Department of Electronic & Telecommunication Engineering
University of Moratuwa

EN3150
Pattern Recognition

Assignment 01

03/09/2024

210642G - Thennakoon T.M.K.R.

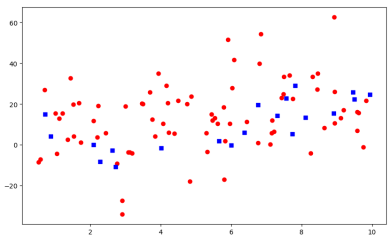
1. Data pre-processing

Feature 1: Feature 1 is centered around 0 and has most values between -1 and 1. Therefore max-abs scaling is a good choice. This scaling method preserve the feature's structure by maintaining the spread in range of -1 to 1, while normalizing it to a similar scale as other features without distorting the distribution.

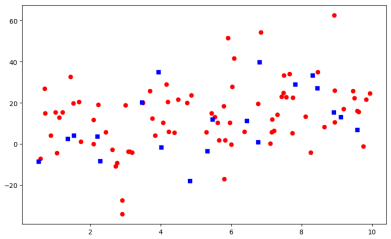
Feature 2: The dataset has a wide range with high variability. Therefore Standard scaling is better. It will standardize this feature by centering it around 0 and normalizing the variance. This scaling will preserve the relative distances and distribution of the data, making it appropriate for models that assume normally distributed data.

2. Learning from data

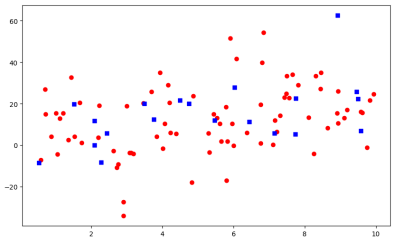
2.
When running the code multiple times, we can observe that the training and testing datasets are different in each run.
This variability in the training and testing datasets across different runs is due to the use of a randomly generated seed (`r`) for the `random_state` parameter in the `train_test_split` function. `random_state` parameter is set to a new random seed value every time the code is executed. This randomness makes data is split differently each time.



1st run

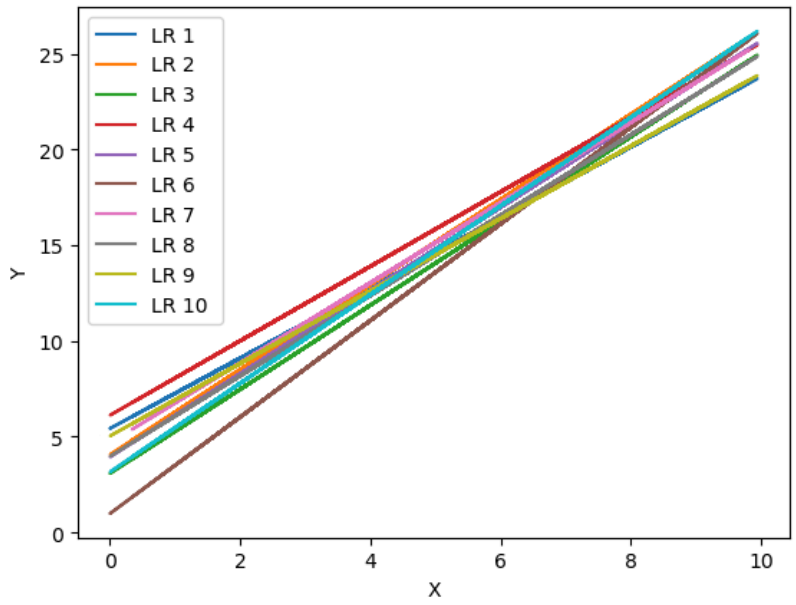


2nd run



3rd run

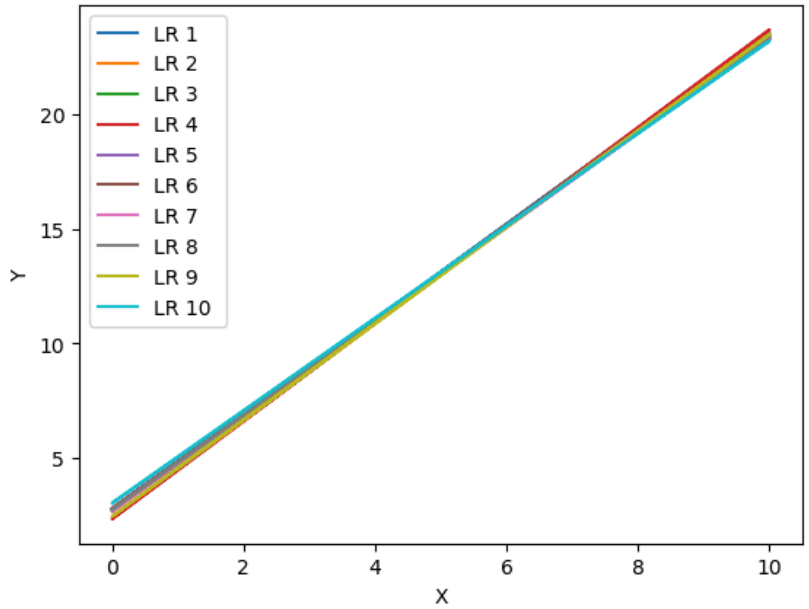
3.
The linear regression models are different across instances because the training data changes with each iteration due to the random splitting process controlled by the `random_state`. Since the training data is different each time, the linear regression model fitted to this data will also be different, leading to variations in the predicted lines.



4.

When the number of data samples is increased to 10,000, we can notice that the variability in the linear regression models across the 10 different instances becomes reduced.

This is because stability provided by the larger and more representative dataset. A large number of dataset helps to reduce variability, stablesized model parameters and decrease sensity of random splits.



3. Linear regression on real world data

2.
Number of independent variables: 33
Number of dependent variables: 2

3.
Linear regression usually works with numerical data. That means features of datasets have numerical values and target values are also have to be numerical. But in this dataset some of the featues are categorical. Therefore we cannot use linear regression for thos features.

Solution for this is to Hot Encoding those categorical features.
First we have to hot encoded features which are categorical. Then it splits those features into sub features which can assign 0 or 1. Then we can input those features for linear regression model.

4.
The approach of dropping missing values separately from `X` and `y` is not entirely correct and could lead to a misalignment between the independent (`X`) and dependent (`y`) variables. Therefore, need to drop the rows with missing values from both `X` and `y` simultaneously to ensure that the indices remain aligned.

It can be done as follows.

```
import pandas as pd
data = pd.concat([X, y], axis=1)
data = data.dropna()
```

7.
Estimated Coefficients:
Coefficient for T_atm: 0.00636477031385094
Coefficient for Humidity: 0.0014354180191543811
Coefficient for Distance: 0.0025143135455469832
Coefficient for T_offset1: 0.18279515887175235
Coefficient for Age_21-25: -0.005293117954369737
Coefficient for Age_21-30: 0.0751627395989575
Coefficient for Age_26-30: -0.14829523423206117
Coefficient for Age_31-40: -0.026279909784433047
Coefficient for Age_41-50: -0.09644371345033831
Coefficient for Age_51-60: -0.32560687421308787
Coefficient for Age_>60: -0.20310069308964

8.
Coefficient for Age_51-60 is highly contributes for dependent feature, Since it has highest absolute weight than others

9.
Estimated Coefficients:
Coefficient for T_OR1: 0.20545776323995343
Coefficient for T_OR_Max1: 0.34819684316001903
Coefficient for T_FHC_Max1: -0.08371846705362082
Coefficient for T_FH_Max1: 0.376564342065323

10.
Residual Sum of Squares (RSS): 15.923399754377353
RSE= 0.14029546674991922
Mean Squared Error (MSE): 0.0780558811489086
R-squared: 0.9287986830855307

Feature	Standard Error	t-value	P-Value
T_OR1	1.559	0.359	0.720
T_OR_Max1	1.559	-0.035	0.972
T_FHC_Max1	0.090	0.594	0.553
T_FH_Max1	0.097	1.874	0.062

11.
All of the p values are above the typical 0.05 significance level. This suggests that none of the features are statistically significant. Therefore, these features should be discarded from the model.

4. Performance evaluation of Linear regression

2.

$$\text{RSE} = \sqrt{\frac{RSS}{N-d-1}}$$

For Model A,

$$\text{RSE} = \sqrt{\frac{9}{10000-2-1}} = 0.03$$

For Model B,

$$\text{RSE} = \sqrt{\frac{2}{10000-4-1}} = 0.0141$$

RSE of Model B is lower than Model A. Therefor Model B performs better.

3.

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

For Model A,

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

For Model B,

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

R^2 is higher for Model A than Model B. Therefore Model A performs better.

4.

R^2 is generally a more fair and interpretable metric for comparing two models, particularly when the models involve different datasets or variables. It provides a clear, relative measure of how well each model explains the variance in the data, making it easier to evaluate and compare their performanc beetween RSE and R^2

5. Linear regression impact on outliers

1.

When a goes to zero L1(w) and L2(w) goes to 1 and it will be a constant value. This results the loss function being independent from $|r_i|$ which gives difference between real y value and predicted value. For a loss function this is not a good scenario to happen because it has to distinguish actual value and predicted value for smaller $|r_i|$ which are not outliers..

2.

The goal is to clamp the loss for residuals where $|r_i| \geq 40$ to ensure that these outliers do not affect the model. L2(w) is likely the better choice for reducing the impact of outliers with $|r_i| \geq 40$, as it can clamp the loss more effectively than L1(w) due to its exponential nature.Thus, **choosing L2 with a range of 3.5 to 18** would be a suitable approach to minimize the influence of outliers in the context described.