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

EN3150 - Pattern Recognition

Assignment 01: Learning from data and related challenges and linear models for regression

210391J - Morawakgoda M.K.I.G.

2nd September 2024

Contents

1 Data pre-processing

1.
   - **Feature 1:**
     For Feature 1, where most values are concentrated around zero and there are a few extreme outliers, **(c) max-abs scaling** is the optimal choice. This scaling method preserves the zero-centered nature of the majority of data points, which is important for maintaining the inherent distribution of the feature. Max-Abs Scaling achieves this by scaling the data relative to the maximum absolute value, thus accommodating the outliers without excessively distorting the scale of the majority of the data.
   - **Feature 2:**
     Feature 2 shows a wide spread of values with significant variability, including both positive and negative numbers. In this case, **(a) standard scaling** is the most suitable approach. By centering the data around zero and scaling it to unit variance, Standard Scaling provides a consistent spread of the values, ensuring that the variability of the feature is appropriately represented.

2 Learning from data

2.
   - **Observations :** Running the code provided in Listing 2 multiple times results in different training and testing data splits for each execution. This variation is evident in the plot, where the red (training data) and blue (testing data) points differ with each run.

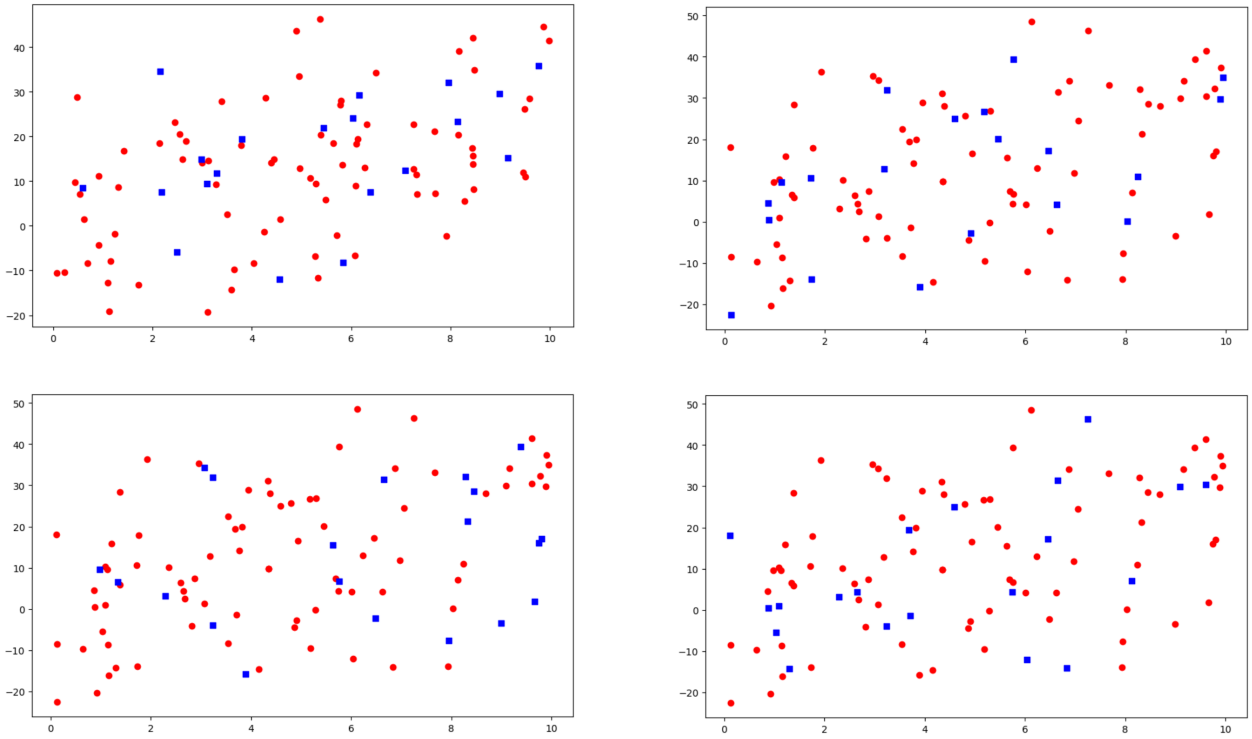

Figure 1: Variation in the training and testing data splits

   - **Reasoning :** The variation occurs because of the use of a randomly generated `random_state` parameter in the `train_test_split` function. The line, `r = np.random.randint(104)` generates a random integer between 0 and 103, which is then used as the `random_state` parameter in `train_test_split`. This parameter controls the shuffling of the data before splitting. Different values of `random_state` will lead to different splits of the data, hence causing variation in the training and testing sets.

3. The linear regression model is different from one instance to another because each instance uses a different subset of the data for training. This results from the random selection of the training set in each iteration, leading to the model learning different relationships between X and Y for each random split.

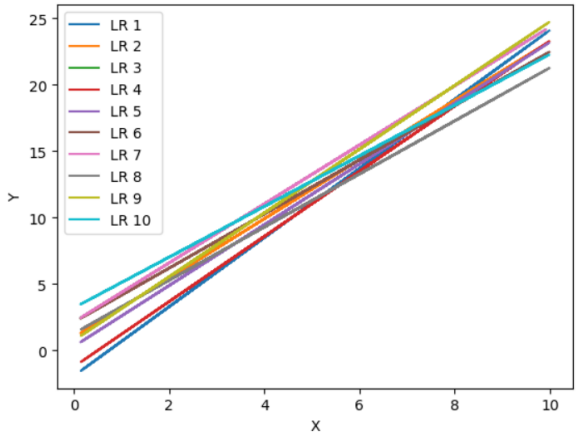

Figure 2: High variation between regression models (for 100 data samples)

4. • **Observations :** When the number of data samples is increased to 10,000 (`n_samples = 10000`), and the code in Listing 3 is repeated, the linear regression lines from different instances become much more similar to each other. Unlike with 100 samples, where each iteration of the linear regression could produce visibly different lines, with 10,000 samples, the variation between the lines is minimal, and they almost overlap.

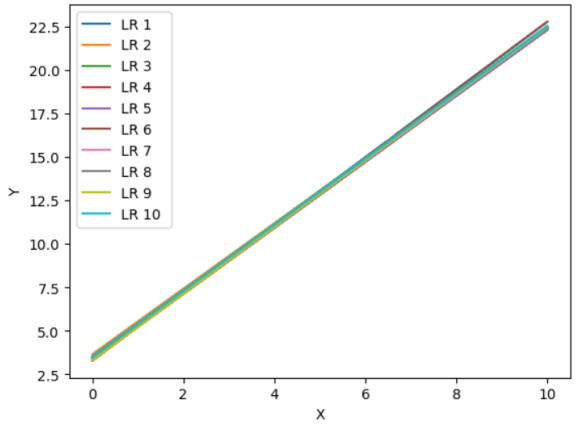

Figure 3: Less variance between regression models (for 10000 data samples)

• **Reasoning :** With 10,000 data samples, the dataset offers greater stability and more accurately represents the relationship between X and Y, capturing the true distribution and reducing the impact of noise and outliers. Larger sample sizes lead to lower variance in training sets, resulting in consistent model outcomes. The average of sample outcomes converges to the expected value as the number of samples increases, meaning random variations diminish, revealing the true relationship between variables. This reduces the impact of random splitting, as each split remains representative of the overall data distribution, yielding similar model results.

3 Linear regression on real world data

2. There are 2 dependent variables(`Target`) and 33 independent variables(`Feature`).

	name	role	type	demographic
0	SubjectID	ID	Categorical	None
1	aveOralF	Target	Continuous	None
2	aveOralM	Target	Continuous	None
3	Gender	Feature	Categorical	Gender
4	Age	Feature	Categorical	Age
5	Ethnicity	Feature	Categorical	Ethnicity
6	T_atm	Feature	Continuous	None
7	Humidity	Feature	Continuous	None
8	Distance	Feature	Continuous	None
9	T_offset1	Feature	Continuous	None
10	Max1R13_1	Feature	Continuous	None
11	Max1L13_1	Feature	Continuous	None
12	aveAll1R13_1	Feature	Continuous	None
13	aveAll1L13_1	Feature	Continuous	None
14	T_RC1	Feature	Continuous	None
15	T_RC_Dry1	Feature	Continuous	None
16	T_RC_Wet1	Feature	Continuous	None
17	T_RC_Max1	Feature	Continuous	None
18	T_LC1	Feature	Continuous	None
19	T_LC_Dry1	Feature	Continuous	None
20	T_LC_Wet1	Feature	Continuous	None
21	T_LC_Max1	Feature	Continuous	None
22	RCC1	Feature	Continuous	None
23	LCC1	Feature	Continuous	None
24	canthiMax1	Feature	Continuous	None
25	canthi4Max1	Feature	Continuous	None
26	T_FHCC1	Feature	Continuous	None
27	T_FHRC1	Feature	Continuous	None
28	T_FHLC1	Feature	Continuous	None
29	T_FHBC1	Feature	Continuous	None
30	T_FHTC1	Feature	Continuous	None
31	T_FH_Max1	Feature	Continuous	None
32	T_FHC_Max1	Feature	Continuous	None
33	T_Max1	Feature	Continuous	None
34	T_OR1	Feature	Continuous	None
35	T_OR_Max1	Feature	Continuous	None

Figure 4: Dependent and Independent variables.

3. No , it is not possible to apply linear regression to this dataset directly. Therefore following steps should be taken before applying linear regression to this dataset.

   - Convert categorical features (`Gender, Age, Ethnicity`) into numerical values using **One-Hot Encoding.**
   - Check for multicollinearity and consider removing or combining highly correlated features.
   - Use feature scaling to ensure that features contribute equally to the model.
   - Handle missing values properly(Drop or Impute).

4.
   - This is not a correct approach as it drops missing values separately for X and Y, which can lead to misalignment between the features and the target variables. After dropping missing values, the indices of X and Y may not match, which could cause errors or unexpected behavior when training the model.
   - Hence, rows with missing values from the combined dataset should be dropped, ensuring that only rows with complete data for both X and y are kept.

Listing 1: Handling Missing Values

```
# Combine X and y into a single DataFrame to ensure alignment
data = pd.concat([X, y], axis=1)

# Drop rows with missing values from the combined DataFrame
data = data.dropna()

# Split back into X and y
X = data.drop(columns=y.columns)
y = data[y.columns]
```

5. **Selected features :** `Age, T_atm, Humidity, Distance, T_RC_Max1`

7. Following table shows the each selected feature with their estimated coefficients after the training process.

Feature	Estimated coefficient
T_atm	-0.053396
Humidity	0.001499
Distance	0.002541
T_RC_Max1	0.765347
Age_18-20	-0.165189
Age_21-25	-0.117728
Age_21-30	0.002035
Age_26-30	-0.106107
Age_31-40	-0.156186
Age_41-50	0.077870
Age_51-60	-0.060913
Age_>60	0.526218

8. T_RC_Max1 feature contributes highly for the dependent feature as its estimated coefficient is the highest (0.765347) among the selected features.

9. Following table shows the each given feature with their estimated coefficients after the training process.

Feature	Estimated coefficient
T_OR1	0.091997
T_OR_Max1	0.464070
T_FHC_Max1	-0.087332
T_FH_Max1	0.370886

10.
- RSS = 79.02804487558424
- RSE = 0.3115863447195744
- MSE = 0.09684809421027481
- R^2 = 0.643312635118277

Feature	Coefficient	Standard Error	t-statistic	p-value
T_OR1	0.091997	0.019513	4.714745	2.846581e-06
T_OR_Max1	0.464070	0.019513	23.782064	2.430323e-95
T_FHC_Max1	-0.087332	0.018743	-4.659507	3.701784e-06
T_FH_Max1	0.370886	0.020652	17.958785	5.287498e-61

11. Based on the calculated p-values, none of the features should be discarded, as they all have p-values near zero, indicating strong evidence against the null hypothesis

4 Performance evaluation of Linear regression

2. • **For Model A:**

$$\begin{aligned} \text{SSE} &= 9 \\ N &= 10000 \\ \text{Number of independent variables } d &= 2 \ (x_1, x_2) \end{aligned}$$

$$\text{RSE}_A = \sqrt{\frac{\text{SSE}}{N-d-1}} = \sqrt{\frac{9}{10000-2-1}} = \sqrt{\frac{9}{9997}} \approx \sqrt{0.00090027} \approx 0.03$$

• **For Model B:**

$$\begin{aligned} \text{SSE} &= 2 \\ N &= 10000 \\ \text{Number of independent variables } d &= 4 \ (\ x_1, x_2, x_3, x_4) \end{aligned}$$

$$\text{RSE}_B = \sqrt{\frac{\text{SSE}}{N-d-1}} = \sqrt{\frac{2}{10000-4-1}} = \sqrt{\frac{2}{9995}} \approx \sqrt{0.00020002} \approx 0.014$$

• Based on RSE, Model B performs better because it has a lower residual standard error.

3. • **For Model A:**

$$\begin{aligned} \text{SSE} &= 9 \\ \text{TSS} &= 90 \\ R_A^2 &= 1 - \frac{\text{SSE}}{\text{TSS}} = 1 - \frac{9}{90} = 1 - 0.1 = 0.9 \end{aligned}$$

• **For Model B:**

$$\begin{aligned} \text{SSE} &= 2 \\ \text{TSS} &= 10 \\ R_B^2 &= 1 - \frac{\text{SSE}}{\text{TSS}} = 1 - \frac{2}{10} = 1 - 0.2 = 0.8 \end{aligned}$$

• Based on R-squared, Model A performs better because it has a higher R-squared value.

4. R^2 is generally more fair for comparing models because it accounts for the number of predictors used. It adjusts for the model's complexity and provides a normalized measure of performance. RSE, on the other hand, can be heavily influenced by the scale of the response variable and does not inherently account for the number of predictors.

5 Linear regression impact on outliers

2. • For $L_1(w)$:

$$L_1(w) = \frac{1}{N}\sum_{i=1}^{N}\frac{r_i^2}{a^2 + r_i^2}$$

As $a \to 0$:

– If r_i is non-zero, a^2 becomes negligible compared to r_i^2.

– Thus, $\frac{r_i^2}{a^2+r_i^2} \approx \frac{r_i^2}{r_i^2} = 1$.

So, the function $L_1(w)$ approaches:

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

The loss function $L_1(w)$ becomes a constant value of 1, independent of the residuals. This means the loss function does not differentiate between residuals anymore, treating all residuals the same regardless of their size.

• For $L_2(w)$:

$$L_2(w) = \frac{1}{N}\sum_{i=1}^{N}\left(1 - \exp\left(-\frac{2|r_i|}{a}\right)\right)$$

As $a \to 0$:

– For large $|r_i|$, $\frac{2|r_i|}{a}$ becomes very large.

– Therefore, $\exp\left(-\frac{2|r_i|}{a}\right) \to 0$, making $1 - \exp\left(-\frac{2|r_i|}{a}\right) \to 1$.

– For small $|r_i|$, $\frac{2|r_i|}{a}$ is still small, so $1 - \exp\left(-\frac{2|r_i|}{a}\right)$ is approximately $\frac{2|r_i|}{a}$.

Thus, the function $L_2(w)$ approaches:

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

Similar to $L_1(w)$, the loss function $L_2(w)$ also becomes constant, treating all residuals uniformly when $a \to 0$.

3. To minimize the influence of data points with $|r_i| \geq 40$, we need to choose a value of a and a function that effectively reduces the impact of these large residuals. Based on the calculations for $L_{1,i}$ and $L_{2,i}$ with $r_i = 40$, let's analyze the impact of different values of a on both loss functions:

• **For $a = 2.5$:**

– $L_{1,i} \approx 0.9961$

– $L_{2,i} \approx 1$

Both L_1 and L_2 are very close to their maximum possible values (1), indicating that with such a small value of a, the influence of the large residual is still quite significant. This implies that $a = 2.5$ does not sufficiently reduce the impact of outliers.

- **For $a = 25$:**
  - $L_{1,i} \approx 0.7191$
  - $L_{2,i} \approx 0.9592$

  Here, L_1 shows a more noticeable reduction, indicating that the influence of the residual is being moderated. L_2 also shows some reduction, though it remains relatively high. This suggests that $a = 25$ starts to have a meaningful impact on reducing the influence of large residuals, but L_2 still allows some influence from the outlier.
- **For $a = 100$:**
  - $L_{1,i} \approx 0.1379$
  - $L_{2,i} \approx 0.5507$

  With $a = 100$, both L_1 and L_2 show significant reductions, particularly L_1. This indicates that larger values of a effectively dampen the influence of large residuals, thus reducing the impact of outliers.
- **Choosing the Function:** L_1 is generally more sensitive to the choice of a for large residuals, as seen from the significant reduction in influence when $a = 100$. This sensitivity makes L_1 a better choice for robustly handling outliers, as it provides a more pronounced reduction in the influence of large residuals.
- **Choosing the Value of a:** Based on the analysis, a larger value of a, such as $a = 100$, is preferable because it significantly reduces the influence of large residuals like $r_i = 40$. Using $a = 100$ minimizes the impact of these outliers without completely ignoring them, providing a balance between robustness and sensitivity.