CART

Prof. Eric A. Suess

Feburary 8, 2021

Introduction

Today we will discuss more details about **Regression Trees**. This is the **RT** part of CART.

Before discussing Regression Trees, we will discuss a little more about Linear Regression.

Prediction or Forecasting

The author uses the word forecasting and prediction interchangeably.

As a Statistician, I have always tried to be specific about which word to use. Prediction is what is done when using Regression, or other similar methods, to predict a mean value or future observation within the range of the data. Forecasting is what is done with time series data when future observations are forecasted beyond the range of the data.

Prediction or Forecasting

See wikipedia for further discussion.

- Prediction
- Forecasting

I would have titled the chapter Predicting Numeric Values - Regression Methods.

This is not that important in the big picture.

Linear Regression

The **simple linear regession model**:

$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$

where

$\epsilon_i \sim N(0, \sigma_\epsilon^2)$

Because of the distributional assumption on the **error terms** we can do Statistics. So we can **test the statistical significance** of the slope β_1 and **compute a confidence interval** for the slope β_1.

Linear Regression

Parameter estimates are usually represented by the parameter with a **hat**. So the estimate of the slope in the simple linear regression mode would be

$\hat{\beta}_1$

The author introduces the **a** and **b** as the estimates.

Linear Regression

The estimates are produced by minimizing the **Sum of Squares Error** for β_1 and β_0.

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

where

$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

Linear Regression

The estimates are

$$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = r \cdot \frac{s_y}{s_x}$$

where

$$s_y^2 = \frac{\sum(y_i - \bar{y})^2}{n-1}$$

and

$$s_y = \sqrt{s_y^2}$$

Linear Regression

and

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \cdot \bar{x}$

So the fitted model would be

$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

We could use this equation to draw a line on a scatterplot.

Final Questions

Question: What is the relationship between the correlation coefficient and the estimated slope coefficient?

Question: Can you define clearly what the correlation coefficient r measures?

Multiple Linear Regression

Multiple Linear Regression includes more predictor variables.

Matrix notation is usually used to write down the model.

$Y = X \cdot \beta + \epsilon$

where

- Y is an $n \times 1$ vector of the y_i values
- X is the $n \times (p + 1)$ design matrix
- β is an $(p + 1) \times 1$ vector of the β_i values
- ϵ is an $n \times 1$ vector of the ϵ_i values

Multiple Linear Regression

The estimates are

$$\hat{\beta} = (X^T \cdot X)^{-1} X^T Y$$

- $^{-1}$ is the inverse
- ^{T} is the transpose

Regression Trees using CART

Trees for Numeric Prediction.

Strengths:

- trees for numeric data
- automatic feature selection
- no model in advance
- may work better than traditional regression
- does not require knowledge of statistics to interpret the results

Regression Trees using CART

Weaknesses:

- not so commonly used
- requires large amount of training data
- difficult to interpret effect of the predictors/features
- may not be as easy to interpret as a traditional regression model

Regression Trees using CART

Partitioning is done using a **divide-and-conquer** strategy according to the feature that will result in the greatest increase in homogeneity in the outcome after a split is preformed.

So the measurement is on the response variable/target variable Y.

Common Splitting Criteria

Standard deviation reduction (SDR)

$SDR = sd(T) - \sum \frac{|T_i|}{|T|} \times sd(T_i)$

where

$sd(T)$ is the standard deviation of the Y values that are in the set T.

$|T|$ stands for the number of observations in the set T.

Today

Today we will look at the wine data example that tries to create a system to mimic expert ratings of wine.

Wine Spectator's 100-point Scale

White Wine

Red Wine

Measuring performance with MAE

Since we are not performing a Classification with Regression Trees, in this example, we cannot use a Confusion Matrix.

We will look at the **correlation** between the *test values* and the *predicted values*.

Another way to measure the error is to use the **Mean Absolute Error** (MAE)

$MAE = \frac{1}{n} \sum |\epsilon_i|$

or **Mean Squared Error** (MSE)

$MSE = \frac{1}{n} \sum \epsilon_i^2$