Class Notes: Intro to Data Science (301033)

Ryan G

2019-03-13

Contents

Chapter 1

Prerequisites

These are topic Notes and Practicals for IntrotoDataSci, the material is:

Week no.	Topic
1	Introduction to Data Science
2	Supervised Learning: Linear Models 1
3	Supervised Learning: Linear Models 2
4	Supervised Learning: Classification (Logistic Regression)
5	Resampling Methods (e.g. Cross-Validation, Training Split etc.)
6	Supervised Learning: Tree Based Methods
7	Supervised learning: Classification (Support Vector Machines)
8	unsupervised learning: Principal Component Analysis
9	STUVAC
10	Unsupervised Learning: Clustering
11	Guest Lecture
12	Guest Lecture
13	Revision
14	Final Exam

In this unit we will cover:

- Supervised learning:
  - Linear Models
  - Classification (KNN and Discrimination)
  - Classification and Regression Trees
- Unsupervised Learning:
  - DimensionReduction: Principal Component Analysis
  - Clustering: K Means and Hierarchical
- Unstrucutred Data:
  - Text Mining
- Resampling
- Visualisation

Chapter 2

(Wk 1) Introduction to Data Science

Material of 5 March 2019, week 1 test ## Types of Data Data is classified as either structured or unstructured:

- ***Structured Data***
  - **Quantitative**/Numeric Data
    - Height, Weight, Salary etc.
  - Qualitative Data (also known as Categorical Data, Factors, or Discrete Variables)
    - be careful, factors usually refer to the variables in a predictive model
    - When dealing with factors in R it is necessary to use a data type called factors discussed below
      - Examples of categorical data include alive/dead, male/femaile, ethnicity, product code, hair colour etc.

2.0.1 Categorical variables in R

In order to deal with discrete variables R uses a data type called factors, to create factors the `factor()` command is used, within this command a vector containing factor levels must be enclosed, e.g.

```r
factor(c("Male", "Female"))
```

```
## [1] Male   Female
## Levels: Female Male
```

2.0.1.1 Categorical Variables and Regression

When performing a multiple linear regression with categorical data, the cqategorical data will be treated as a boolean `1/0` variable, basically choosing between different categorical variables is choosing a different intercept for the regression (i.e. adding a constant value)

> under the hood 1 corresponds to True and 0 corresponds to false

2.0.1.1.1 Not as accurate

This is not as accurate as using a linear regression seperately on the data within that category, so, unless there is a good reason i.e. the different categories would have a trend with the same rate but a different intercept, e.g. the ambient temperature of a location would have different mean values for each month:

$$\text{Response} = -74 = 3.109 \cdot X_{\text{wind}} - 1.875 \cdot X_{\text{temp}} + 14.76 \cdot (\text{Jan}) + 8.75 \cdot (\text{Feb}) + 4.197 \cdot (\text{Mar}))$$

in this case Jan, Feb etc. would be a `True`/`False` ($\equiv$ `1`/`0`) value indicating whether or not to include that constant in the equation (because it is or isn't that date) so if in this example Sydney is an everage of 23 degrees, maybe January is on average 14 degrees hotter and the coefficient for July might be -14 because in July it is 14 degrees cooler.

2.1 Supervised vs Unsupervised Problems

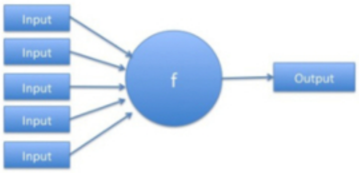
Machine Learning problems are often split into two categories, supervised and unsupervised.

Supervised Learning involves data where each observational unit has one special variable, (e.g. survive/perish, amount spent).

Unsupervised Learning Is about pattern discovery, there is no clear special variable (e.g. trying to detect unusual spending or grouping spenders into seperate groups).

2.1.1 Supervised Learning

Supervised Learning has a response variable (also known as outcome or in calculus as an independent variable, Y) and the idea is to predicting the relationship between the output and several inputs:



knitr::include_graphics(rep("images/knit-logo.png", 3))

In a classification problem Y takes the value of a discrete variable (e.g. survived/perished), In a regression problem Y takes a $\mathbb{R}$ quantitative value (e.g. $72.45 or 55 kg)

In Mathematical terms, if y was the output and $x_1, x_2, x_3 \dots$ the input, the model would be the expected value of y given the inputs, denoted $E(y)$, defined by some function f:

$$E(y) = f(x_1, x_2, x_3, \dots)$$

The observed output is expected to vary in value owing to errors in measurement (so for example even though we have very strong mathematical evidence for a relationship, e.g. $S = \frac{1}{2}at^2$) we would expect the observed values to contain error.

```
n <- 100
x <- 1:n
y <- 0.5*9.81*x^2 + rnorm(n, mean = 0, sd = 2000)

##layout(matrix(1:2, ncol =2))

#Using baseplot
```

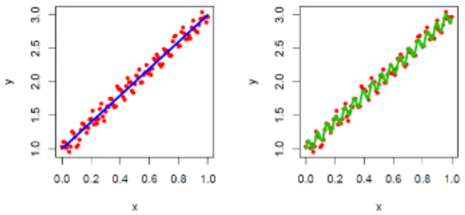
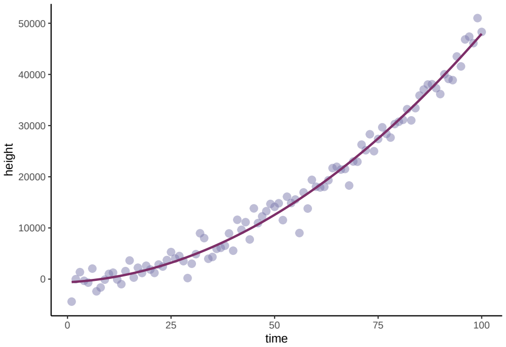


Figure 2.1: Comparison of more complex and simpler models

```
  #model <- lm(y ~ poly(x, degree = 2))
  #plot(y~x)
  #Expy <- predict(object = model)
  #lines(Expy)

#Using GGplot2
egdata <- data.frame(height = y, time = x) #create dataframe
ggplot(data = egdata, aes(x = time, y = height)) + # call ggplot2
  geom_point(col = "#7f7caf", size = 3, alpha = 0.5) + #plot the points for the data
  stat_smooth(col = "#7d2e68", method = 'lm', formula = y ~ poly(x, 2, raw = TRUE), se = FALSE) + # dra
  theme_classic() # change the theme
```



```
  # use https://coolors.co/app for colour themes
```

2.1.1.1 Types of Errors (Stochastic Trends)

- Random Error
  - unforseeable fluctuations in Data
- Systemic Error
  - Shortcomings of the capacity to measure accurately
    - e.g. Measuring using a ruler that is ±1 mm

2.1.1.2 Choosing the right Model

It’s also necessary to choose the right type of model, for example below the function chosen on the left is a simple linear regression, but maybe it’s appropriate to assume that there is a seasonal or cyclical trend (a cyclical trend being less predictable like a recession and a seasonal being more predictable like seasons).

2.1.1.3 Bias and Variance

The more complex a function, the more variance in fitting that function (because there is more uncertainty around the fitted parameters). However fitting a simpler function can introduce more bias because there may be a more fundamental difference between the predicted and actual values.

2.1.1.4 Model Evaluation

Prediction accuracy is estimated from the same sample that was used to fit the function, two strategies are used to offset the bias that this would introduce:

- Splitting the Data
  - Use **training data** to create the model
  - **validation data** to validate the model accuracy
  - **testing data** to measure the accuracy of the model predictions
- Cross Validation
  - Cross Validation only tests the modelling process while splitting the data evaluates the final model

2.1.1.5 Classification and Regression

2.1.1.5.1 Regression

When the output is a numeric variable, supervised learning can be referred to as regression, some examples of regression are:

- Linear Regression (Simple or Multiple)
- Generalised Linear Models (`glm`)
- Neural Networks
  - Neural Networks is an example of a non-paramentric method, the above two rely on statistical assumptions

2.1.1.5.2 Classification

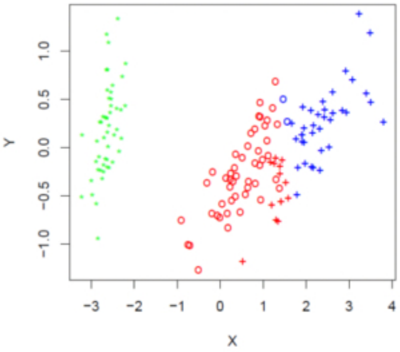
When the output is a categorical variable supervised learning is known as classification

- K-Nearest Neighbours
- Generalised Linear Models (Logistic Regression)

2.1.2 Unsupervised Problems

When there is no output variable the problem is usually one of **unsupervised Learning** a common example is clustering data, for example:

Clustering the Iris data



In this example the goal would be to classify the data according to the colours.

In this unit we will cover:

- Supervised learning:
  - Linear Models
  - Classification (KNN and Discrimination)
  - Classification and Regression Trees
- Unsupervised Learning:
  - DimensionReduction: Principal Component Analysis
  - Clustering: K Means and Hierarchical
- Unstrucutred Data:
  - Text Mining
- Resampling
- Visualisation

Chapter 3

(Wk 2) Introduction to Data Science

Material of Tue 12 2019, week 2 test

3.1 Heading 1

3.1.1 Sub Heading 1

3.2 Heading 2

3.2.1 Sub Heading 1

3.3 Heading 3

3.3.1 Sub Heading 1

Chapter 4

(Wk 3) Introduction to Data Science

Material of Tue 19 March2019, week 3 testlkj

4.1 Heading 1

4.1.1 Sub Heading 1

4.2 Heading 2

4.2.1 Sub Heading 1

4.3 Heading 3

4.3.1 Sub Heading 1

Chapter 5

(Wk 4) Introduction to Data Science

Material of Tue 26 March 2019, week 4

5.1 Heading 1

5.1.1 Sub Heading 1

5.2 Heading 2

5.2.1 Sub Heading 1

5.3 Heading 3

5.3.1 Sub Heading 1

Chapter 6

(Wk 5) Introduction to Data Science

Material of Tue 2 April 2019, week 5

6.1 Heading 1

6.1.1 Sub Heading 1

6.2 Heading 2

6.2.1 Sub Heading 1

6.3 Heading 3

6.3.1 Sub Heading 1

Chapter 7

(Wk n) Introduction to Data Science

Material of Tue 9 April 2019, week 6

7.1 Heading 1

7.1.1 Sub Heading 1

7.2 Heading 2

7.2.1 Sub Heading 1

7.3 Heading 3

7.3.1 Sub Heading 1

Chapter 8

(Wk n) Introduction to Data Science

Material of Tue 16 April 2019, week 7

8.1 Heading 1

8.1.1 Sub Heading 1

8.2 Heading 2

8.2.1 Sub Heading 1

8.3 Heading 3

8.3.1 Sub Heading 1

Chapter 9

(Wk 8) Introduction to Data Science

Material of Tue 23 April 2019, week 8

9.1 Heading 1

9.1.1 Sub Heading 1

9.2 Heading 2

9.2.1 Sub Heading 1

9.3 Heading 3

9.3.1 Sub Heading 1

Chapter 10

(Wk 10) Introduction to Data Science

Material of Tue 6 May 2019, week 10

10.1 Heading 1

10.1.1 Sub Heading 1

10.2 Heading 2

10.2.1 Sub Heading 1

10.3 Heading 3

10.3.1 Sub Heading 1

Chapter 11

(Wk 11) Introduction to Data Science

Material of Tue 13 May 2019, week 11

11.1 Heading 1

11.1.1 Sub Heading 1

11.2 Heading 2

11.2.1 Sub Heading 1

11.3 Heading 3

11.3.1 Sub Heading 1

Chapter 12

(Wk 12) Introduction to Data Science

Material of Tue 20 May 2019, week 12

12.1 Heading 1

12.1.1 Sub Heading 1

12.2 Heading 2

12.2.1 Sub Heading 1

12.3 Heading 3

12.3.1 Sub Heading 1

Chapter 13

(Wk 13) Introduction to Data Science

Material of Tue 27 May 2019, week 13

13.1 Heading 1

13.1.1 Sub Heading 1

13.2 Heading 2

13.2.1 Sub Heading 1

13.3 Heading 3

13.3.1 Sub Heading 1