Notes and exercises from "An Introduction to Statistical Leaning"" by James, Witten, Hastie & Tibshirani

Nicholas McGlincy

July 21st 2016

Contents

Chapter 1: Introduction

The introduction outlines three basic tasks in statistical learning:

1. **Regression** - predicting *continuous numerical* values from other continuous values
2. **Classification** - predicting class membership; has a *categorical* or *qualitative* output.

In both these tasks, we are trying to predict an output.

3. **Clustering** - here we are not trying to predict an output variable, but instead are trying to understand how similar/different a set of observations are, usually based on a number of variables.

   It is clear that I need to revise matrix multiplication, I didn't really understand the example on p. 12.

The book's website is here.

Chapter 2: Statistical Learning

Broadly, the goal of statistical learning is to estimate f in the equation:

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

Where:

Y is the *response/output/dependent variable.*
X is one or more *input/independent variables*, or *predictors* or *features.*
f is a function that relates X to Y, and
ϵ is a random error term, that is independent of X and has mean zero.

We estimate f for two ends:

1. To predict Y

Here we are not primarily concerned with the exact nature of the function, rather how accurately it predicts Y. Thus our estimate of Y comes from our estimate of f:

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

The accuracy of Y depends on two quantities:

I. The reducible error

$\hat{f}$ will be a more-a-less imperfect estimate of f. This inaccuracy is termed the *reducible error* as it is reducible by determining a better $\hat{f}$. This reduction is the main point of statistical learning.

II. The irreducible error

Even if our estimate of f was perfect, it would still have error in it, as Y is also a function of ϵ, which by definition cannot be predicted from X - hence *irreducible error.* This error may come from unmeasured (or unmeasurable) variables that might be useful in predicting Y, but since we didn't measure them, they are not included in f. The irreducible error will always provide an upper bound on the accuracy of our prediction of Y.

2. To infer the relationship between Y and $X_1, ..., X_p$

In order to do this we need to know the exact form of $\hat{f}$. We can them aim to answer the following questions: 1. Which predictors are associated with the response? 2. What is the relationship between the response and each predictor? 3. Can the relationship between Y and each predictor be adequately summarised using a linear equation, or is the relationship more complicated?

Some examples will combine both prediction and inference.

How do we estimate f?

We estimate f based on a subset of observations called the *training data.* We then use this function to predict the values of Y for new values of X.

There are two classes of methods for producing $\hat{f}$.

1. Parametric Methods

Parametric methods involve a two-step model-base approach. First, we make an assumption about the form of f; for example, that is it linear:

$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p$$

Where p is the number of variables X. The second step uses the the training data to *fit* or *train* the model. In this case, this means estimating the parameters $\beta_0, \beta_1, ..., \beta_p$ (the coefficients), such that:

$$Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p$$

The most common method for doing this is called *(ordinary) least squares*, but there are many others. Parametric approaches are termed this because they reduce the problem of estimating f to one of estimating a set of *parameters*. This makes things simplier, but has the potential disadvantage that the model we pick will usually not match the true form of f. We can try to fit a more flexible model, but this requires the estimation of more parameters; also, there is the risk of the model following the errors in the training set too closely, which is termed *overfitting*.

2. Non-parametric Methods

These methods do not make explicit assumptions about the functional form of f, instead they seek an estimate of f that gets as close to the data points as possible without being too wiggly. By avoiding assumptions about the form of f they have the potential for accurately fit a wider range of possible shapes of f. The disadvantage of non-parametric approaches is that they generally require a greater number of observations to obtain an accurate estimate of f.

The trade-off between prediction accuracy and model interpretability

Model flexibility described the variety of possible shapes of f that the method can estimate. More restrictive models are much more interpretable. Also, more flexible model, while they can initially seem more attractive, are much more prone to overfitting.

Supervised versus unsupervised learning

The majority of what has been discussed to this point has pertained to supervised methods, where were have predictor and response varibles, and we are trying to model how the predictor pertain to the response - either to predict the response to future observations, or to understand/infer the relationship between the response and the predictors.

Unsupervised learning describes situations where for each observation $i = 1, ..., n$ there are a number of observed measurements x_i, but no associated response y_i. In this situation, we can seek to understand the relationships between the variables or between the observations. An example is *clustering analysis*; the goal of clustering analysis is to ascertain, on the basis of $x_1, ..., x_n$, whether the observations fall into relatively distinct groups.

Semi-supervised learning methods exist – such as where collecting response data is more expensive than collecting predictor data – but they will not be discussed in the book.

Assessing model accuracy

No one method dominates all others for all possible datasets. Therefore we need methods to assess model accuracy, and be able make choices between different models.

Measuring the quality of fit

For a given observation, we need to quantify how close the predicted response was to the true response value for that observation. For regression, the most commonly used measure is the *mean squared error* (MSE), which is

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

Conceptual exercises

1. For each of the parts (a) through (d), indicate whether we would gnerally expect the prformance of a flexible statistical learning method to be better or worse than an inflexible method. Justify your answer.

   (a) The sample size n is extremely large, and the number of predictors p is small.
   (b) The number of predictors p is extremely large, and the number of observations n is small.
   (c) The relationship between the predictors and response is highly non-linear.
   (d) The variance of the error terms, i.e. $\sigma^2 = Var(\epsilon)$, is extremely high.

2.