Department of Electronic & Telecommunication Engineering,
University of Moratuwa, Sri Lanka.

Assignment 01

Learning from data and related challenges and linear models for regression

De Silva A.N.T. 210097M

Submitted in partial fulfillment of the requirements for the module
EN 3150 Pattern Recognition

02 September 2024

1 Data Pre-processing

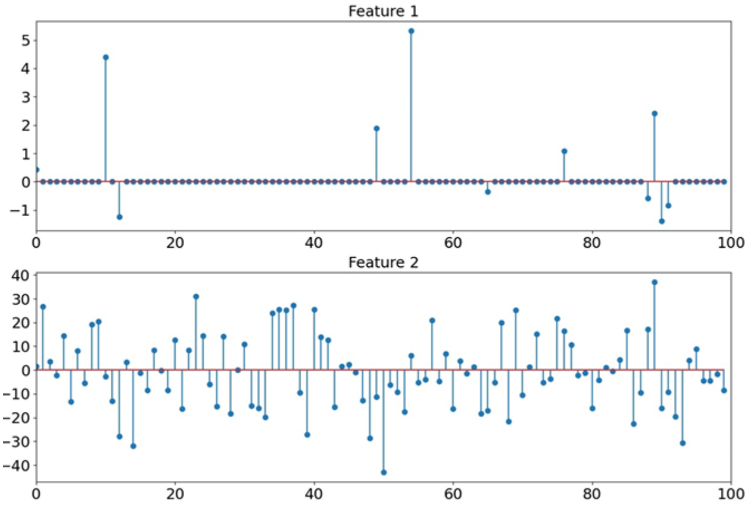

Figure 1: Feature values of a dataset

- **Feature 1:** Max-Abs scaling is more suitable. Majority of the feature values are placed around zero and we can observe some outliers in the data. This is more likely to be a sparse dataset. By using Max-Abs scaling, we can maintain the structure of the feature.
- **Feature 2:** Standard scaling method is suitable for this dataset. This feature is more likely to have a continuous distribution and also has more evenly spread values.

2 Learning from Data

2.1 Why is training and testing data different in each run?

Here we have assigned the random state of the train-test split into a variable that is assigned to a random integer generator. So in each run, it will give a random dataset. There can be 104 multiple unique datasets because the random integer will generate an integer between 0 and 103.

2.2 Why is the linear regression model different from one instance to another?

The regression model is trained using the given dataset. Since we are giving multiple random datasets by changing the random state, the linear regression model is also producing different regression models.

2.3 Increase the number of data samples to 10,000 and repeat the task. What is your observation compared to 100 data samples? State a reason for the different behavior compared to 100 data samples.

When the number of data samples increased to 10,000, the different linear regression models came closer to each other. The reason for this is that the likelihood of the data samples has increased. More data samples provide more data points, helping to capture the true pattern of the data points with less noise and variance.

3 Linear Regression on Real World Data

3.1 How many independent variables and dependent variables are there in the dataset?

Number of Independent Variables: 33
Number of Dependent Variables: 2

3.2 Is it possible to apply linear regression on this dataset? If not, what steps would you follow before applying linear regression?

We can't directly apply linear regression to this model.

- Need to handle the missing values in the dataset.
- Linear regression models require numerical values as input, so if there are any categorical values, we need to encode them into numerical values.
- Need to apply feature scaling methods to standardize and normalize data to improve model performance.

3.3 Code given is used to remove NaN missing values. Is this a correct approach? If not, correct it.

This approach might lead to a misalignment between X and Y indices. We need to combine X and Y values into a single data frame and remove the NaN values. Using this method, there won't be any misalignment in the dataset.

```
data = pd.concat([X, y], axis=1)
data.dropna(inplace=True)
```

3.4 Estimated Coefficients for the Independent Variables:

- Age: 0.0229
- Humidity: 0.0017
- Max1L13_1: -0.1995
- T_LC1: 0.4772
- RCC1: 0.1367

3.5 Which independent variable contributes the most to the dependent feature?

T_LC1 has the highest contribution to the dataset.

3.6 Estimated Coefficients for the Independent Variables:

- Age: 0.0131
- T_OR1: 0.0822
- T_OR_Max1: 0.2311
- T_FHC_Max1: -0.0454
- T_FH_Max1: 0.1964

3.7 Statistical Analyzing Values

Residual Sum of Squares (RSS): 16.1230
Residual Standard Error (RSE): 0.2854
Mean Squared Error (MSE): 0.079
R-squared Value: 0.6027
Standard Errors for each feature:

- Intercept: 0.0109
- Age: 0.0110
- T_OR1: 0.5059
- T_OR_Max1: 0.5048
- T_FHC_Max1: 0.0253
- T_FH_Max1: 0.0254

t-statistics for each feature:

- Intercept: 3403.4438
- Age: 1.1890
- T_OR1: 0.1626
- T_OR_Max1: 0.4578
- T_FHC_Max1: -1.7945
- T_FH_Max1: 7.7210

p-values for each feature:

- Intercept: 0.0000
- Age: 0.2348
- T_OR1: 0.8709
- T_OR_Max1: 0.6472
- T_FHC_Max1: 0.0731
- T_FH_Max1: 0.0000

3.8 Will you be able to discard any features based on p-value ?

High p-values do not provide sufficient evidence to suggest that their coefficients are significantly different from zero.

- *So we can discard the following features from the data set:*
  - Age
  - T_OR1
  - T_OR_Max1

4 Performance evaluation of Linear regression:

4.1 Compute Residual standard error (RSE) for models A and B. Based on RSE for which model performs better?

$$\text{RSE}_A = \sqrt{\frac{9}{10000-2-1}} = \sqrt{\frac{9}{9997}} \approx \sqrt{0.0009003} \approx 0.03$$

$$\text{RSE}_B = \sqrt{\frac{2}{10000-4-1}} = \sqrt{\frac{2}{9995}} \approx \sqrt{0.0002001} \approx 0.01414$$

Model B performs better because it has a lower RSE value.

4.2 Compute R-squared (R^2) for models A and B. Based on R^2, which model performs better?

$$R^2 = 1 - \frac{\text{SSE}}{\text{TSS}}$$

$$R_A^2 = 1 - \frac{9}{90} = 1 - 0.1 = 0.9$$

$$R_B^2 = 1 - \frac{2}{10} = 1 - 0.2 = 0.8$$

Model A performs better because it has higher R squared value.

4.3 Between RSE and R-squared (R2) , which performance metric is more fair for com paring two models and why?

R2 is more fair when comparing, because it gives a measure of how data values are fitted to the model. R2 is a statistical measure that represents the propotion of the variance in the independent variable that is predictable from the independent variable. It is also scale independent.

5 Linearregression impact on outliers

5.1 What happens for $L_1(w)$ and $L_2(w)$ when a→0 ?

$L_1(w)$ when $a \to 0$,

$$L_1(w) \approx \frac{1}{N}\sum_{i=1}^{N}\left(\frac{r_i^2}{r_i^2}\right) = \frac{1}{N}\sum_{i=1}^{N} 1 = 1$$

For large residuals ($a \ll r_i$), $L_1(w)$ loses sensitivity to the magnitude of r_i, treating both large and small residuals similarly. As a result, the loss function is unable to distinguish between them.

$L_2(w)$ when $a \to 0$,

$$L_2(w) \approx \frac{1}{N}\sum_{i=1}^{N}(1-0) = 1$$

Just as with $L_1(w)$, when residuals are large $(a \ll r_i)$ becomes less sensitive to the size of r_i. This means $L_2(w)$ essentially treats both large and small residuals in the same way, making it unable to discern differences between them.

5.2 Minimizing the Influence of Data Points

Since $L_2(w)$ becomes less responsive to very large residuals as a increases, it is recommended to reduce the impact of large residuals. Compared to $L_1(w)$, the $L_2(w)$ function with a $= 25$ effectively balances the penalty for large residuals while maintaining sensitivity to smaller ones, making it more efficient at minimizing the effect of outliers.