Allison Molitor

M.L. Learning Journal

Start Date: 01/05/2022

Most Recent Edit: 08/24/2023

Machine Learning Regression Models

Linear Regression Modeling

- Mathematical Representation
    - $Y = \beta0 + \beta1*X + \varepsilon$
        - Y = dependent variable
        - B0 = Y int when X=0
        - β1 = Slope coefficient (How much Y changes for 1 unit change in X)
        - X = independent variable
        - E = (Error) Variability not explained by model
- Real-world example
    - Required bandwidth and Network users
        - Required Bandwidth = β0 + β1* Network users + ε
            - Ordinary Least Squares determine β0 and β1
- Mathematical computation

Network users (X)	Required bandwidth in Mbps(Y)
1	3
2	5
3	7
4	9
5	11

Equation setup
$\Sigma Y - n\beta0 - \beta1\Sigma X = 0$

$\Sigma Y = 3 + 5 + 7 + 9 + 11 = 35$
$\Sigma X = 1 + 2 + 3 + 4 + 5 = 15$
n = 5 (sample size)

$35 - 5\beta0 - \beta115 = 0$
$5\beta0 + 15\beta1 = 35$

$\Sigma(XY) - \beta0\Sigma X - \beta1\Sigma(X\char`^2) = 0$

$\Sigma(XY) = = (1 * 3) + (2 * 5) + (3 * 7) + (4 * 9) + (5 * 11) = 115$
$\Sigma(X\char`^2) = = (1\char`^2) + (2\char`^2) + (3\char`^2) + (4\char`^2) + (5\char`^2) = 55$
$115 - \beta015 - \beta155 = 0$
$15\beta0 + 55\beta1 = 115$

Solve for β1 and β0 given
$5\beta0 + 15\beta1 = 35, 15\beta0 + 55\beta1 = 115$
B1 = 1
B0 = 4

$Y = \beta0 + \beta1*X + \varepsilon$

$Y = 4+1*X + e$

Meaning for every 1 unit increase in network users, approximately 4 more Mbps is required.

If a model was fitted to a dataset using the predict function, one could predict the required Mbps given the number of network users like so

(Predicted bandwidth requirement) = 4+1*(Network users) + e

RSS or Residual Sum of Squares can be used to determine a model's accuracy, it displays the overall discrepancy between predicted values and observed values

The lower the RSS, the better

The calculation for RSS in this model, first one must calculate Y_hat
$X = 1$
$4 + 1*1 = \underline{5}$
$X = 2$
$4 + 1*2 = \underline{6}$
$X = 3$
$4 + 1*3 = \underline{7}$
$X = 4$
$4 + 1*4 = \underline{8}$
$X = 5$

4 + 1*5 = 9

Then subtract the Y_hat values from the actual values in the table
X = 1
3 - 5 = -2
X = 2
5 - 6 = -1
X = 3
7 - 7 = 0
X = 4
9 - 8 = 1
X = 5
11 - 9 = 2

Then, square all values and sum them
(-2)^2 = 4
(-1)^2 = 1
0^2 = 0
1^2 = 1
2^2 = 4
RSS = 4 + 1 + 0 + 1 + 4
RSS = 10

Another measure to determine model accuracy is MSE or Mean Squared Error. This is calculated by dividing the RSS by the number of observations. For the scenario above, this would be

RSS/n = MSE
10/5 = MSE
MSE = 2

If one were to improve the model, one could use a multi-linear or non-linear regression model that considers multiple features like user demographics. One could confirm findings of a higher accuracy by comparing RSS values and MSE

- Primary Advantages
  - Simplicity
    - Straightforward
    - Simple mathematical approach
  - Easily Interpretable
    - Coefficients represent the variable relationship
  - Efficiency
    - Light computational load in relativity to more complex models
    - Quicker training on large datasets

    - Valuable insight on feature importance
        - Given coefficient values, impact significance can be determined
    - Works better on smaller sample sizes than more complex models
- Primary Disadvantages
    - Assumption of Linearity
        - If a relationship is potentially non-linear, the model will be inaccurate
    - Assumption of Independence
        - Non-independence between variables can lead to inaccuracies and potentially invalidate conclusions
    - Poor for complex relationships
        - Complex patterns can be hard to discover through a Linear Regression.
    - Susceptible to Overfitting or Underfitting
        - Underfitting: Not enough data to capture the relationship
        - Overfitting: too much data that causes noise/distortion to model accuracy
    - Very susceptible to outliers
        - Outliers can heavily distort coefficients
    - Potentially requires feature engineering
        - Fitting the data with the best features may need the implementation of other processes
- Common uses
    - Predictive Analysis
        - Forecasting for sales
        - Forecasting demand
        - Financial trend analysis
    - Relationship Analysis
        - Measures strength of correlation between two variables
    - Risk assessment
        - In insurance, it can predict expected claims given age, location, health, etc
    - Price evaluation
        - Given stats on competing products, one can determine the fair value of their own product
    - Performance evaluation
        - Can analyze correlations between performance and variables like hours of training
    - Quality Control
        - Helps find factors that are most tied to product quality
    - Economic analysis
        - Relationships between varying economic indicators