Introduction to Statistical Learning

Topic 3. Statistical learning - a high-level overview and illustrative examples

3.1. What is statistical learning?

Part One

Sonja Petrović
Created for ITMD/ITMS/STAT 514

Spring 2021.

Goals of this lecture

- Setting the context: data mining
- A few illustrations on applications of statistical learning
- Stat.Learn.:
  - what is it?
  - how is it done?
- Resources and links

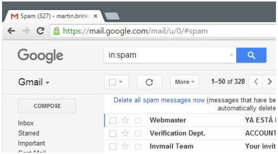
Section 1

Setting the context: data mining

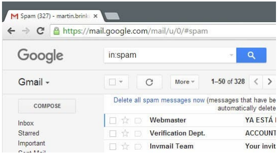
What is data mining?

Data mining is the science of discovering structure and making predictions in large or complex data sets.

Spam filtering, Fraud detection, Event detection

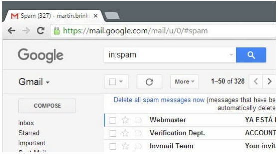


Spam filtering, Fraud detection, Outbreak detection



- How can we tell apart **spam** from real emails?
- How do we identify **fraudulent** transactions?
- Is the president's **tweet** going viral?

Spam filtering, Fraud detection, Outbreak detection



- How can we tell apart **spam** from real emails?
- How do we identify **fraudulent** transactions?
- Is the president's **tweet** going viral? Is the **flu** going viral?

Recommendation systems

- Which **movies** should I recommend to my customers?
- How can I identify individuals with **similar viewing/purchasing** preferences?
- Which **products** should I recommend to my customers?
- Which **promotional offers** should I send out, and to whom?

Precision medicine, health analytics

... And many more applications (content tagging in images; text mining; ...)

Thinking about Data Mining problems

Data mining problems are often divided into predictive tasks and descriptive tasks.

- Predictive Analytics (Supervised learning):
  - Q: To whom should I extend credit?

Thinking about Data Mining problems

Data mining problems are often divided into predictive tasks and descriptive tasks.

- Predictive Analytics (Supervised learning):
  - Q: To whom should I extend credit?
    - Task: Predict how likely an applicant is to repay loan.
  - Q: What characterizes customers who are likely to churn?

Thinking about Data Mining problems

Data mining problems are often divided into predictive tasks and descriptive tasks.

- Predictive Analytics (Supervised learning):
  - Q: To whom should I extend credit?
    - Task: Predict how likely an applicant is to repay loan.
  - Q: What characterizes customers who are likely to churn?
    - Task: Identify variables that are predictive of churn.
  - Q: How profitable will this subscription customer be?

Thinking about Data Mining problems

Data mining problems are often divided into predictive tasks and descriptive tasks.

- Predictive Analytics (Supervised learning):
    - Q: To whom should I extend credit?
        - Task: Predict how likely an applicant is to repay loan.
    - Q: What characterizes customers who are likely to churn?
        - Task: Identify variables that are predictive of churn.
    - Q: How profitable will this subscription customer be?
        - Task: Predict how long customer will remain subscribed.

Thinking about Data Mining problems

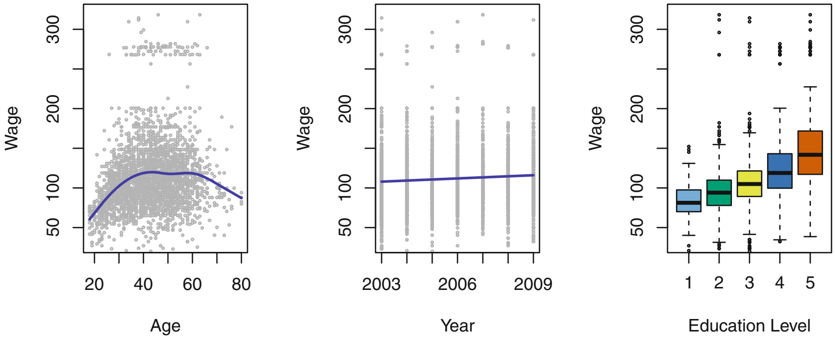
Data mining problems are often divided into predictive tasks and descriptive tasks.

- Predictive Analytics (Supervised learning):
  - Q: To whom should I extend credit?
    - Task: Predict how likely an applicant is to repay loan.
  - Q: What characterizes customers who are likely to churn?
    - Task: Identify variables that are predictive of churn.
  - Q: How profitable will this subscription customer be?
    - Task: Predict how long customer will remain subscribed.
- Descriptive Analytics (Unsupervised learning):
  - Clustering customers into groups with similar spending habits
  - Learning association rules: E.g., 50% of clients who {recently got promoted, had a baby} want to {get a mortgage}

Section 2

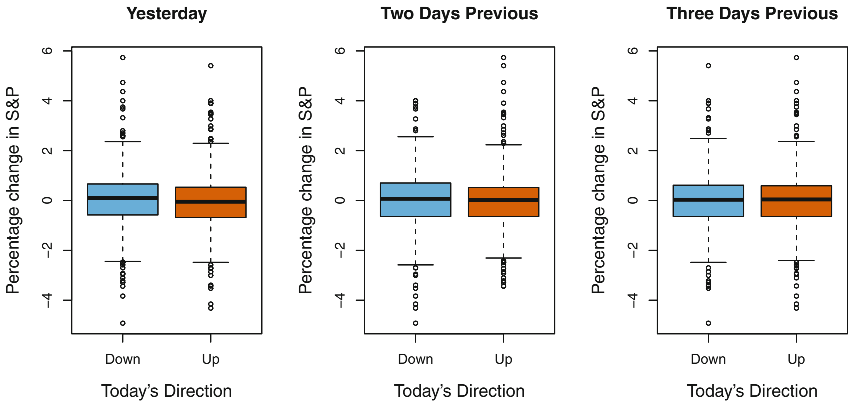
A few illustrations on applications of statistical learning

Wage data



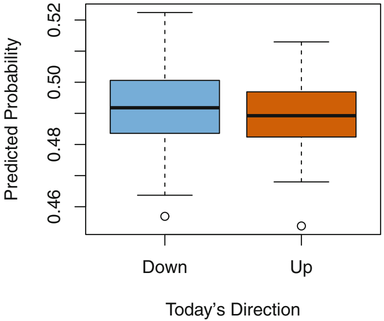
Wage data, which contains income survey information for males from the central Atlantic region of the United States. **Left**: `wage` as a function of `age`. On average, `wage` increases with `age` until about 60 years of age, at which point it begins to decline. **Center**: `wage` as a function of `year`. 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`.

Stock Market data



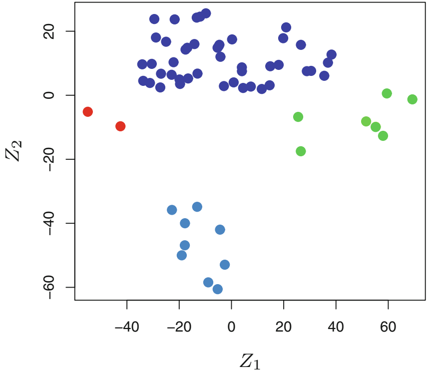
Left: Boxplots of the previous day's percentage change in the S&P index for the days for which the market increased or decreased, obtained from the `Smarket` data. **Center and Right**: Same as left panel, but the percentage changes for 2 and 3 days previous are shown.

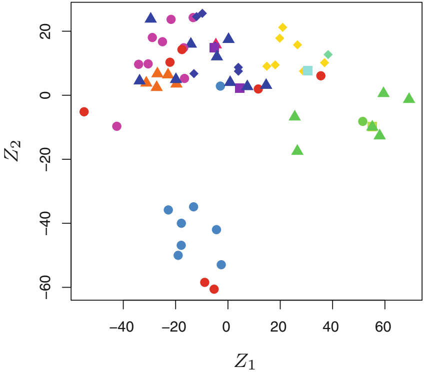
Stock Market data



We fit a quadratic discriminant analysis model to the subset of the `Smarket` data corresponding to the 2001–2004 time period, and predicted 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 60% of the time.

Gene Expression Data





Left: Representation of the NCI60 gene expression data set in a two-dimensional space, Z1 and Z2. Each point corresponds to one of the 64 cell lines. There appear to be four groups of cell lines, which we have represented using different colors. **Right**: Same as left panel except that we have represented each of the 14 different types of cancer using a different colored symbol. Cell lines corresponding to the same cancer type tend to be nearby in the two-dimensional space.

Section 3

What is statistical learning?

Uncovering relationships

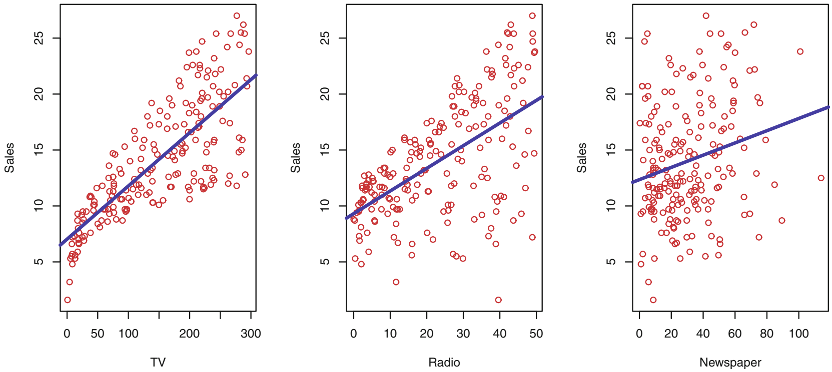


Figure 1: ISLRfig2.1: The `Advertising` data set. The plot displays `sales`, in thousands of units, as a function of `TV`, `radio`, and `newspaper` budgets, in thousands of dollars, for 200 different markets. In each plot we show the simple least squares fit of `sales` to that variable... In other words, each blue line represents a simple model that can be used to predict `sales` using `TV`, `radio`, and

Estimating f

- More generally, suppose that we observe a quantitative response Y and p different predictors, $X_1, X_2, \ldots, X_p$.
- We assume that there is some relationship between Y and $X = (X_1, X_2, ..., X_p)$, which can be written in the very general form

$$Y = f(X) + \epsilon.$$

- f is some fixed but unknown function of $X_1, X_2, \ldots, X_p$.
- ϵ is a random error term, which is independent of X and has mean zero.
- In this formulation, f represents the *systematic* information that X provides about Y.

Another example

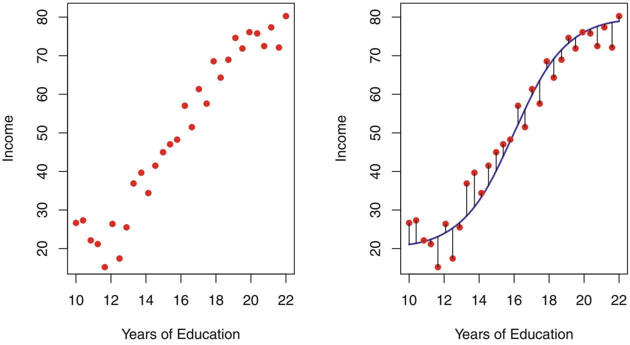


Figure 2: ISLRfig2.2: The `Income` data set. *Left*: The red dots are the observed values of `income` (in tens of thousands of dollars) and `years of education` for 30 indi- viduals. *Right*: The blue curve represents the true underlying relationship between `income` and `years of education`, which is generally unknown (but is known in this case because the data were simulated). The black lines represent the error associated with each observation. Note that some errors are positive (if an ob- servation lies above the blue curve) and some are negative (if an observation lies below the curve). Overall, these errors have approximately mean zero.

Summary: essence of statistical learning

Takeaway

Statistical learning refers to a set of approaches for estimating f.

- Let's outline some of the key theoretical concepts that arise in estimating f,
- as well as tools for evaluating the estimates obtained.

Why estimate f?

Two reasons: prediction and inference.

Why estimate f?

Two reasons: prediction and inference.

Why estimate f?

Prediction

- A set of inputs X are readily available,
- but the output Y cannot be easily obtained.
- Since the error term averages to zero, we can predict Y using

$$\hat{Y} = \hat{f}(X),$$

where $\hat{f}$ represents our estimate for f, and $\hat{Y}$ represents the resulting prediction for Y.

  - $\hat{f}$ = a black box.

Why estimate f?

Prediction

- A set of inputs X are readily available,
- but the output Y cannot be easily obtained.
- Since the error term averages to zero, we can predict Y using

$$\hat{Y} = \hat{f}(X),$$

where $\hat{f}$ represents our estimate for f, and $\hat{Y}$ represents the resulting prediction for Y.
  - $\hat{f}$ = a black box.

Example

- $X_1, X_2, \ldots, X_p$ are characteristics of a patient's blood sample that can be easily measured in a lab,
- Y is a variable encoding the patient's risk for a severe adverse reaction to a particular drug.

It is natural to seek to predict Y using X: we can then avoid giving the drug in question to patients who are at high risk of an adverse

Why estimate f?

Two reasons: prediction and inference.

Why estimate f?

Two reasons: prediction and inference.

Why estimate f?

Inference

- We are often interested in *understanding the way* that Y is affected as $X_1, X_2, \ldots, X_p$ change.
- Estimate f; not necessarily to make predicitons for Y, but **understand the relationship** between X and Y:
  - how Y changes as a function of $X_1, X_2, \ldots, X_p$.
  - $\hat{f}$ can't be a black box.
  - Questions:
    - Which predictors are associated with the response?
    - What is the relationship between the response and each predictor?
    - an the relationship between Y and each predictor be adequately summarized using a linear equation, or is the relationship more complicated?

Why estimate f?

Inference

- We are often interested in *understanding the way* that Y is affected as $X_1, X_2, \ldots, X_p$ change.
- Estimate f; not necessarily to make predicitons for Y, but **understand the relationship** between X and Y:
  - how Y changes as a function of $X_1, X_2, \ldots, X_p$.
  - $\hat{f}$ can't be a black box.
  - Questions:
    - Which predictors are associated with the response?
    - What is the relationship between the response and each predictor?
    - an the relationship between Y and each predictor be adequately summarized using a linear equation, or is the relationship more complicated?

Example - Advertising data set

- Which media contribute to sales?
- Which media generate the biggest boost in sales? or
- How much increase in sales is associated with a given increase in TV advertising?

How Do We Estimate f?

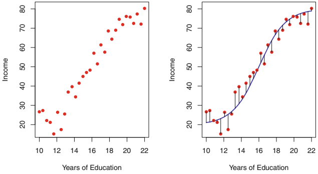


Figure 3: Recall this example

- Observed n = 30 data points.
- These observations are called the training data because we will use these observations to *train*, or *teach*, our method how to estimate f.
- we apply a statistical learning method to the training data in order to estimate the unknown function f.

"We apply a statistical learning method to the training data in order to estimate the unknown function f:"

Find a function $\hat{f}$ such that $Y \approx \hat{f}(X)$ for any observation (X, Y).

"We apply a statistical learning method to the training data in order to estimate the unknown function f:"

Find a function $\hat{f}$ such that $Y \approx \hat{f}(X)$ for any observation (X, Y).

1. Parametric methods
   - Select model (make an assumption about the functional form, or shape, of f; e.g., linear, say)
   - Train/fit the model (e.g. ordinary least squares, say).
2. Non-parametric methods:
   - do not make explicit assumptions about the functional form of f.
   - Seek an estimate of f that gets as close to the data points as possible without being too rough or wiggly.

Parametric vs. nonparametric:

- Non-parametric advantage:
  - by avoiding the assumption of a particular functional form for f, they have the potential to accurately fit a wider range of possible shapes for f.
  - Any parametric approach brings with it the possibility that the functional form used to estimate f is very different from the true f, in which case the resulting model will not fit the data well.
  - In contrast, non-parametric approaches completely avoid this danger, since essentially no assumption about the form of f is made.
- Non-parametric disadvantage:
  - since they do not reduce the problem of estimating f to a small number of parameters, a very large number of observations (far more than is typically needed for a parametric approach) is required in order to obtain an accurate estimate for f.

Example

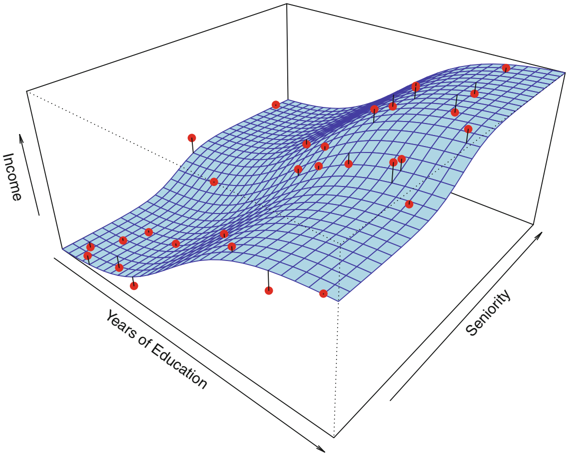


Figure 4: ISLRfig2.3. The plot displays income as a function of years of education and seniority in the Income data set. The blue surface represents the true underlying relationship between income and years of education and seniority, which

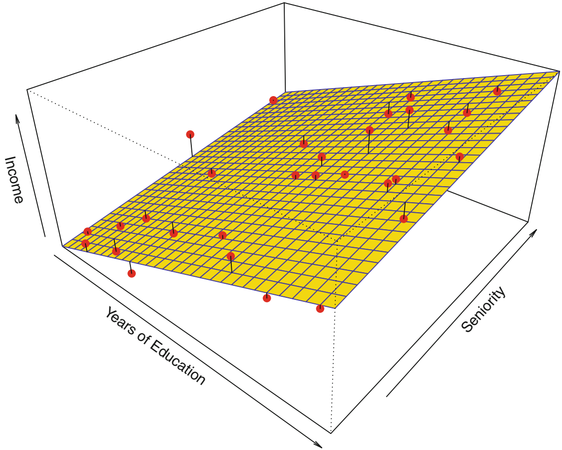


Figure 5: A linear model fit by least squares to the Income data from Figure prev page. The observations are shown in red, and the yellow plane indicates the least squares fit to the data.

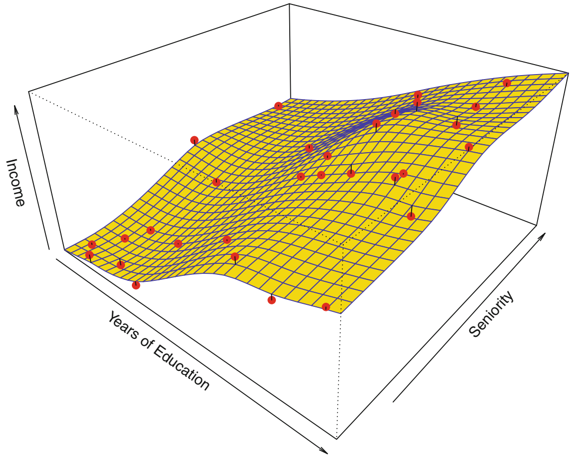


Figure 6: ISLRfig2.5. A smooth thin-plate spline fit to the Income data is shown in yellow; the observations are displayed in red.

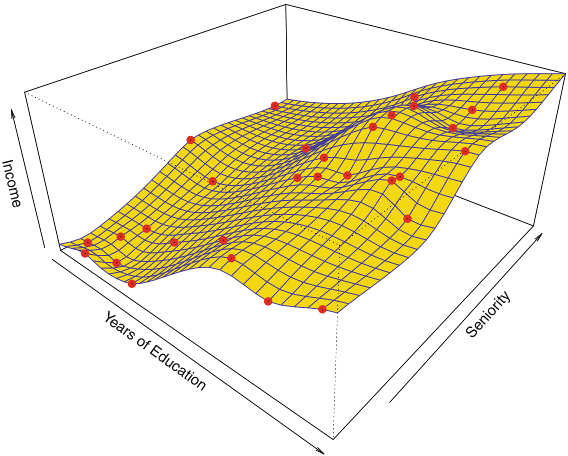


Figure 7: ISLRfig2.6. A rough thin-plate spline fit to the Income data. This fit makes zero errors on the training data.

Conclusion and where to next

What does it mean to be a *good predictor*?

stay tuned; here's a quick example.

License

Content of this lecture is based on the first two chapters of the textbook Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani, ' An Introduction to Statistical Learning: with Applications in R'. The book is available online.

Part of this lecture notes are extracted from Prof. Alexandra Chouldechova data mining notes CMU-95791, released under a Attribution-NonCommercial-ShareAlike 3.0 United States license.

Code for generating the stat plots will be released on the course site.