Introduction to Statistical Machine Learning

Rebecca C. Steorts, Duke University

STA 325, Chapter 2 ISL

Agenda

- Motivation
- Exploring the data
- Statistical Machine Learning
- Prediction, Inference, Uncertainty Quantification

Motivation

- You are an analyst hired by a client to improve the sales on a product.
- The Advertising data set consists of sales of a product in 200 different markets.
- There are advertising budgets for three different media sources: TV, radio, newspaper.

Advertising data set

```r
ad <- read.table("data/Advertising.csv",
    header=TRUE, sep=",")
head(ad)
```

```
##   X    TV Radio Newspaper Sales
## 1 1 230.1  37.8      69.2  22.1
## 2 2  44.5  39.3      45.1  10.4
## 3 3  17.2  45.9      69.3   9.3
## 4 4 151.5  41.3      58.5  18.5
## 5 5 180.8  10.8      58.4  12.9
## 6 6   8.7  48.9      75.0   7.2
```

Exploratory data analysis (EDA)

```r
x11(width=5, height=2, pointsize=12)
pdf("examples/sales.pdf",width=5,height=3)
par(mfrow=c(1,3))
plot(ad$TV, ad$Sales, xlab="TV", ylab="Sales")
plot(ad$Radio, ad$Sales, xlab="Advertising", ylab="Sales")
plot(ad$Newspaper, ad$Sales, xlab="Newspaper", ylab="Sales'
dev.off()
```

```
## pdf
##   2
```

Plotting the EDA

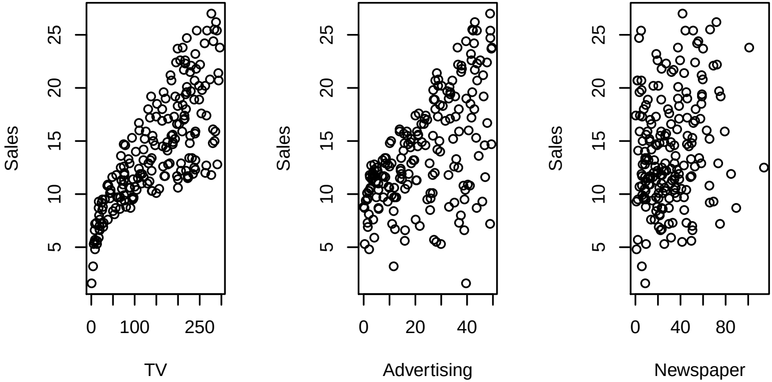

Figure 1: The Advertising data set, where we see sales (thousands of dollars) plotted against TV, Advertising, Newspaper, respectively. In Chapter 3, we will learn one simple way (regression) to perform prediction.

Terminology and notation

- In our example, the advertising budgets are **input variables** while sales is an **output variable**
- X: input variables (predictors, features, indepedent variables)
- We distinguish these via indices (X_1, X_2, X_3) for example as in the case of the TV, advertising, newspaper variables.
- Y: **ouput variable** (response or dependent variable).

Terminology and Notation

- Suppose we observe a quantitiative response Y and p different predictor variables
$$X_1, \ldots, X_p.$$
- We assume there is some relationship between Y and $(X_1, \ldots, X_p)$, which can be written as

$$Y = f(X) + \epsilon. \tag{1}$$

- f is some **fixed** but **unknown** function of $X_1, \ldots, X_p$.
- ϵ is a random variable that is independent of X and has **mean zero**.

Statistical Learning Recap

Statistical learning at it's core refers to a set of approaches for estimating or learning f.

We now turn to key ways we can estimate f and evaluate the accuracy of our estimate.

Statitical Learning

Statistical Learning refers to the set of approaches for estimating f.

We now outline

- some of the key mathematical concepts that are needed to estimate f and
- tools for evaluating the corresponding estimates that are obtained.

Prediction versus Estimation of f

Depending on the motivation of the data set, we may wish to perform prediction or estimation (or both)!

We discuss both.

Prediction of f

In many situations, a set of inputs X are available but the output Y is difficult to obtain.

One easy way to predict Y is using

$$\hat{Y} = \hat{f}(X), \tag{2}$$

- $\hat{f}$ is our estimate of f
- $\hat{Y}$ is our prediction of Y.

Remark: In this setting, we are treating the form of $\hat{f}$ as a black box as one is not concerned about the form **provided** it gives **accurate predictions** for Y.

Accuracy of $\hat{Y}$

We measure the accuracy of $\hat{Y}$ predicting Y using two quantities. Speficially, there are two error that we quantify.

1. $\hat{f}$ will not be a perfect estimation for f, and introduces **reducible** error since it is possible to improve the accuracy of $\hat{f}$ (using a better machine learning technique).
2. However, even if we can perfect our estimate of $\hat{f}$ such that $\hat{f} = f(X)$, our prediction has error!

- Recall that Y is a function of ϵ which cannot be predicted using X.
- The variability of ϵ affects the accuracy of our predictions. This is the **irreducible** error since we cannot reduce the error that is introduced by ϵ.

Inference

On the other hand, we are often interested in the way that Y is affected as $X_1, \ldots, X_p$ change.

- We wish to estimate f but we don't necessarily wish to make predictions for Y.
- Instead, we want to understand the **relationship** between X and Y.
- That is, how does Y change as a function of $X_1, \ldots, X_p$.

Remark: Now $\hat{f}$ cannot be treated as a black box because we **must know** it's **exact form.**

Inferential questions

Some inferential questions of interest are the following:

1. Which predictors are associated with the response? Specifically, what are the **most important features** among a large set of possible variables?
2. What is the relationship between the response and each preditor? Some precitors have a postiive relationship with Y and others have the oppositive relationship.
3. Can the relationship between Y and each predictor be adequately summarized using a linear equation or is the realtionship more complicated. Many methods for estimating f take a linear form in the predictors. In some cases, such an assumption is undesirable since the true relationship is very complex.

How does one estimate f?

This is what the main goal of this course will address

- Parametric methods (e.g., regression)
- Nonparametric methods (e.g., splines)

We first introduce terminology that we use for the remainder of the course.

Terminology

- We will always assume that we have ovserved a set of n different **data points** (or **training data**).
- These are called **training data** since we use these observations to train our method how to estimate f.
- Our training data consist of

$$\{(x_1, y_1) \ldots (x_n, y_n)\}$$

where $x_i = (x_{i1}, \ldots, x_{ip})^T$.

- Our goal is to apply a machine learning method to the training data to estimate the unknown function f.

Parametric vs Non-Parametric Methods

A **parametric** method has a two-step approach

1. Make an assumption of the functional form of f (perhaps linear).
2. After the form of f is selected, we need a procedure in place to fit or train our model.

They assumed functional form could be quite different from the true form, leading to low accuracy, however, a large number of observations is not needed.

Parametric vs Non-Parametric Methods

Non-parametric methods do not make any explict assumptions about the functional form of f.

- They seek to be **flexbile** and get as close as possible to the underlying data points.

Very flexible and can potentially have better accuracy, but a large number of observations is needed in order to acheive high accuracy.

Assessing Model Accuracy

We discuss some important concepts that arise for selecting machine learning methods in practice. The first we present are

- The mean square error
- And the variance bias trade-off

The Mean Squared Error (MSE)

- We need a way to quantify for a given data set the extent to which the predicted response value for a given observation is close to the true response value for that same observation.
- In the regression setting, the most commonly used measure is the **mean squared error (MSE)**, given by

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2, \tag{3}$$

where $\hat{f}(x_i)$ is the prediction that $\hat{f}$ gives of the ith observation.

Remark: The MSE will be **small** if the predicted responses are close to the true responses and will be **large** if some of the observations, the predicted and true differ greatly.

Remark: The MSE above is calculated using the training data, so technically this is the **training MSE.**

The Mean Squared Error (MSE)

- In general, we do not really care how well the method works on the training data.
- We are interested in the accuracy of the predictions that we obtain when we apply our method to previously unseen test data.

The Mean Squared Error (MSE)

1. Fit our statistical model on $\{(x_1, y_1), \ldots (x_n, y_n)\}$ and obtain $\hat{f}$.
2. Then compute $\hat{f}(x_1), \ldots, \hat{f}(x_n)$.

- What we really want to know is if $\hat{f}(x_o)$ is approximately x_o where (x_o, y_o) is a previously unseen test observation not used to train the statistical learning method.
- We wish to choose the method with the lowest test MSE. If we have a large number of observations, we can compute

$$\text{Ave}(\hat{f}(x_o) - y_o)^2.$$

The Mean Squared Error (MSE)

3. We'd like to select the model for which the average of this quantity — the test MSE — is as small as possible.

Remark: In practice, one can usually compute the training MSE with relative ease, but estimating test MSE is considerably more difficult because usually no test data are available

- We will discuss through the course how to compute the test MSE (such as cross validation).

Bias Variance Trade Off

What do we mean by the variance and bias of a statistical learning method?

- Variance refers to the amount by which $\hat{f}$ would change if we estimated it using a different training data set.
- Bias refers to the error that is introduced by approximating a real-life problem, which may be extremely complicated, by a much simpler model.

Bias Variance Trade Off

It is possible to show that the expected test MSE (for a given value of x_o) can always be decomposed into three fundamental quantities:

- the **variance** of $\hat{f}(x_o)$
- the squared **bias** of $\hat{f}(x_o)$
- and the **variance** of the estimator in terms of ϵ.

$$E[(y_o - \hat{f}(x_o))^2] = \mathbb{V}(\hat{f}(x_o)) + [\mathtt{Bias}(f(x_o))]^2 + \mathbb{V}(\epsilon) \tag{4}$$

Supervised versus Unsupervised Learning

Most statistical learning techniques fall into two categories:

1. Supervised
2. Unsupervised

Supervised versus Unsupervised Learning

Most of the methods in ISLR are **supervised learning techniques**.

This means that for each predictor x_i $i = 1, \ldots, n$ there is an associated response variable y_i,

Our goal is to fit a model relating the response (y) to the predictors (x) such that we can accurately predict future responses (prediction) or such that we can better understand the relationship between the response and the predictor (inference).

Examples of this include regression, logistic regression, boosting, and support vector machines.

Supervised versus Unsupervised Learning

Chapter 10 covers unsupervised learning techniques.

Unsupervised learning covers a more challenging situation where for every observation x_i that we observe, we do not observe a response variable y_i.

Therefore, we cannot fit a linear regression model since we don't have a reponse variable!

In this setting, we lack the supervision of a response y_i to guide the x_i for model fitting, so we develop other methods to guide our analysis.

Typically we are guided by the data!

Supervised versus Unsupervised Learning

One statistical learning tool that we will learn about is clustering.

The goal of clustering is to ascertain, on the basis of $x_1, \ldots, x_n$, where the observations fall into relatively distinct groups.

Another goal may just be to do an exploratory data analysis on $x_1, \ldots, x_n$ and perform some dimension reduction and visualize the features. This is known as principle components analysis.