Overview

Maryclare Griffin

2024-09-05

This material is based on Chapter 1 of Introduction to Statistical Learning (ISL) and Chapter 1 of Elements of Statistical Learning (ESL). We will tend to follow ISL more closely, and look to ESL for occasional additional higher level material.

The phrase **statistical learning** is a broad term that is not very well defined. The authors of ISL define it as "a set of tools for understanding data." It can be divided into:

- **Supervised learning**: We observe inputs and (some) outputs, and wish to relate them via a statistical model, e.g. trying to predict a continuous output (**linear regression**) or trying to predict a categorical or qualitative output (**classification**).
- **Unsupervised learning**: We observe inputs and no outputs, and wish to describe the relationships within/structure of the inputs, e.g. clustering.

Throughout, we will use the following notation:

- x_{ij} denotes the value of the j-th variable for observation i, i.e. the i-th observation of input j
- y_i denotes i-th observation of the quantity we want to predict, i.e. the i-th output
- n denotes the number of observations
- p denotes the number of variables
- $\boldsymbol{X} = \begin{pmatrix} x_{11} & \dots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \dots & x_{np} \end{pmatrix}$ denotes the matrix of inputs
- $x_i = \begin{pmatrix} x_{i1} \\ \vdots \\ x_{ip} \end{pmatrix}$ denotes the p dimensional vector of input variables for observation i
- $\boldsymbol{x}_j = \begin{pmatrix} x_{1j} \\ \vdots \\ x_{nj} \end{pmatrix}$ denotes the n dimensional vector of all observations of input variable j
- $\boldsymbol{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$ denotes the n dimensional vector observations of the quantity we want to predict or output variable
- When outputs are observed, we denote our observed data as $\{(x1, y_1), \dots, (x_n, y_n)\}$.
- We may sometimes drop the subscripts and capitalize to refer to an arbitrary, potentially not yet observed, input or output e.g. $X = (X_1, \dots, X_p)$ or Y

Note that the terminology can vary. For instance, what we are referring to as inputs are sometimes also called **predictors**, **independent variables**, or **features**. What we are referring to as outputs are sometimes called **responses** or **dependent variables**.