EN3150 Assignment 01: Learning from data and related challenges and linear models for regression

Sampath K. Perera

August 13, 2024

1 Data pre-processing

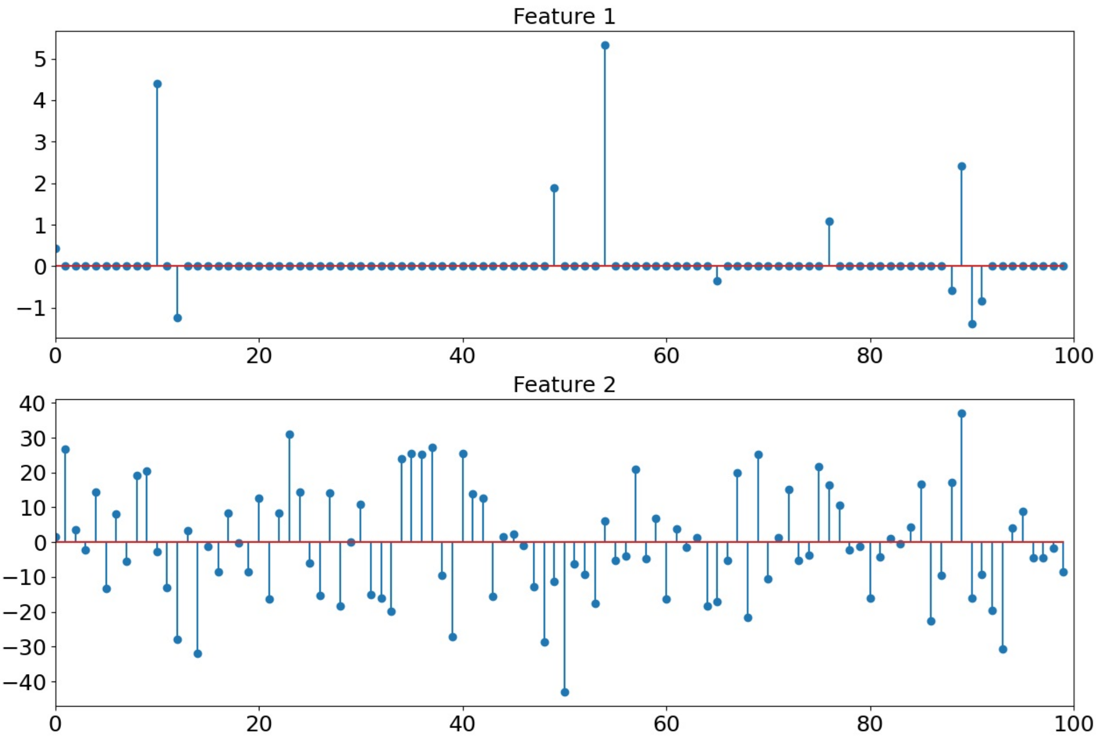

Figure 1: Feature values of a dataset.

1. Feature values of two features is shown in Fig.1. Considering the scaling methods of (a) standard scaling, (b) min-max scaling, and (c) max-abs scaling. Select one scaling method for feature 1 and 2, ensuring that the chosen method preserves the structure/properties of the features. Justify your answer. [5 marks]

2 Learning from data

1. Use the code given in listing 1 to generate data.

```
import numpy as np
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression

# Generate 100 samples
n_samples = 100

# Generate X values (uniformly distributed between 0
    and 10)
X = 10 * np.random.rand(n_samples, 1)

# Generate epsilon values (normally distributed with
    mean 0 and standard deviation 15)
epsilon = np.random.normal(0, 15, n_samples)

# Generate Y values using the model Y = 3 + 3X +
    epsilon
Y = 3 + 2 * X + epsilon[:, np.newaxis]
```

Listing 1: Data generation.

2. Run the code given in listing 2 multiple times and write down your observation. Why training and testing data is different in each run? [2.5 marks]

```
r=np.random.randint(104)

# Split the data into training and test sets (80% train,
    20% test)
X_train, X_test, Y_train, Y_test = train_test_split(X, Y,
    test_size=0.2, random_state=r)

# Plot the data points
plt.figure(figsize=(10, 6))
plt.scatter(X_train, Y_train, alpha=1, marker='o',color='
    red',label='Training Data')
plt.scatter(X_test, Y_test, alpha=1, marker='s',color='
    blue',label='Testing Data')
plt.show()
```

Listing 2: Data visualization.

3. Use the code given in listing 3 to fit a linear regression model. Why linear regression model is different from one instance to other instance ? [2.5 marks]

```
for i in range(10):  # Plotting 10 different instances
    X_train, X_test, Y_train, Y_test = train_test_split(X,
        Y, test_size=0.2, random_state=np.random.randint
        (104))
    model = LinearRegression()
    model.fit(X_train, Y_train)
    Y_pred_train = model.predict(X_train)
    plt.plot(X_train, Y_pred_train, label=f'LR {i+1}')
plt.xlabel('X')
plt.ylabel('Y')
plt.legend()
plt.show()
```

Listing 3: Linear regression.

4. Increase the number of data samples to 10,000 (n_samples = 10000 in listing 1) and repeat the task 3. What is your observation in comparison to 100 data samples ? State a reason for the different behavior compared to 100 data samples. [10 marks]

3 Linear regression on real world data

1. Load the dataset given in this url. Use the code given in listing 4 to load data.
2. How many independent variables and dependent variables are there in the data set? [2 marks]
3. Is it possible to apply linear regression on this dataset? If not, what steps would you follow before applying linear regression? [3 marks]
4. Code given in is used to remove NaN/missing values. Is this a correct approach? If not correct it. [5 marks]
5. Select "aveOralM" as the dependent feature. For the independent features, select 'Age' and four other features based on your preference.
6. Split the data into training and testing sets with 80% of data points for training and 20% of data points for testing.
7. Train a linear regression model and estimate the coefficient corresponds to independent variables. List the estimated coefficients. [15 marks]
8. Which independent variable contributes highly for the dependent feature? [5 marks]
9. Select 'T_OR1', 'T_OR_Max1', 'T_FHC_Max1', 'T_FH_Max1' features as independent features. Train a linear regression model and estimate the coefficient corresponds to independent variables. [5 marks]

10. Calculate followings [10 marks]

    - Residual sum of squares (RSS)
    - Residual Standard Error (RSE)
    - Mean Squared Error (MSE)
    - R^2 statistic
    - Standard error for each feature
    - t-statistic for each feature
    - p-value for each feature

    Note that RSE is given by

    $$\mathrm{RSE} = \sqrt{\frac{\mathrm{RSS}}{N-d-1}}. \tag{1}$$

    Here, N is the total number of data samples and d is the number of independent features.

11. Will you be able to discard any features based on p-value ? [5 marks]

```
# If package not installed, install it using pip install ucimlrepo
from ucimlrepo import fetch_ucirepo

# fetch dataset
infrared_thermography_temperature = fetch_ucirepo(id=925)

# data (as pandas dataframes)
X = infrared_thermography_temperature.data.features
y = infrared_thermography_temperature.data.targets

# metadata
print(infrared_thermography_temperature.metadata)

# variable information
print(infrared_thermography_temperature.variables)
```

Listing 4: Load data

```
    # Drop rows with missing values from both X and y
    X = X.dropna()
    y = y.dropna()
```

Listing 5: Data cleaning

4 Performance evaluation of Linear regression

1. Consider the linear regression models Model A: $y = w_0 + w_1x_1 + w_2x_2$ and Model B:$y = w_0 + w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4$. Sum of squared errors (SSE) and total sum of squares (TSS) of these models are given in Table 1.

Table 1: SSE and TSS of linear regression models.

	Model A	Model B
SSE= $\sum_{i=1}^{N}(y_i - \boldsymbol{w}^T\boldsymbol{x}_i)^2$	9	2
TSS= $\sum_{i=1}^{N}(y_i - \tilde{y}_i)^2$	90	10
Number of data samples (N)	10000	10000

2. Compute residual standard error (RSE) for models A and B. Based on RSE for which model performs better? [1 mark]

3. Compute R-squared (R^2) for models A and B. Based on R^2 for which model performs better? [1 mark]

4. Between RSE and R-squared (R^2) , which performance metric is more fair for comparing two models and why? [3 marks]

5 Linear regression impact on outliers

1. Linear regression is known to be less robust in the presence of outliers. To reduce the impact of outliers, a modified loss function is introduced. Suppose, following two modified functions

$$L_1(\boldsymbol{w}) = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{r_i^2}{a^2 + r_i^2}\right) = \frac{1}{N}\sum_{i=1}^{N}\left(L_{1,i}\right).$$

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

Here, residual is given by $r_i = \hat{y}_i - y_i$, where y_i and $\hat{y}_i = \boldsymbol{w}^T\boldsymbol{x}_i$ are true and liner regression model outputs for $i-$th data sample. Further, "a" (≥ 0) is a hyper-parameter. Figure 2 shows behavior of $L_1(\boldsymbol{w})$ and $L_2(\boldsymbol{w})$ with respect to different "a" values.

2. What happens when $a \to 0$? [1] [10 marks]

3. Suppose we need to minimize the influence of data points with $|r_i| \geq 40$. What value(s) of "a" and what function(s) would you choose, and why? [15 marks]

[1] You may compare behavior of LS with this modified loss function and LS with standard loss function

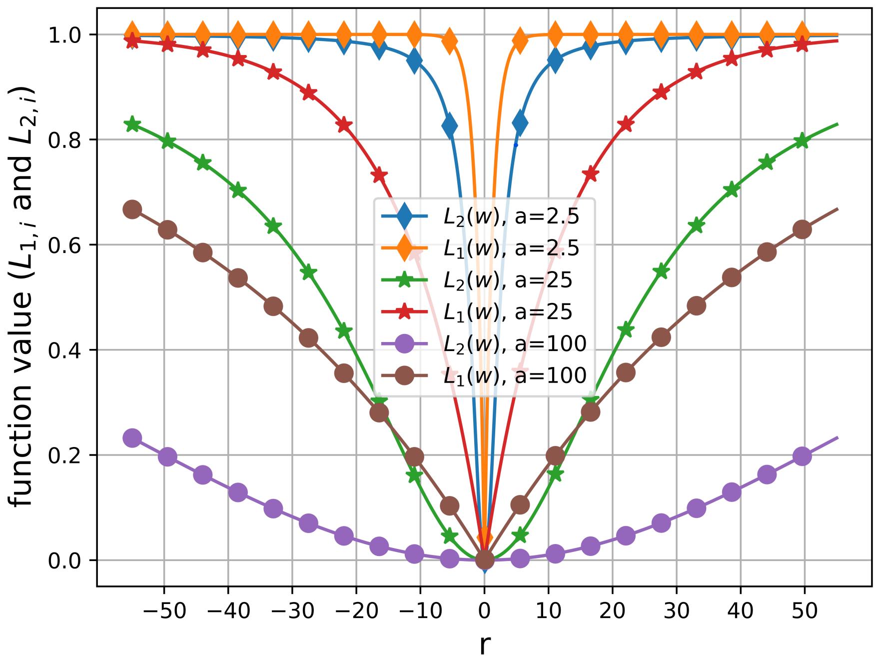

Figure 2: $L_1(\boldsymbol{w})$ and $L_2(\boldsymbol{w})$ with respect to different "a" values.

6 Additional Resources

1. Scikit-learn preprocessing data
2. Introduction to sparsity in signal processing
3. sklearn linear regression

7 Submission

- Upload a report and your codes as a zip file named as "EN3150_your_indexno_A01.zip". Include the index number and the name within the report as well. Please include all your answers in the report.
- Pay careful attention to formatting such as font size, spacing, and margins.
- Include a title page with necessary information (e.g., title, author, date, index no).
- Use consistent and professional formatting throughout the document.
- Plagiarism will be checked and in cases of plagiarism, an extra penalty of 50% will be applied. In case of copying from each other, both parties involved will receive a

grade of zero for the assignment. Academic integrity is of utmost importance, and any form of plagiarism[2] or cheating will not be tolerated.

- An extra penalty of 15% is applied for late submission.

[2] https://en.wikipedia.org/wiki/Plagiarism