CORK INSTITUTE OF TECHNOLOGY
INSTITIÚID TEICNEOLAÍOCHTA CHORCAÍ

COMP9033 – Data Analytics: Research Project II

Dr. Donagh Horgan

February 22, 2017

1 Outline

Linear regression is a useful tool for understanding complex multivariate data. In this project, you will use a data set consisting of the physical measurements of several thousand Abalone specimens to build a predictive model that can be used to accurately predict the age of an individual. By modelling the data accurately, you will be able to reliably predict the age of new specimens, given their physical measurements. Further details of the data are given in Section 2.

In order to complete this project, you are required to submit a report detailing your treatment, modelling and assessment of the data, explaining and critiquing your methodology, and outlining competing solutions. Further details of the report requirements are given in Section 3. The marking scheme is given in Section 4.

It should be noted that certain elements of this project **require** you to conduct **independent research** outside of the material covered in class. Where appropriate, you should identify reputable, peer-reviewed sources of information to use. External sources of information **must** be referenced. Guidelines for reference styles are given in Section 3.

Your project report must be submitted by **17:00 Irish Standard Time (IST), May 12, 2017**. Please note that, as college policy **strictly prohibits plagiarism**, all reports **must** be processed using Turnitin after submission. Further details on submission requirements are given in Section 5.

2 Data

You are given a single data file: `abalone.csv`. The data in the file describes the physical measurements, gender and age of 4177 Abalone specimens. The physical measurements include the length, diameter and height of the specimens, as well as several measures of weight. The age is not given directly, but can be calculated from the number of rings as

$$\text{Age} = \text{Rings} + 1.5. \tag{1}$$

A breakdown of the attributes is given in Table 1.

3 Requirements

In order to complete this project, you **must** write a structured report in the style of an academic paper. The report should be approximately 5000 words long (±10%), excluding references and captions. The structure of the report should be as follows:

Abstract: Summarise the content of your report, outlining the problem you have solved, the tools you have used and the main conclusions you have reached. *(200 words max.)*

Name	Units	Description
Sex	-	M (male), F (female), and I (infant)
Length	mm	Longest shell measurement
Diameter	mm	Perpendicular to length
Height	mm	With meat in shell
Whole weight	grams	Whole abalone
Shucked weight	grams	Weight of meat
Viscera weight	grams	Gut weight (after bleeding)
Shell weight	grams	After being dried
Rings	-	+1.5 gives the age in years

Table 1: Attribute names, units and descriptions.

Introduction: Describe the operation of linear ridge regression; list at least **three** advantages and **three** disadvantages of linear regression algorithms over other kinds of regression algorithms; explain whether the presence of outlying data affects the operation of ridge regression and, if so, how; find **three** example uses of linear regression in academia and/or industry, with appropriate references, and give a brief explanation of **each**.

Exploratory data analysis: Explain the importance of exploratory data analysis. Next, for the data problems a) erroneous data, b) outlying data and c) missing data:

- Explain **two** causes of this type of data problem.
- Outline **two** strategies for dealing with this type of data problem.
- Explain **one** advantage and **one** disadvantage of **each** strategy.

Finally, complete the following tasks:

1. Determine whether any records in the data set contain missing data and, if so, remove them. List any records you remove in your report.
2. Based on the information provided in Section 2, determine if any of the records in the data set contain erroneous values and, if so, then remove these records from the data set. List any records you remove in your report.
3. After removing any records with missing or erroneous data, prepare a scatter plot matrix of the data and a short report on each attribute. Each report should include:
   (a) An analysis of the attribute values and data type.
   (b) A histogram and boxplot of the attribute values.
   (c) Some brief comments summarising your impression of the attribute data and plots (*e.g.* are there dependencies, outliers, *etc.*).

   Make sure to choose appropriate visual cues when creating your charts. The clarity of the presentation style is an important factor: you should use visual elements (*e.g.* axis labels, chart titles, captions and annotations) judiciously.

Preprocessing and modelling: Complete the following tasks:

1. Build a ridge regression model to predict the number of rings of the Abalone specimens. Select the parameters of your model using 10-fold cross validation and detail the optimised parameter values in your report. You may need to preprocess your data before constructing your model. Detail any preprocessing steps you choose to take and explain your reasoning for doing so.
2. Validate your final model using hold-out cross validation and detail the error measurements in your report. Make sure to explain your reasoning for choosing a particular train-test split ratio.
3. List the coefficients of your final model and the features they relate to. Which feature is the most important? Which is the least?

Your report should also include some additional research beyond these requirements, *e.g.* you could quantify the improvement that additional or alternative preprocessing steps make and/or compare your model to a model generated using a different regression algorithm.

Conclusion: Evaluate the success of your approach and identify possible sources of error and/or improvements that could be made.

You may freely reference **but not copy** material from external sources. If you reference material from an external source, you **must** cite it in a consistent and accepted style (*e.g.* IEEE, Chicago, MLA). Wikipedia and personal blogs are **not** acceptable sources for citation. Instead, you should try to identify reputable, peer-reviewed sources of information.

4 Marking scheme

The project report will be graded as follows:

- Completion of the core requirements, as given in Section 3: **60%**.
- Additional preprocessing and/or modelling work beyond these requirements: **25%**.
- General presentation of the report, including layout, diagrams and references: **15%**.

Overall, this project is worth **70%** of the marks for this module. For more information, see bit.ly/2gxFuD2.

5 Submission

The deadline for submission of the project report is **17:00 IST, May 12, 2017**. When you submit your report, you **must** ensure that:

1. Your report is in Portable Document Format (PDF), *i.e.* `.pdf`.
2. You submit a Jupyter Notebook containing **all** of the code that you used to complete your project.
3. Your filenames follow the naming convention "`<NAME> <STUDENT_NUMBER>.<FORMAT>`", *e.g.* "`Donagh Horgan R00012345.pdf`".
4. You submit your report and notebook via Blackboard. You can find a link to the submission tool under *Assessments*.

Standard **penalties** apply for late submission. For more information, see the *Regulations for Modules and Programmes (Marks and Standards)* at bit.ly/2gJe0vq.