Statistical Learning Assignment

MIT 801

Connor McDonald
u16040725

Introduction To Machine And Statistical Learning

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Department of Computer Science
University of Pretoria
21 May 2021

1 Linear Regression

Linear regression is arguably one of the most well-known and most well-understood algorithms in the field of machine and statistical learning. Essentially, linear regression is a linear model which takes one or more input variables X_i and constructs a linear combination of these variables with the regression coefficients (β) and the error (ϵ) to model an output variable Y_i with the formula shown below.

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \tag{1}$$

Generally speaking, linear regression is classified into two separate types, known as simple and multiple linear regression. Simple linear regression only takes one explanatory variable whilst multiple linear regression can take infinitely many variables. However, both of these methods use the same underlying processes of fitting a line in space that minimizes the sum of squares of the deviation of each point from the line. The distance between any given point and the fitted line is known as a residual and will be referred to as such for the remainder of this report. By plotting the line that minimizes the sum of the squared residuals we model the single dependent variable Y.

Whilst linear regression can model a wide variety of data, certain assumptions need to be in place to guarantee the effectiveness of the model, this includes the following:

- *Linearity* - there must be some degree of linear relationship between the explanatory variables X_i and the dependent variable Y.
- *Normality* - the dependent variable must be normally distributed for any fixed X_i.
- *Constant Variance* - also known as "Homoscedasticity" assumes the variance of the residuals should be constant for any given X_i
- *Independence* - all observations are independent of one another.

Correlation is often used in conjunction with linear regression to test the strength of the linear relationship between variables. Correlation ranges between -1 and +1 where 0 represents no correlation at all, -1 represents perfect negative correlation and +1 represents perfect positive correlation. However, where linear regression assumes X to be a fixed variable and Y to be a randomly distributed variable, correlation assumes both X and Y to be randomly distributed variables.

The reason in which correlation and linear regression are often used together is due to the fact that data with a strong linear relationship will be modelled much more effectively than data with a weak or non-linear relationship. The correlation is often computed before the linear regression analysis in order to test the first assumption shown above.

Since the absolute value of the correlation coefficient (r) determines the strength of the linear relationship, it is often squared to create the R-squared (R^2) metric which measures the proportion of variance in the dependent variable that can be attributed to the independent variable(s). R-squared always ranges between 0 and 1, where 0 indicates that the model explains none of the variability in the dependent variable around its mean, and 1, which indicates that the model explains all of the variability of the dependent variable around its mean. It is important to note that linear regression models only approximate the dependent variable, and therefore it is very rare for the model to predict the exact value of y unless the model has R-squared value of 1, which is arguably impossible to achieve in real-world applications. Due to this fact, if the accuracy of linear regression model was determined based on number of correct predictions divided by total number of predictions, most models would have extremely low accuracy values. Hence R-squared must be used in place of a traditional accuracy calculation as it is a measure of goodness of fit, or how closely the data lies to the line of best fit. Essentially, a higher R-squared value implies that the observations lie closer to the mean and thus can be approximated closer to their real values (more accurately) than a data set with a lower R-squared value.

When dealing with multiple linear regression we are often faced with numerous parameters, some of which may not be contributing any significant information to the model. In this case we can improve our models performance by identifying these insignificant parameters and excluding them from the model all together. When constructing a regression model, a regression coefficient is estimated for each parameter. Looking at equation 1, we only have one parameter (X_1) and one regression coefficient (β_1), β_0 is the y-intercept of the line. However, in multiple linear regression we can have as many regression coefficients as there are parameters, and since regression is a form of inferential statistics, we can calculate a p-value for each regression coefficient to test its significance. We start by assuming a null hypothesis (H_0): a certain parameter has no correlation to the the dependent variable. Next, a significance level is chosen (the most commonly used value is 0.05), if our p-value is lower than the significance level this infers that there was enough evidence in the sample data to conclude that the parameter is not insignificant and that we should include it in our model.

Validation is a crucial step in any model building process, as the models results need to be somewhat reproducible in order for the model to provide any use. The most common method of validation is splitting the data set into a training set and a testing set. Typically the training set is larger than the testing set as having more data for a model to "learn" with allows for a more generalized model. The goal of the test set is to see if the model can accurately estimate the dependent variable on new data which has not been used to construct the model. This gives a realistic perspective of how the model will perform when applied in a real world scenario. However, this will not work if the number of outliers is disproportionate between the training and testing data. To combat this problem we can use k-fold cross validation which is a resampling procedure where the data are split into k groups where the training and testing sets are selected from different groups each time.

Building on this concept, suppose we create a model that fits the training dataset perfectly with an R-squared value of 1, this model is said to have very low bias as it has a sum of squares equal to 0, however if we had to validate the model on a test set we would likely encounter a sum of squares value much greater than zero and thus a large amount of error, this is because the model has been over-fit and has high variance. This error can be reduced to some extent by finding a balance between bias and variance in the model. The amount of error that can be predicted by X_i is known as reducible error, and since it can be predicted by X_i we are able to reduce it by fine tuning our model with various statistical techniques. However, there is also an error component in the regression function shown in 1 which is denoted by ϵ. This component cannot be predicted by X_i and hence cannot be reduced, this is known as irreducible error. The reason this component is irreducible is because it is comprised of either unmeasured or immeasurable quantities and therefore cannot be used for prediction, a good example of this is natural variation.

Interpreting the results of a linear regression model can be an iterative process where the methods and techniques mentioned above are used to either understand, improve or validate the model. As stated previously, it is important to remember that linear regression is only an estimation of the dependent variable, and therefore it would be unwise to judge the models performance on whether it made a correct prediction or not. A better approach to evaluating and interpreting results would be to analyse the regression coefficients in order to get an idea of whether they are positively or negatively correlated to their respective parameters as well as their weighted influence on the model prediction (shown by their magnitude). Next would be to assess the R-squared value to get an idea of how well your model estimates your data. Lastly, eliminating your insignificant parameters based on their respective p-values may further improve the model by minimizing the reducible error. As stated before this is an iterative process and may need to be done numerous times before arriving at an optimal model.

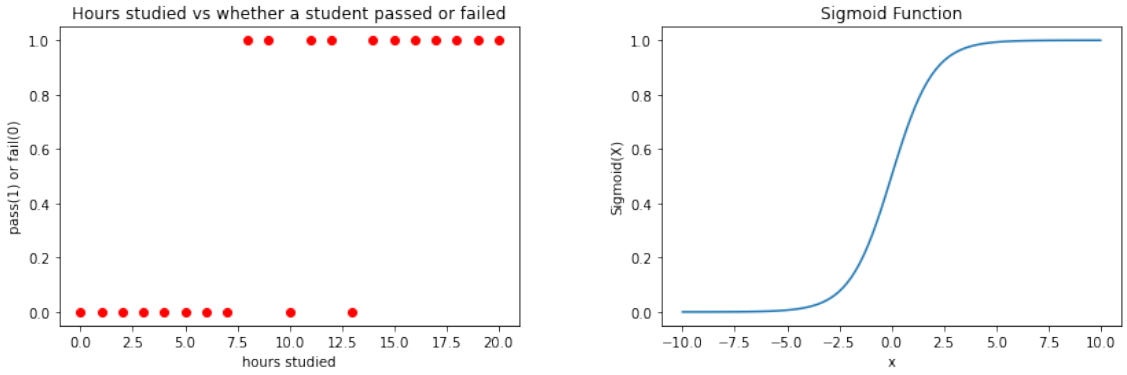
Linear regression has its fair share of advantages and disadvantages. For example, it can be a very powerful algorithm regardless of dataset size, however it requires all the assumptions detailed on page 1 to be met. Linear regression, unlike many other algorithms, also provides insight into the relevance of parameters and thus can be used as an intermediate model in the modeling process which is used solely to filter out insignificant parameters. Lastly, linear regression is prone to underfitting (due to its high bias) and can be sensitive to outliers, therefore it is of utmost importance to standardize your data and remove any known or obvious outliers if you want to get the most our of your model.

2 Logistic Regression

Logistic regression is an increasingly popular statistical technique used to model the probability of a dichotomous dependent variable such as (yes/no), (0/1), (pass/fail) etc. The underlying mathematical concept used in logistic regression is that of the logit function which is the natural logarithm of the odds of each outcome, where p is a probability. This function is shown below in equation 2.

$$\text{logit(p)} = \ln(\frac{p}{1-p}) = \beta_0 + \beta_1 X_1 +, ..., +\beta_n X_N \tag{2}$$

The logit function is used to model a categorical dependent variable on a continuous interval, which is why logistic regression returns the probability of an observation belonging to one specific class or another and not the class of the observation itself. Logistic regression is best suited to modeling relationships between a single categorical outcome and numerous categorical or continuous inputs. Logistic regression solves one of the major downfalls of linear regression, which is modelling a categorical outcome. Even if our explanatory variables fall on a continuous interval, we cannot fit a line of best fit through the data if it has a dichotomous dependent variable as seen in figure 1a. However, if we split the explanatory variables into groups and plot the mean of each group, an "S-shape" will form as we the number of groups tends to infinity over a the same initial interval. This shape is shown in figure 1b and is known as the Sigmoid function. The Sigmoid function appears to be linear in the middle section but curves off at the extremes of the function.



(a) (b)

Figure 1

This Sigmoidal shape would fit the data better than a straight line, however it is difficult to model as the extremes do not follow a linear trend and the the errors are neither normally distributed nor constant across the entire range of data. However, by transforming the dependent variable with the logit function we can address these problems whilst also creating a linear relationship between logit(Y) and X.

Many people wonder why logistic regression isn't called logistic classification since it is mainly used to predict categorical variables. To understand why this is the case we must first understand what classification is. Classification is the problem of identifying which set of categories an observation belongs to. Logistic regression does not do this, but rather it returns the **probability** that an observation belongs to a specific class. On its own, logistic regression is in fact regression, however by combining logistic regression with a decision rule such as: observations with a probability less than 0.5 belong to class A and observations with a probability greater than 0.5 belong to class B, we create a classification algorithm.

Since prediction values are so heavily rooted in the Sigmoid and logit functions, one might suspect that logistic regression cannot be used in cases where there are more than two output categories. This is not the case, for example, assume we want to predict if an observation belongs to one of three classes, A,B or C. To do this we create three logistic regression models. The first model predicts if an observation belongs to category *A or not*, the second predicts if the observation belongs to *B or not* and the last model predicts if the observation belongs to *C or not*. All three models are then used in conjunction to predict the category of an observation and the one with the highest probability will be used as the outcome of the observation.

Logistic regression can be advantageous over other algorithms as it trains efficiently and is easy to implement and interpret. It is arguably simpler than linear regression as it needs not make any assumptions about the distributions of classes in the feature space and can make predictions on unseen data incredibly quickly.

However, one must be care not to overfit their model when classifying large numbers of categories especially when using subsets of data that may contain more features than observations. Another drawback of logistic regression is that, much like linear regression, there must be an assumption of linearity between the dependent and independent variables.

3 Linear Discriminant Analysis

Linear discriminant analysis is similar to principal component analysis but it focuses on maximising the separability of the classes. Linear discriminant analysis was originally created by Ronald A. Fisher in 1936 and was designed to be applied to a two class problem. For example consider a sample dataset with two dimensions and two outcome classes shown below:

hours slept	hours studied	class
6	10	pass
8	8	pass
7	9	pass
6	4	fail
7	8	fail

We can reduce this dataset to one dimension by using linear discriminant analysis which will use the information from both dimensions to create a new axis on which the data will be projected. This will be done in such a way that maximises the separability of the two classes. This new axis is created by simultaneously considering the two criteria shown below:

- maximising the distance between the means of each class.
- minimising the variation (also known as scatter) of each class.

This concept can be applied to any number of dimensions provided that the class is dichotomous, and it wasn't until 1948 that C.R. Rao generalized the model to work for datasets with more than two classes.

When dealing with more than two classes, some changes need to be made to the approach, but the underlying concept remains the same. Suppose we had three dimensions, firstly we can no longer use the distances between the means of each class. We have to identify a central point and then maximise the distance between that point and means of each class. Furthermore, whilst with a two-dimensional dataset we identified a **single** new axis. We will now identify a new **set of axes** as we are reducing the dimensions from three to two and thus will be projecting our data onto a plane. We can create a maximum of n axes for an n-dimensional dataset, but typically linear discriminant analysis is used to reduce large dimensional datasets to a number of dimensions that can be visualised which usually means two or three dimensions.

Linear discriminant analysis uses Bayes' theorem to estimate the probabilities of a new set of inputs belonging to each class the class that achieves the highest probability is then the output class of the model. Whilst logistic regression and linear regression are were both designed for use with dichotomous classes, linear discriminant analysis often outperforms logistic regression when dealing with well separated classes or small datasets as logistic regression is known to become unstable in such cases.