Learning

6.3 Function Approximation

7 Structured Regression Models

7.1 Difficulty of the Problem

8 Classes for Restricted Estimators

8.1 Roughness Penalty and Bayesian Methods

8.2 Kernel Methods and Local Regression

8.3 Basis Function and Dictionary Methods

9 Model Selection and Bias-Variance Trade-off