6.1: Introduction to Statistical Learning

Navona Calarco

The University of Toronto

Table of contents I

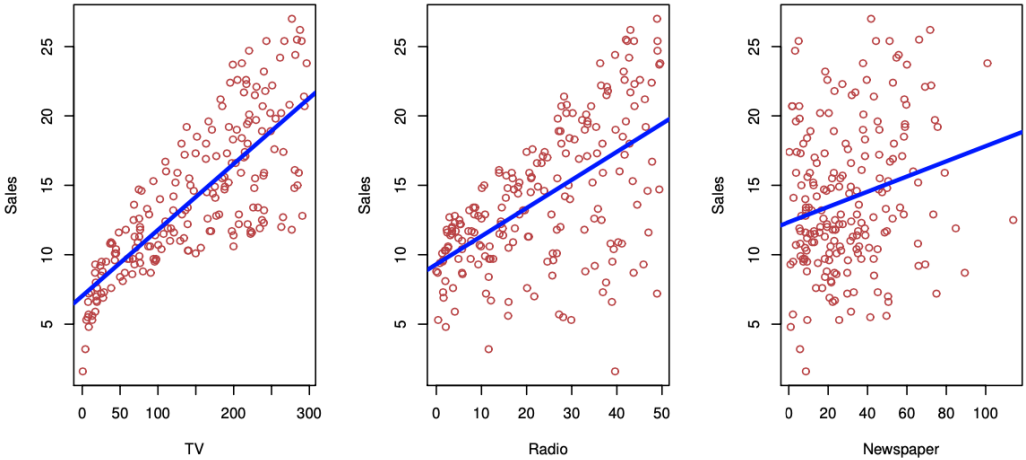
What is Statistical Learning?

Suppose we want to figure out the association between the allocation of advertising budgets and sales in order to increase sales for a client.

- There are three types of advertising: TV, radio, and newspaper which can be labelled $X_1, X_2, and\ X_3$ respectively.
- The advertising budgets are the independent variables, or **predictor variables**.
- The sales is the dependent variable, or **response variable** which we label Y.When you click the **Render** button a document will be generated that includes:
- Each observation in the dataset has a value for the predictors X_1, X_2, X_3, and the response Y.

What is Statistical Learning?

The sales in relation to each of the advertising budgets are shown along with a simple fitted line for the relationships.



What is Statistical Learning?

We want to find the relationship between the predictor variables and the response variable.

We can assume that there is a relationship between $X = (X_1, X_2, X_3)$ and Y which can be written in the general form

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

- f is a fixed unknown function of the predictor variables.
- ϵ is a random error term which has mean zero.

Statistical learning is summarized by the set of approached which are used to estimate f.

Prediction vs Inference

The are two main reasons for why we want to estimate f:

1. **Prediction** We want to know what response is expected given a set of predictors. Ex: What income is expected for a given level of education and seniority?
2. **Inference** We want to understand how the response variable is affected by changes in the predictors. Ex: To what extent is income associated with education?

Prediction

Prediction problems often arise when the predictor variables X are known but the response Y is not easily obtained. Recall the general form for the relationship between the predictor and response variables

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

Since the error term ϵ averages to zero, we can go about predicting Y using

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

We use " \$\hat \ \$ " to denote estimates. That is, $\hat{Y}$ is an estimate for Y and $\hat{f}$ is an estimate for f.

Prediction

The accuracy of our prediction $\hat{Y}$ depends on:

1. The **reducible error**: the error in our estimate $\hat{f}$. This error is reducible since estimates can always be improved.
2. The **irreducible error**: the random error associated with the true response $Y = f(x) + \epsilon$. (Even if $\hat{f} = f$, $\hat{Y}$ will still have error associated with its prediction since ϵ is not a function of X.)

Inference

We want to know how the predictor variables and the response variable are related.

1. *Which predictors affect the response?*
2. *Is a linear equation a good approximation for the relationship between the predictors and the response?*

In each case, we do not want to make predictions for Y using $\hat{f}$, we want to find the true form of f.

How do we Estimate f?

Assume that we have n observations in our data set. The standard approach is to split the data set into training data and testing data.

- **training data** is used to train or teach the statistical method we are using to estimate f.
- **testing data** is used to test the accuracy of the resulting estimate for f on new data.

In general, the statistical learning methods used to estimate f can be characterized as **parametric**, and **non-parametric**.

Parametric Methods

This approach is implemented in two steps:

- Make an assumption about the functional form of f.
- Develop a procedure to fit the model to the training data.

Example

- Suppose f is linear in X: $f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$. This assumption has reduced the number of parameters that need to be fit significantly compared to fitting a generic p-dimensional function.
- We need to estimate $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ such that $Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$. One approach is to use least squares which attempts to minimize the difference between the data and our estimate for f.

Non-Parametric Methods

This method does not make an assumption about the form of f. Instead, the goal of non-parametric fitting it to create an estimate for f that follows the data as close as possible while staying as "smooth" as possible.

- This avoids the danger seen in parametric approaches where the functional form of the estimate could be completely different from the true f.
- This approach requires many parameters to be fit since the from of f needs to be flexible.
- As a result, many observations are needed in order to get an accurate estimate for f.
- This approach could lead to **overfitting**, in which f follows the noise and random variation in our data too closely.
- Splines are an example of non-parametric fitting.

Accuracy-Interpretability Trade-Off

- The methods we will introduce have different levels of restrictiveness or flexibility.
- Choosing a model on the basis of flexibility will depend on the problem at hand.
  - If we are interested in inference, restrictive models are much more interpretable (i.e. the relationship between the predictors and the response is more clear)
  - If we are only interested in prediction accuracy, flexible models **might** perform better.
- Flexible models will not always provide better predictions since they are very prone to overfitting!

Supervised vs Unsupervised Learning

- **Supervised learning** involves models for predicting a response based on predictor variables
  - Examples: linear regression, boosting, support vector machines (SVM)
- **Unsupervised learning** refers to models used to investigate features associated with observations.
  - There is no response variable to predict.
  - The goal is to understand the relationship between variables or observations.
  - Example: clustering
- **Semi-supervised learning** involves a set of observations, some of which have both predictor and response variables and some with only predictor variables. (Not covered in this module)

Regression vs Classification Problems

- Variables can be either qualitative or quantitative.
  - **Quantitative** variables have numerical values (ex: age, monetary value, etc.)
  - **Qualitative** variables are categorical values (ex: {small, medium, large} or {yes, no})
- Problems that involve quantitative response variables are **regression** problems.
- Problems that involve qualitative response variables are **lassification** problems.
- This is a bit of a generalisation (logistic regression is a classification method but its output is numerical so can be thought of as a regression method as well)

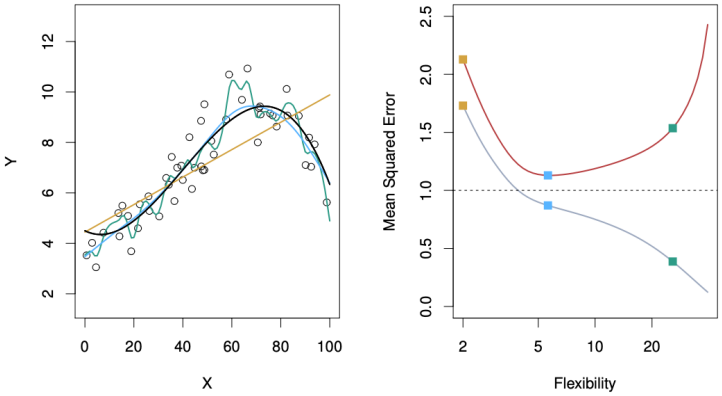
Assessing Model Accuracy

There is no method that works the best on all data sets so we need a way to assess the quality of the model's fit to the data.

- Recall that we use the training data set to fit the model.
- Then the test data set is used to see how well the model performs on new data by computing some test statistic.
- For regression, the **mean squared error** (MSE) is most common which measures how close the predicted responses are to the true responses.
- We can compute the MSE for both the training data and the test data.
  - The training MSE will usually be lower (better) than the test MSE.
  - We want to choose the model that minimizes the test MSE since we only really care about the performance of the model on new data.

Assessing Model Accuracy

The left plot shows three different models fit to data. The right plot shows the training MSE (grey) and test MSE (red) versus the flexibility of the model.



- The least flexible model (orange) has the worst training and test MSE so it does not fit the data well.
- The most flexible model (green) has the lowest training MSE but a much higher test MSE so the model is overfit.
- The blue model has the lowest test MSE so it is the winner.

Assessing Model Accuracy

What if no test observations are available?

- As we saw, there is no guarantee that the model that minimises the training MSE will minimise the test MSE.
- As model flexibility increases, the training MSE decreases but the test MSE may not.
- A model is said to be overfit if a less flexible model would results in a lower test MSE.
- An overfit model is picking up on patterns in the training data that are not in the test data.
- We will cover many approaches for estimating the test MSE from training data.

The Bias-Variance Trade-Off

The test MSE will always exhibit a U-shaped curve as a function of model flexibility. This is because the expected test MSE for some observation x_0 is the sum of:

- $\text{Var}\left(\hat{f}\left(x_0\right)\right)$: the variance of $\hat{f}\left(x_0\right)$ (the variance of the predicted response for the test observation x_0 given many $\hat{f}$ fit on different training sets).
- $\left[\text{Bias}\left(\hat{f}\left(x_0\right)\right)\right]^2$: The squared bias of $\hat{f}\left(x_0\right)$.
- $\text{Var}(\epsilon)$: the variance of the error terms ϵ.

The Bias-Variance Trade-Off

In order to minimise the expected test error we need to use statistical learning methods that result in low bias *and* low variance.

- **Variance** is the amount $\hat{f}$ would change if we fit it using a different training set.
  - More flexible models have higher variance since they fit the training data more closely.
- **Bias** is the error from approximating a complicated relationship with a simpler model.
  - More restrictive models have higher bias since they make more assumptions about the form of f.

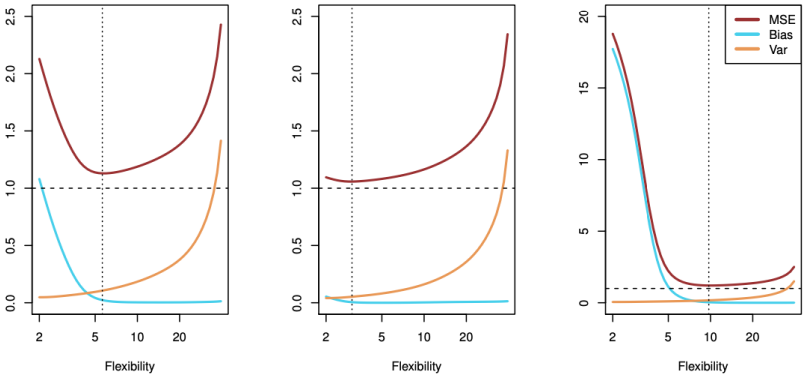
The Bias-Variance Trade-Off

The rate of change of the bias versus variance determines whether the test MSE will decrease or increase with flexibility.

- Initially, as the flexibility of the method increases, the bias decreases faster than the variance increases.
  - $\Rightarrow$ the test MSE declines.
- At some point the increasing flexibility has little impact on the bias but the variance increases significantly.
  - $\Rightarrow$ The test MSE increases. -This results in a U-shaped curve for test MSE vs method flexibility.

The Bias-Variance Trade-Off

A plot of the bias (blue), variance (orange), variance of the error (dashed line), and the test MSE (red) versus method flexibility for three different data sets.



- The MSE is the sum of the other three curves.
- The MSE cannot be smaller than the variance of the error (irreducible error).
- The middle model is close to linear so the test MSE immediately increases with flexibility.

Classification Model Accuracy

So far we have discussed model accuracy in the context of regression but the same idea apply to classification.

- The most common approach for assessing model accuracy is the **training error rate** which is the proportion of misclassified training observations.
- The **test error rate** is what we are actually interested in and hoping to minimise.
- The bias-variance trade-off is what controls the test error rate in the classification context as well.

References

Chapter 2 of the ISLR2 book:

James, Gareth, et al. "Statistical Learning." An Introduction to Statistical Learning: with Applications in R, 2nd ed., Springer, 2021.