BUAD 5072

Machine Learning I

Introduction to Machine Learning

ISLR Chapter 1

Agenda

- A Quick Overview
- Typical Datasets and Models
- A Brief History (very quickly)
- The Three Premises of the Course
- Access to data for labs and exercises

A Brief Overview

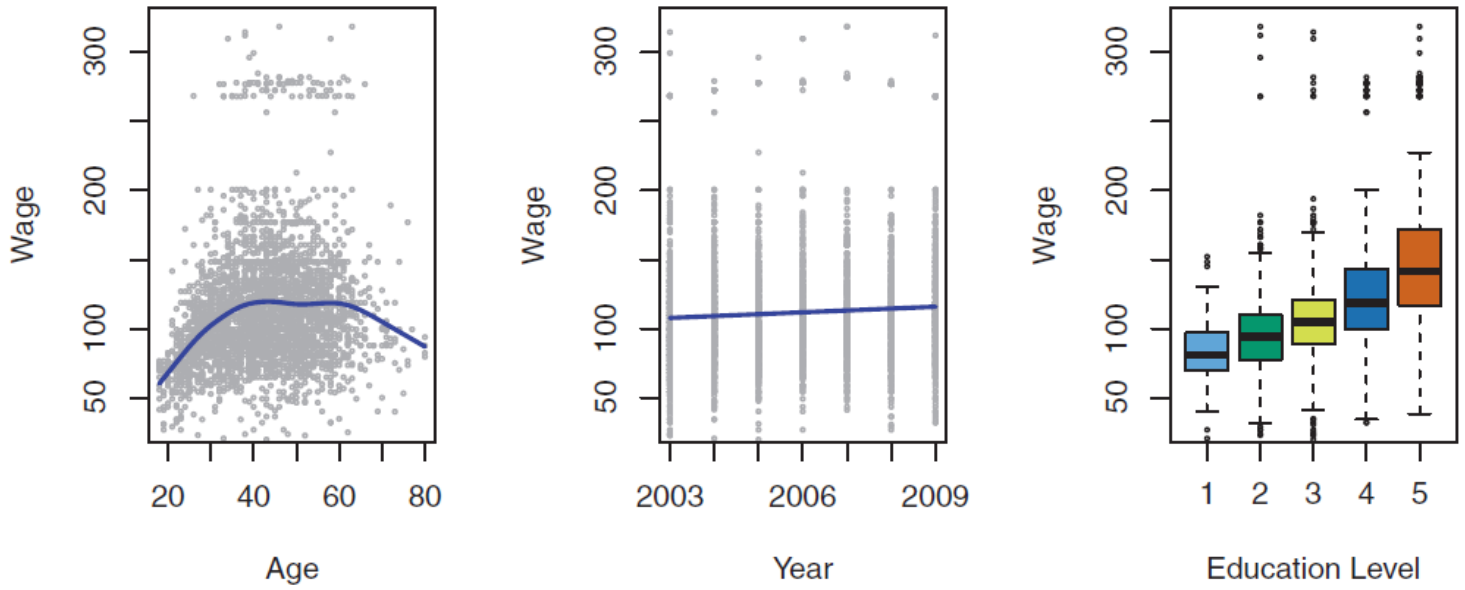
- Machine Learning tools can be classified as *supervised* or *unsupervised*
  - Supervised machine learning involves building a statistical model for predicting, or estimating, an *output* based on one or more *inputs*.
    - Problems of this nature are ubiquitous in business – examples include:
      - Predicting which product features are most desirable for an individual…so-called “recommender systems”,
      - estimating the future economic value of a particular customer,
      - predicting the time and nature of the next machine failure,
      - forecasting the direction of the stock market, and
      - identifying anomalies in financial transactions.

A Brief Overview

- Machine Learning tools can be classified as *supervised* or *unsupervised*
  - Unsupervised machine learning requires inputs but does not require that we specify a supervising output; that is, we do not ask the model to make a prediction about something.
    - Even though nothing is predicted, these models can be very valuable in revealing relationships and structure within the input data.
    - A typical application may be the clustering of existing customers to identify groups of customers that are similar within groups but different between groups. Such analysis can lead to provocative questions about why apparently dissimilar customers were clustered together, or why apparently similar customers were not.

Three Real-World Datasets

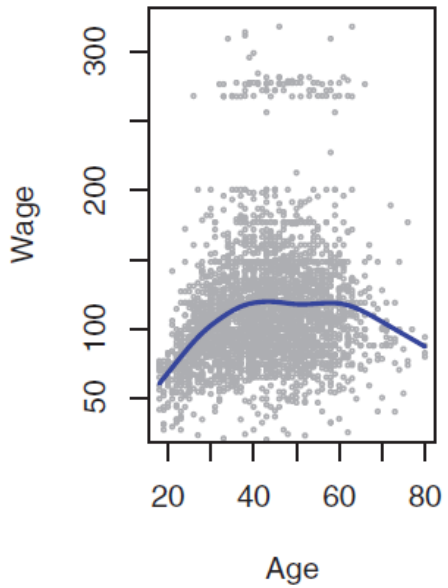
- Wage Data: A Regression Problem



Wage data, which contains income survey information for males from the central Atlantic region of the United States. Left: `wage` *as a function of* `age`*. On average,* `wage` *increases with* `age` *until about* 60 *years of age, at which point it begins to decline.* Center: `wage` *as a function of* `year`*. There is a slow but steady increase of approximately* $10,000 *in the average* `wage` *between* 2003 *and* 2009. Right: *Boxplots displaying* `wage` *as a function of* `education`*, with* 1 *indicating the lowest level (no high school diploma) and* 5 *the highest level (an advanced graduate degree). On average,* `wage` *increases with the level of education.*

Three Real-World Datasets

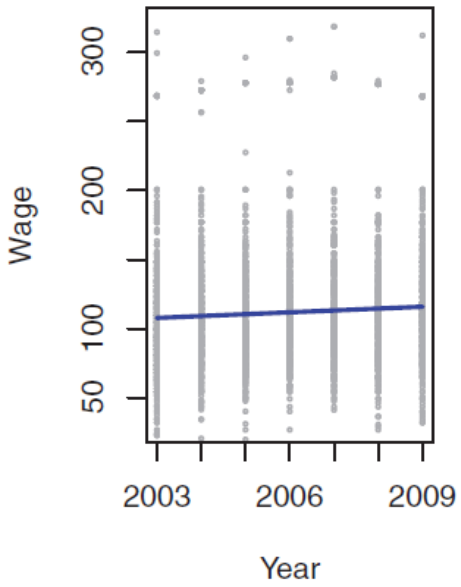
- Consider wages as a function of age



We could use the blue “mean” line to predict wage for a given age, but observe the high level of variability around the mean. It suggests that age alone is unlikely to provide an accurate prediction of a particular person’s wage.

Three Real-World Datasets

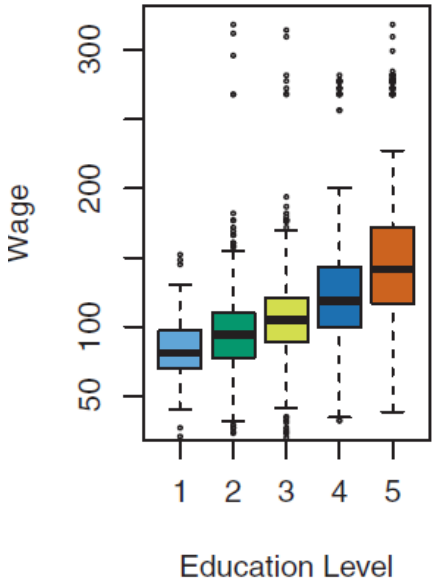
- Consider wages as a function of the year



We note that wages increase by approximately $10,000, in a roughly linear (or straight-line) fashion, between 2003 and 2009, though this rise is very slight relative to the variability in the data.

Three Real-World Datasets

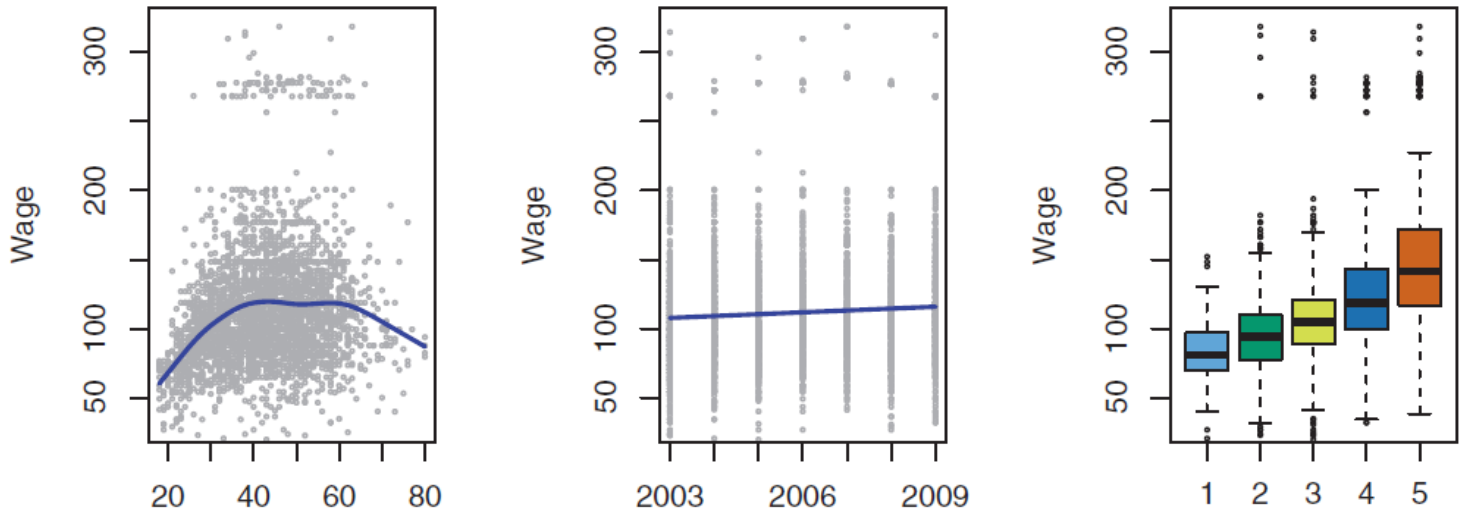
- Consider wages as a function of the education



It seems that wages are also typically greater for individuals with higher education levels: men with the lowest education level (1) tend to have substantially lower wages than those with the highest education level (5), but once again observe the high level of variability for each category.

Three Real-World Datasets

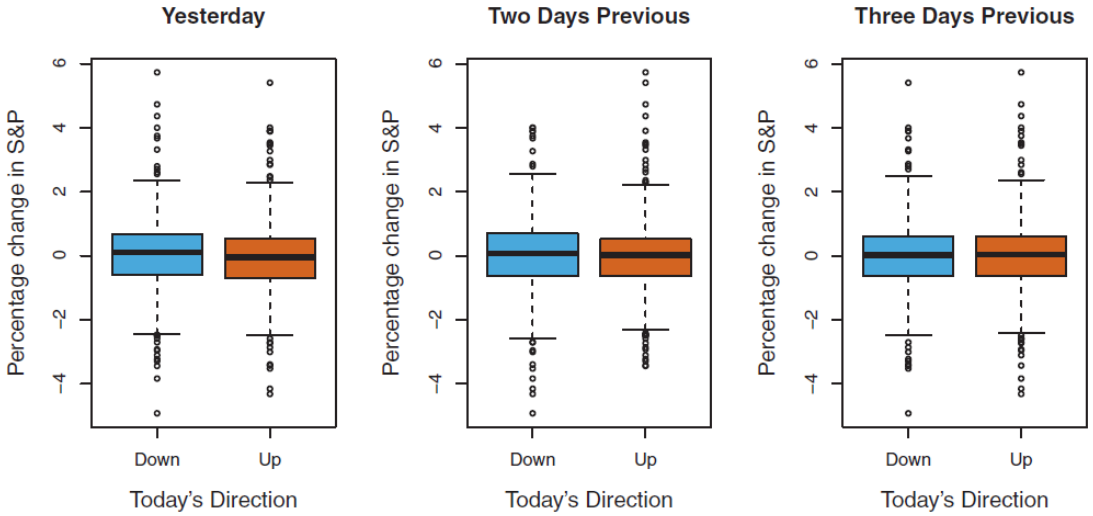
- Wage Data: A Regression Problem



- Clearly, the most accurate prediction of wages will be obtained by combining age, education, and year. When we discuss linear regression in this class, we will predict wage from this data set.
- Ideally, we should predict wage in a way that accounts for the non-linear relationship between wage and age. In Machine Learning 2, we will discuss a variety of approaches for addressing these non-linear situations.

Three Real-World Datasets

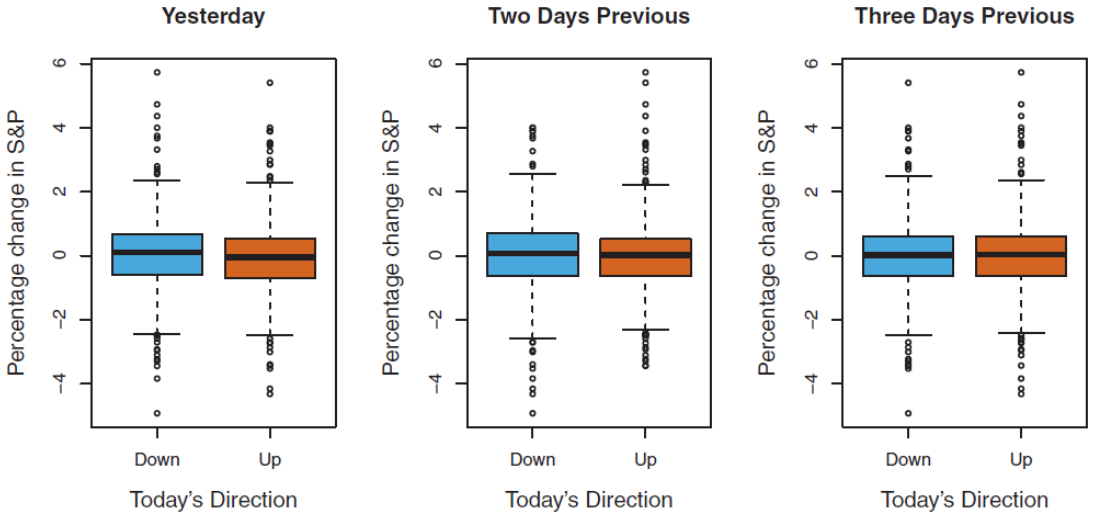
- Stock Market Data:



Left: *Boxplots of the previous day's percentage change in the S&P index for the days for which the market increased or decreased, obtained from the* `Smarket` *data.* Center and Right: *Same as left panel, but the percentage changes for 2 and 3 days previous are shown.*

Three Real-World Datasets

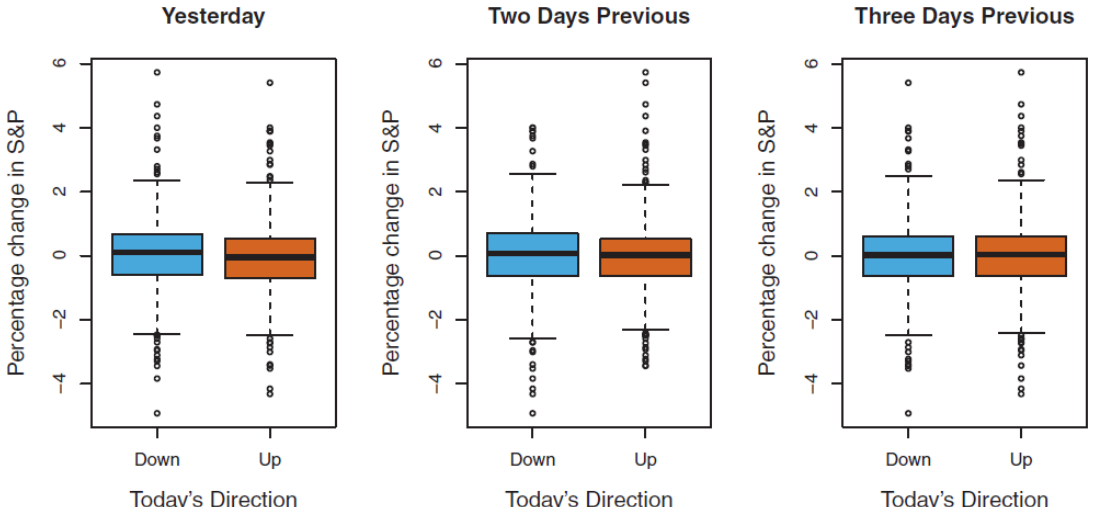
- Stock Market Data:



The previous Wage data involves predicting a *continuous* or *quantitative* output value. This is often referred to as a *regression* problem. However, in certain cases we may instead wish to predict a non-numerical value—that is, a *categorical* or *qualitative* output.

Three Real-World Datasets

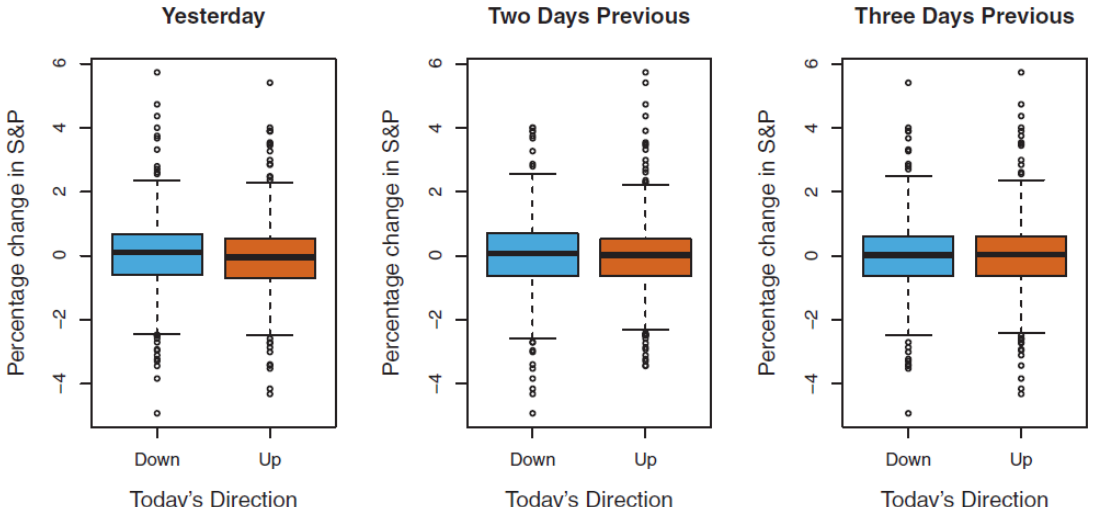
- Stock Market Data: A Classification Problem



Later, we will examine a stock market data set that contains the daily movements in the S&P stock index over a 5-year period between 2001 and 2005. The goal will be to predict whether the index will *increase* or *decrease* on a given day using the past 5 days' percentage changes. Here the problem does not involve predicting a numerical value, but rather whether a given day's stock market performance will be Up or Down. This is known as a ***classification*** problem.

Three Real-World Datasets

- Stock Market Data: A Classification Problem



We will fit a quadratic discriminant analysis (QDA) model to the subset of the data corresponding to the 2001–2004 time period, and predict the probability of a stock market decrease using the 2005 data. On average, the predicted probability of decrease is higher for the days in which the market does decrease. Based on these results, we are able to correctly predict the direction of movement in the market in 2005 60% of the time.

Three Real-World Datasets

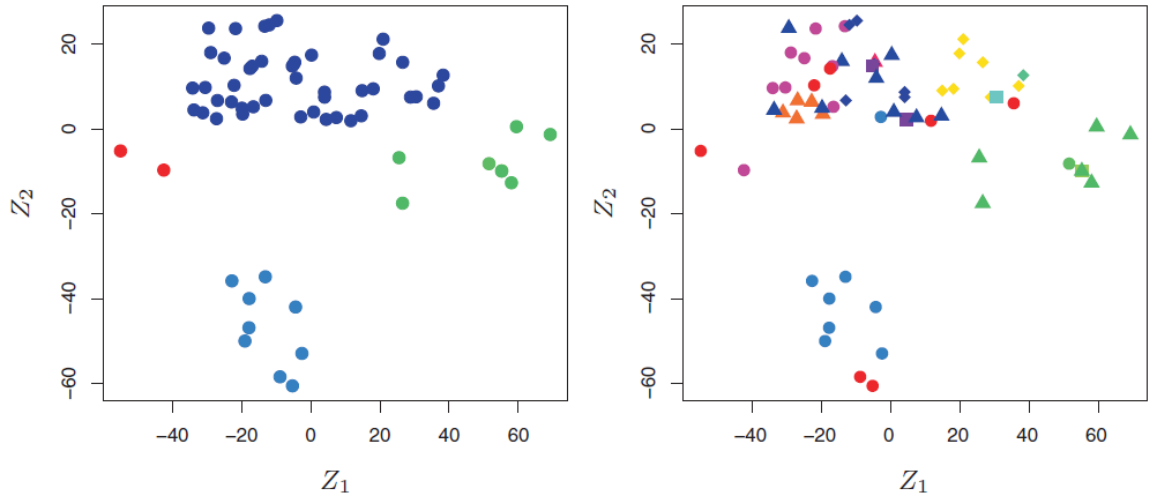
- Gene Expression Data
  - The previous two applications illustrate datasets with both input and output variables. Another important class of problems involves situations in which we only observe input variables, with no corresponding output.
  - For example, in marketing, we might have demographic information for a number of current or potential customers. We may wish to understand which types of customers are similar to each other by grouping individuals according to their observed characteristics. This is known as a *clustering* problem.
  - Unlike the previous examples, here we are not trying to predict an output variable.

Three Real-World Datasets

- Gene Expression Data
  - In Machine Learning 2 we discuss machine learning methods for problems in which no natural output variable is available.
  - As an example here, we consider the NCI60 data set, which consists of 6,830 gene expression measurements (the so-called “feature set”) for each of 64 cancer cell lines.
  - Instead of predicting a particular output variable, we are interested in determining whether there are groups, or clusters, among the cell lines based on their gene expression measurements.
  - This is a difficult question to address, in part because there are thousands of gene expression measurements per cell line, making it hard to visualize the data.

Three Real-World Datasets

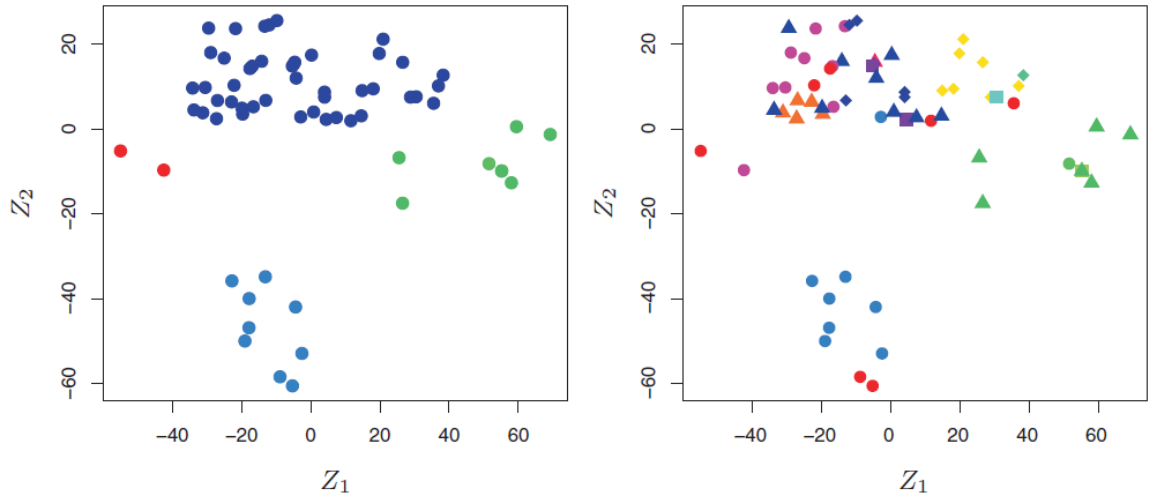
- Gene Expression Data



- The left-hand panel of above represents each of the 64 cell lines using just two numbers, *Z*1 and *Z*2.
- These are the first two *principal components* of the data, which summarize the 6,830 expression measurements for each cell line down to two numbers or *dimensions*.
- This “dimension reduction” has resulted in some loss of information, but it is now possible to visually examine the data for evidence of clustering.

Three Real-World Datasets

- Gene Expression Data



- The right panel is the same as left panel except that we have represented each of the 14 different types of cancer using a different colored symbol.
- Observe that cell lines corresponding to the same cancer type tend to be nearby in the two-dimensional space.

A Brief History

- At the beginning of the nineteenth century, Legendre and Gauss published papers on the method of least squares, which implemented the earliest form of what is now known as linear regression.
- In order to predict categorical values, Fisher proposed linear discriminant analysis in 1936. In the1940s, various authors put forth an alternative approach, logistic regression.
- In the early 1970s, Nelder and Wedderburn coined the term generalized linear models (GLM) for an entire class of statistical learning methods that include both linear and logistic regression as special cases.

A Brief History

- By the end of the 1970s, many more techniques for learning from data were available. However, they were almost exclusively linear methods because fitting non-linear relationships was computationally infeasible at the time.
- By the 1980s, computing technology had finally improved sufficiently that non-linear methods were computationally tractable. In mid-1980s Breiman, Friedman, Olshen and Stone introduced classification and regression trees and cross-validation for model selection.

A Brief History

- **Hastie and Tibshirani** coined the term generalized additive models (GAMs) in 1986 for a class of non-linear extensions to generalized linear models, and also provided a practical software implementation.
- Since that time, machine learning has emerged as a new subfield in statistics, focused on supervised and unsupervised modeling and prediction. In recent years, progress has been marked by the increasing availability of powerful and relatively user-friendly software, such as R.
- This has the potential to continue the transformation of the field from a set of techniques used and developed by statisticians and computer scientists to an essential toolkit for a much broader community.

The Premises of the Course

1. Machine learning methods should not be viewed as a series of black boxes. No single approach will perform well in all possible applications.
   - Without understanding all of the cogs inside the box, or the interaction between those cogs, it is impossible to select the best box.
   - Hence, we will always carefully describe the model mechanics, and the intuition, assumptions, and trade-offs behind each of the methods that we consider.

The Premises of the Course

2. While it is important to know what job is performed by each cog in the box, it is not necessary to have the skills to construct the machine inside the box!
    - Thus, we will minimize our discussion of technical details related to fitting procedures and theoretical properties. We assume that students are comfortable with basic mathematical concepts, but we do not assume a graduate degree in mathematics.

The Premises of the Course

3. It is presumed that students are interested in applying machine learning methods to real-world problems.
   - In order to facilitate this, as well as to motivate the techniques discussed, the course includes extensive hands-on work whereby students work through realistic applications of each method considered.

Access To Data For Labs And Exercises

- We will require data from various sources for the labs and assignments that accompany each major topic. Our sources will include:
  - Data provided in the base R distribution
  - Data provided by various R packages, including data associated with the exercises in the text.
    - These latter files are contained an a package named ISLR, or on the author's web site at http://www.StatLearning.com
  - Data from other sources, for which links or data downloads will be provided on the course web site.