Practical Machine Learning with R

Lecture 5 : Regression and Regression Trees

Objectives:

- ❖ Comprehend Regression in context of predictive scope of machine learning algorithms.
- ❖ Perceive Ordinary Linear Regression and the estimation of regression coefficients.
- ❖ Adaptation of classification Trees to depict regression trees for the purpose of quantification of the prediction.

Regression Methods for Forecasting:

- These methods help quantify the numeric relationship between the attributes/features of interest.
- Predictions are inherently more insightful and substantive for decisions that depend on concrete numerical values like profit or investments in terms of dollars.
- Regression entails elucidating relationships between one or multiple independent variable and a dependent variable. The relationship therefore can depicted by a technique called Simple Regression or Multiple Regression.
- The inherent premise/assumption for Simple Regression is that the relationship between the Independent and Dependent variable is linear ie follows a straight line pattern.
- Simple Linear Regression is a scientific paradigm introduced by the genetic scientist Sir Francis Galton. He empirically proved that son’s height regresses to the mean of the father’s and mother’s height.
- Regression technique, although a simple technique can model complex relationships that facilitate ascertaining the strength, direction as well as numerical quantification of the response variable including futuristically achieving extrapolation.
- This methodology can be used across a spectrum of domain areas including but not limited to insurance claims, crime rates, election results, profit projection etc.
- Regression encompasses a family of algorithms addressing different combination of Categorical and Quantitative data.
- In context of Simple or Multiple Regression the dependent (response variable) variables are quantitative and the independent variable can be categorical or quantitative.
- If the dependent variable is a binary categorical variable and independent variable is nominal, ordinal, interval or ratio then for classification purposes the technique titled Logistic Regression is used whereas multinomial Logistic regression is used if the classification entails a categorical dependent response variable and multiple independent variable is is nominal, ordinal, interval or ratio.

Simple Linear Regression For representing the regression equation between a dependent and independent variable the following equation is used:

Population:

$y = \alpha + \beta x + \varepsilon$ Epsilon is the error term

Alpha α is the y intercept and Beta β is slope of the line

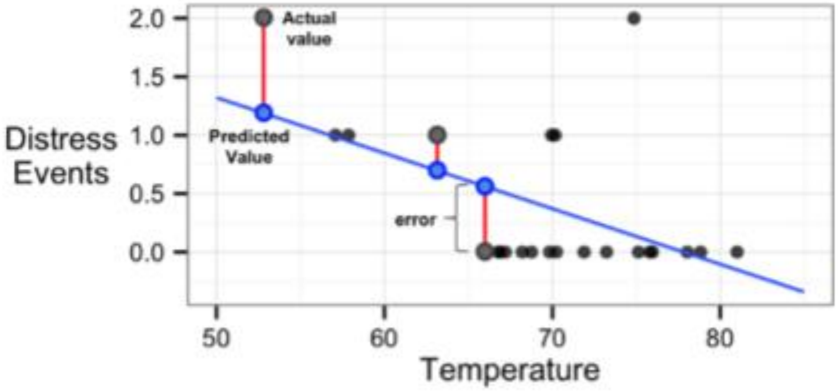
Sample:

$\hat{y} = a + bx$ where this is the Line of Best Fit which estimates the real response variable y.

Where a and b are the estimates of α and β

The assumptions for Regression to work include that for every x , the y values are normally distributed and have equal variances.

Ordinary Least Squares Estimation(OLS) is a process of ascertaining the line of best fit that minimizes cumulative residual error.



Residual error $y_i - \hat{y}_i = e_i$

- The line of best fit is found by computing a and b by minimizing the error.

$$\sum (y_i - \hat{y}_i)^2 = \sum {e_i}^2$$

This can also be interpreted as minimizing the mean squared error in context to testing data

$$\boldsymbol{MSE} = \sqrt{\frac{(\boldsymbol{y}_i - \widehat{\boldsymbol{y}})^2}{\boldsymbol{N}}}$$

This is minimized by using calculus and the value of a and b are derived as follows:

$$a = \bar{y} - b\bar{x} \quad \text{and} \quad b = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

The value of b can be transformed by the following formula of variance and covariance:

$$Var(x) = \frac{\sum(x_i - \bar{x})^2}{n}$$

And

$$Cov(x,y) = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n}$$

$$b = \frac{Cov(x,y)}{Var(x)}$$

a and b estimates α and β

Correlation The correlation between two variables determines the strength and direction of a relationship between two variables. This quantified by a measure called Pearson's correlation coefficient named after a scientist called Karl Pearson. Pearson correlation ranges from – to 1 where 0 signifies no correlation.

Conventionally 0--.49 Weak .5---.69 Moderate .7---1.0 Strong

$\rho_{x,y}$ Correlation Coefficient

$$\rho_{x,y} = Corr(x,y) = \frac{Cov(x,y)}{\sqrt{Var(x)}\sqrt{Var(y)}} = \frac{Cov(x,y)}{\sigma_x \sigma_y}$$

Multiple Linear Regression If there are several independent /predictor variables for an outcome /response variable then Multiple Linear Regression is very useful.

This a good technique that can model any numeric data though the model formulation must be specified by the user.

This technique works intrinsically for numeric data and requires additional processing for categorical data.

The algorithm makes strong assumptions of the underlying data, does not handle missing data.

$$Y = \alpha + \beta X + \varepsilon$$

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \ldots\ldots\ldots\ldots \beta_i X_i + \varepsilon$$

This can also be represented as

$$Y = \beta_0 X_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \ldots\ldots\ldots\ldots \beta_i X_i + \varepsilon$$

Where α = β_0

Therefore

$$Y = \beta X + \varepsilon$$

β_0	β_1	β_2	β_3

Y

Xo	X_1	X_2	X_3

ε

Each row is an example , X_i s are features Y is the outcome and betas are the regression coefficients.

This can be depicted in vector notation as:

$$\boldsymbol{Y} = \boldsymbol{\beta X} + \boldsymbol{\varepsilon}$$

The model for the multiple regression equation is

$$\widehat{\boldsymbol{Y}} = \widehat{\boldsymbol{\beta}}\boldsymbol{X}$$

Where

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X^T X})^{-1}(\boldsymbol{X^T Y})$$

Regression Trees and Model Trees

If the classification tree is adjusted the tree can be used to make numeric predictions. There are two types of trees used for numeric predictions: Regression trees and Model trees.

Paradoxically Regression trees (CART package) use the average value of the examples that reach the leaf rather than linear regression techniques to make predictions.

Model trees are more powerful and these use Regression techniques. Specifically at each leaf node a multi Regression model is built from the examples reaching the nodes. These might be more accurate but sometimes could possible be very cumbersome if they are too many leaves.

https://towardsdatascience.com/introduction-to-model-trees-6e396259379a

Regression trees and model trees are less known as compared to Linear regression. Trees require significantly large data for training the algorithm.

The tree automatically chooses the features for the final prediction to create a multiple regression framework but this comes with the tradeoff of not being able to ascertain effect of individual features. Even though the regression tree might fit the data to a greater extent the trees are more challenging to interpret as compared to a conventional regression model.

Model Trees using the regression model are better for scenarios where there are multiple features and these showcase non linear relationships between features and outcomes . Trees also do not need to satisfy distribution constraints like normality for the outcome data which are the inherent assumptions for Regression models.

The data is partitioned using the Divide and Conquer strategy. Unlike Trees which uses Entropy to access the homogeneity, trees use a measure called Standard Deviation reduction.

https://www.saedsayad.com/decision_tree_reg.htm