Lecture 1 - Machine Learning Introduction

Sondre Elstad

2022-06-25

Introduction

We are surrounded by data and questions that requires data to be answered. Economists are discussing economic growth, we might try to estimate the effects of education on social class or political preferences or one might happen to run some large-scale experiment in the natural sciences. We need to understand data, and statistical learning aids in this.

We distinguish between supervised and unsupervised learning problems. Albeit we will have a look at both, our main focus will be on supervised learning. These techniques aims to facilitate the process of inference and prediction in the traditional sense of the word. They entail investigating the relationship between a set of regressors (or features in machine learning terminology) and an outcome variable (or response). In unsupervised learning, we do not have a response variable. Rather, the task is to uncover inherent clusters or structures within the set of features.

In the coming weeks we will talk more about what a statistical model is. We will talk about regression and classification. We will talk about model selection and validation - and, in particular, the usefulness of cross-validation. We will touch on unsupervised learning and deep learning. We will also force ourselves to lift our gazes for a moment and consider the ethical ramification of what we are doing with said models. Finally, we will consider, as a particular case of machine learning, techniques for working with textual data. These lectures are more about learning key ideas in machine learning. Math is necessary to make certain points (hopefully) as clear as possible, but the main concern are the ideas that the mathematical formalism represents.

Syllabus

In the course of these lectures, we will look closer at some data sets. These data sets are analyzed in *An Introduction to Statistical Learning with Applications in R* (published in 2017, and henceforth referred to as ISL), written by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani. There are a broad range of sources available in the reading lists. These offer different perspectives on the topics covered and each have their own advantages and disadvantages. In the first 6 lectures, I will draw extensively from the presentation in *An Introduction to Statistical Learning with Applications in R*. Now, for the datasets:

1. Wage data: A number of factors affecting wages for a group of men from the Atlantic region of the United States. The relationship between age and education, on the one hand, and wages is particularly dominant. Figure 1.1 (the pictures are placed at the last pages of these notes) illustrates these data. Wage is measured in USD and, as such, it has continuous output values. We refer to it as a quantitative output. This problem is a classical example of a regression problem.

2. Sometimes, we might be more interested in considering a qualitative output variable. An output variable with the values 'Up' and 'Down' is such an output variable, and the problem of predicting whether stock prices are going up or down (i.e., problems involving such qualitative output variables) are called classification problems. The Smarket data, illustrated in Figure 1.2, provides us with a

qualitative output variable and our task is to use these data (daily movements in the S&P 500 stock index from 2001 to 2005) to predict whether the stock prices are going up or down.

3. Cases 1 and 2 sort under the label 'supervised learning'. Unsupervised learning is an increasingly important field of machine learning. We will discuss unsupervised learning several times during our lectures. However, coming from a traditional empirical science, these kind of problems might seem rather strange or alien at first. We will see that they are useful in many settings. This field of machine learning provides us with techniques to cluster data, and one example we will look closer at is the NC160 data set. This consists of 6,830 gene expression measurements for 64 cancer cell lines. Figure 1.3 presents different clusters for the data set.

We can find all of these data sets in the ISRL2 package.

Mathematics

Even though the mathematical exposition is kept as rudimentary as possible, we nevertheless have to express certain key ideas in a formal manner. Mathematics allows us to do this. "A picture says a thousand words!" This is indeed true for mathematics. In fact, we could explain all of the models in the coming lectures using verbal prose. However, such a presentation would undoubtedly become quite tedious. Certain ideas are much clearer phrased using a formal equation or a matrix or some mathematical function. This is one of the main reasons for using mathematics. However, some people are quite used to mathematics. Others do everything they can to steer clear of it. What we will be doing over the next few weeks is to discuss ideas. I hope you will see that the mathematical definition of these ideas will make it easier for you to understand (and remember!) them. However, if you disagree, my biggest wish is that you understand the models and ideas to be discussed at an intuitive level.

Some key definitions: n is used to represent the number of observations in a sample, whereas p denotes the number of predictors. x_{ij} represents the value of the jth variable for the ith observation ($i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, p$). $\mathbf{X}$ denotes a $n \times p$ matrix with the (i, j)th element being x_{ij}:

$$\boldsymbol{X} = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix}.$$

You might think of a matrix as a spreadsheet of numbers (I dare to say we've all used Excel). A vector is like a list of numbers and, by default, a vector is defined as a column of numbers (as opposed as a row of numbers). A row in $\mathbf{X}$ (which we write as $x_1, x_2, \ldots, x_n$). x_i is of length p.

$$\boldsymbol{x_i} = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{pmatrix}$$

The element $x_{i,j}$ belongs to row i and column j in $\mathbf{X}$. This implies that you should not be tricked by the fact that the numbers in x_i above are sorted in a column. As a matter of fact, they all belong to the same row, namely row i in $\mathbf{X}$. We write the columns of $\mathbf{X}$ as $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_p$ (all of them are of length n). In fact, we will write all vectors of length n in bold types. As such,

$$\boldsymbol{x_j} = \begin{pmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{nj} \end{pmatrix}$$

Thus, we can write $\mathbf{X}$ as $\mathbf{X} = (\mathbf{x}_1 \quad \mathbf{x}_2 \quad \cdots \quad \mathbf{x}_p)$, or

$$\boldsymbol{X} = \begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{pmatrix}$$

where the T notation denotes the transpose of a matrix or vector. Transpose means pivot - transposing a column turns it into a row, and vice versa. So

$$\boldsymbol{X}^T = \begin{pmatrix} x_{11} & x_{21} & \cdots & x_{n1} \\ x_{12} & x_{22} & \cdots & x_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1p} & x_{2p} & \cdots & x_{np} \end{pmatrix}$$

and $x_i^T = (x_{i1} \quad x_{i2} \quad \cdots \quad x_{ip})$. We typically use x to represent features and y to represent response or output variable. In other words, y_i is the ith observation of the variable for which we make predictions. We can list all the responses accordingly

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

Observed data thus consists of $\{(x_1, y_1), (_2, y_2), ..., (x_n, y_n)\}$. Recall, each x_i is of length p. Sometimes, we need to be explicit about the dimension of a particular math (a crucial part of being formal is being tidy. . .). If an object is a scalar, we write $a \in \mathbb{R}$. For a vector of length p, we use $a \in \mathbb{R}^k$ (or $\mathbf{a} \in \mathbb{R}^n$ if it is of length n). A matrix of order $r \times s$ (i.e., with r rows and s columns) is denoted by $\mathbf{A} \in \mathbb{R}^{r \times s}$.

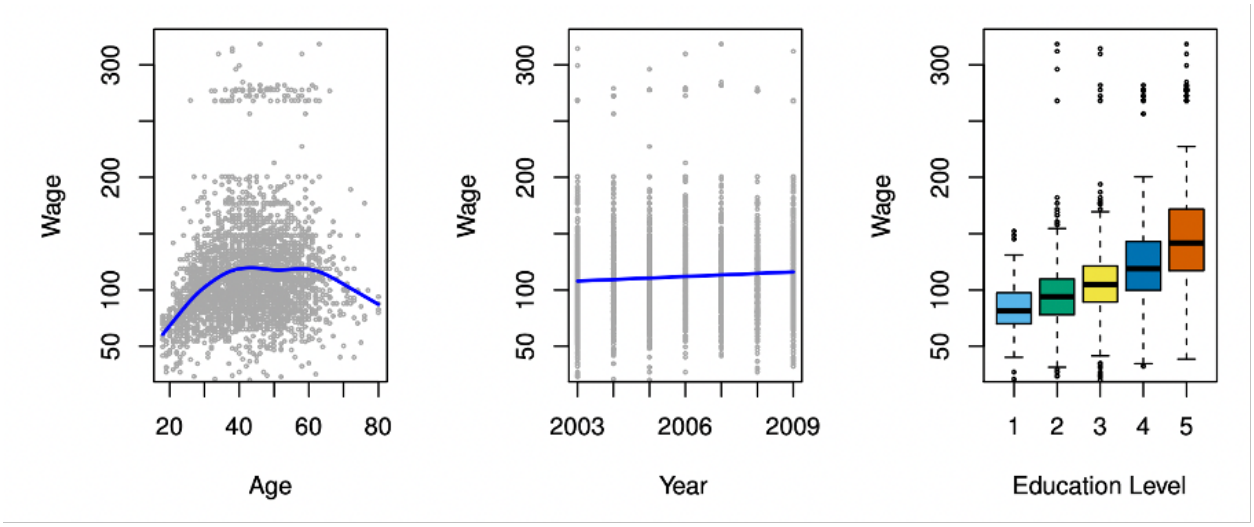

Figure 1: Wage data for men from the Atlantic region of the United States. From left to right: 1) wage as a function of age, 2) wage as a function of year and 3) wage as a function of education, on a scale from 1-5 (1: no high school diploma and 5: an advanced graduate degree). Source: ISL (2017)

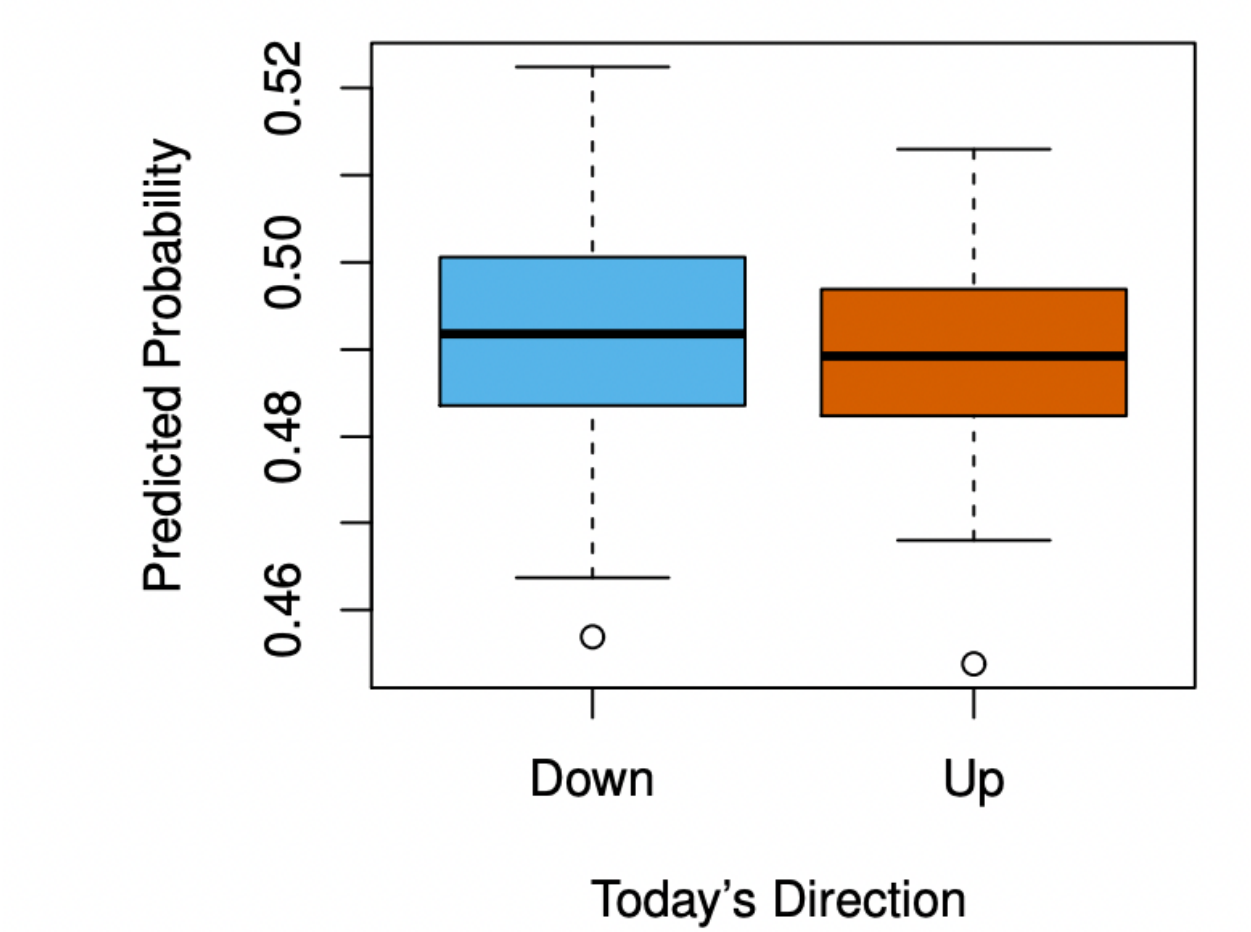

Figure 2: Boxplot of predicted probability of stock market decrease using the 2005 data. The probability of a decrease is predicted to be higher for days in which the market actually does decrease. The predictions are correct about 60% of the time. Source: ISL (2017)

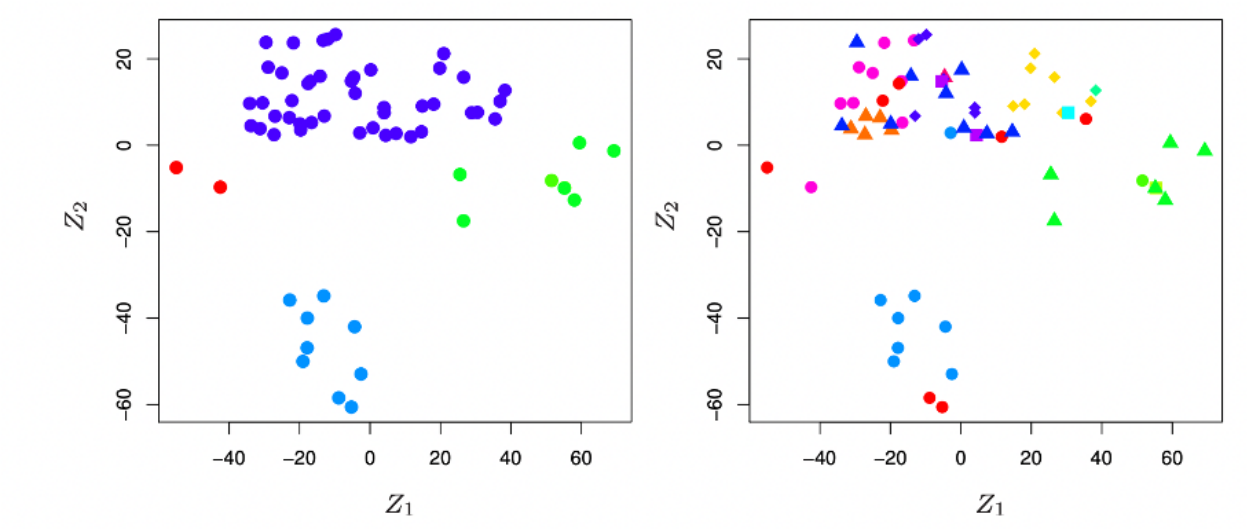

Figure 3: Two-dimensional represenations of the NC160 gene expression data set. Each point corresponds to one of 64 cell lines. Left: 4 clusters, represented in different colours. Right: 14 different clusters. We see much more overlap in the right panel. Cell lines corresponding to the same kind of cancer are close to one another. Source: ISL (2017).