IS71104A Statistics and Statistical Data Mining – Coursework

Elliot Walker [SN: 3368 6408]1

1Goldsmiths, University of London

8 Lewisham Way, London SE14 6NW

**Abstract.**

# 1 Overview

Linear regression and Logistic regression are statistical modelling techniques which yield expressions that provide predictions for environmental response variables. The former is utilized for continuous, numerical data, whereas the latter is a form of probabilistic classification (e.g., predictions are provided as probabilities, then rounded to either 0% or 100%).

Both techniques are utilized in this report and applied to relevant data sets over a series of three analytical tasks:

* **Task 1 – Multiple Linear Regression – Auto.csv**
* **Task 2 – Multiple Linear Regression – Carseats.csv**
* **Task 3 – Logistic Regression & KNN – Weekly.csv**

The code, written in the R Programming Language, along with the findings of each conducted task and associated visualizations of data are provided throughout the report.

# 2 Theory

## 2.1 Regression

Regression applies adjustable coefficients to environmental predictors (inputs) defined in the set…

to yield a prediction . This prediction is generated with the use of test data (usually derived from the original dataset) representing the set of predictors , which are in turn adjusted courtesy of their corresponding coefficients. The values of these coefficients are constant and derived from the modelling process using a training subset.

A univariate (single input) linear regression model takes the following form…

where is the model’s irreducible error (e.g., environmental noise causing skews in prediction accuracy), and is the model’s y-intercept for . The remaining coefficients of the form are gradients (rates of change) for the predictors.

The multivariate counterpart for predictors is given as…

which simplifies to…

Accounting for dummy variables (cross-field numeric representations of categorical data), the model extends to…

with as thedummy variable instance, for dummy variables.

## 2.2 Classification

Classification follows a probabilistic paradigm. It is prudent to determine the probability of a response belonging to a specific category (or discrete value), as opposed to outright stating which one it belongs to.

Models are referred to as classifiers, which, using these calculated probabilities, sort response variables into predefined categories/classes.

Logistic regression is one such form of classification model, and always yields a value between 0 and 1 (representing 0% and 100% probabilities). Below is the univariate form…

and the multivariate form for classes…

The exponent for Euler’s number used in this classifier is identical to a linear regression model, which can be made the argument of the equation by taking the logit…

## 2.3 K-Nearest Neighbors

Among the other classification techniques used in this report is KNN, k-nearest neighbors, which is an unsupervised learning algorithm that classifies observations based on their proximity to others in a dimensional plane.

The unsupervised aspect of this classification arises from the k-nearest neighbors to a given uncategorized (usually test) observation, whose classes are already known. The test observation is assigned the most common class among its neighbors, where the proximity may be defined as either Euclidean (hypotenuse) or Manhattan distance (vertical and horizontal components only), making KNN nonparametric.

The n-dimensional Euclidean distance between two points and is given as follows…

and their n-dimensional Manhattan distance is given by…

The go to distance measurement is usually the Euclidean distance.

## 2.4 Hypothesis Testing

Under what conditions does a model indicate relevance and accuracy towards its represented environment? Hypothesis testing and its constituent tools grant this insight.

Regression coefficients may be of focus where the goal is to determine if the null hypothesis applies, such that the corresponding predictor does not affect the response variable , and is cancelled out. For example, in the case of univariate linear regression…

This value corresponds to the statistical significance of in the model, which is given by its p-value – the lower it is, the more likely the predictor is to be significant to the model predictions. The p-value is the probability of observing an extreme (very big or very small) value for the test statistic of the model, provided that is true.

In this report’s toolkit RStudio, this value is stated under each model summary as , where is student t-test statistic given by the following equation for observations in a sample…

is the sample mean of observations, is the population mean of all observations in , and is the standard deviation of the field. Respectively, these are all given by…

In hypothesis testing aimed towards the gradient coefficients of the model, the test statistic numerator takes the following form…

Where applies to the model (null hypothesis is considered true for ), this statistic then exhibits a t-distribution with degrees of freedom. If the null hypothesis is shown to be false, then a solid relationship between predictors and responses may be established based on the intuition and judgement of the scientist.

Rudimentarily put, disproving a null hypothesis requires infinitesimal p-values, and values for which stray far from a value of 0.

Associated with the hypothesis test is the confidence interval, which is the range of prediction values a model can be expected to yield upon being tested. It is given by…

Where is the confidence value, predefined by the confidence interval’s percentage (e.g., corresponds to ).

Complimentary to the p-value is the -value, the significance level and probability of conducting a type I error (rejecting, perhaps prematurely, when it is in fact true)…

## 2.5 Model Accuracy

Model accuracy plays a significant role in hypothesis testing with its focus on statistics which measure the errors (or in the case of classifiers, confusion), exhibited by the model.

The irreducible error cannot be mitigated, as its origin is that of stochastic, disruptive phenomena. Naturally, all models are imperfect and cannot perfectly capture the nature of the environments they designed for. In the hands of the scientist, though, is the reducible error, given by the difference between an actual response and a prediction of said response given by the model…

This error can be further applied to calculate statistics such as the residual sum of squares (RSS)…

The -statistic measures the goodness of fit for a given model by explaining how much responses vary with the given predictors, and utilizes the RSS with the total sum of squares (TSS)…

where RSS and TSS may be interpreted as the modelled variation of data currently being investigated, and TSS is the overall variation. A high -statistic implies that the predicted response variables are affected significantly by the predictors .

For multivariate regression models it is necessary to adjust the -statistic to help account for possible overfitting, wherein the model is hyperfocused on generating correct predictions only when accepting its training data, and not test data. Adjusted is given as follows for observations and predictors…

Model accuracy inherently relies on the training data and the predictors used. The training data is utilized to help determine an estimate for each predictor’s regression coefficient .

In linear regression this involves selecting values for which minimize the value of RSS…

whereas even though this same method is valid, for logistic regression models the maximum likelihood approach is preferred…

with values for the regression coefficients necessarily maximizing the likelihood .

# 3 Tasks

## 3.1 Task 1 of 3

The **Auto.csv** dataset contains statistics recorded over a period of 12 years between 1970 and 1982, denominated with one response and nine inputs:

|  |  |
| --- | --- |
| **Field Name** | **Outline** |
| mpg | Miles per gallon of fuel consumed by the automobile’s engine. Response variable for the model. |
| cylinders | The number of cylinders possessed by the automobile’s engine. |
| displacement | Engine displacement. |
| horsepower | Horsepower (hp) produced by the engine. |
| weight | Weight (mass under gravity) of the engine. |
| acceleration | How much the automobile progresses between two velocities. |
| year | Year which the automobile was manufactured. |
| origin | Region from which the automobile originates from. Encoded as for America (USA), for Europe, and for Japan. |
| name | The brand and model of automobile. |

The dataset contains 397 observations, one of which possesses a duplicate (in name, year of manufacture, and cylinder count). As part of the data preprocessing for this task, these duplicates were filtered from the original dataset, the arithmetic means of their fields calculated, then the resulting row appended to the original data set.

## 3.2 Task 2 of 3

## 3.3 Task 3 of 3

# 4 Discussion

*Don’t stray to far, keep it to the requirements outlined in the brief 😉*

# 5 Conclusion

# References

# Appendix A – Code for Task 1

# Appendix B – Code for Task 2

# Appendix C – Code for Task 3