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

Sampath K. Perera

August 23, 2023

1 Data preprocessing

1. Use the code given in listing 1 to generate data. Please use your **index number** for data generation.

```
import numpy as np
import matplotlib.pyplot as plt

def generate_signal(signal_length, num_nonzero):
signal = np.zeros(signal_length)
nonzero_indices = np.random.choice(signal_length,
   num_nonzero, replace=False)
nonzero_values = 50*np.random.randn(num_nonzero)
signal[nonzero_indices] = nonzero_values
return signal

#Data generation
signal_length = 100  # Total length of the signal
num_nonzero = 10
your_index_no=40266 # Enter your index number (without
    English letter and without leading zeros)
signal = generate_signal(signal_length, num_nonzero)
signal[10] = (your_index_no % 10)*10 + 10
if  your_index_no % 10 == 0:
signal[10] = np.random.randn(1) + 30


signal=signal.reshape(signal_length,1)

#plotting
plt.figure(figsize=(15,5))
plt.subplot(1, 1, 1)
plt.title("Data")
plt.stem(signal)
```

Listing 1: Data generation.

2. Plot the generated data (signal).

3. Apply following normalization methods

   - MaxAbsScaler (preprocessing.MaxAbsScaler() from *sklearn.preprocessing*)
   - Implement min-max and standard normalization yourself and apply the normalization on data. The relvent equations are listed below. Further, as an example min-max scale function can be implemented as given in listing 2.

   Min-Max scaling equation

$$x_{\text{scaled}} = \frac{x - \min(x)}{\max(x) - \min(x)}. \tag{1}$$

   Where:

x_{scaled} is the scaled value of x.
x is the original value.
$\min(x)$ is the minimum value in the data.
$\max(x)$ is the maximum value in the data.

   Standardization scaling equation

$$x_{\text{scaled}} = \frac{x - \mu}{\sigma}. \tag{2}$$

   Where:

x_{scaled} is the scaled value of x.
x is the original value.
μ is the mean (average) of the data.
σ is the standard deviation of the data.

```
def min_max_scale(data):
min_val = np.min(data)
max_val = np.max(data)
scaled_data = (data - min_val) / (max_val - min_val)
return scaled_data
```

Listing 2: Min-max scale function.

4. Visualize the data before and after normalization. Create stem plots of the original and normalized data to visualize the effects of each normalization method on the data.

5. How many none-zero elements in the data before the normalization and after the normalization.

6. Compare how each normalization method scales the data and its impact on structure of the data.

7. Discuss the effects of each normalization method on the data's distribution, structure, and scale. Which normalization approach you recommend for this kind of data and what is the reason behind this?

2 Linear regression on real world data

1. Load the dataset given in this url. The data illustrates the relationship between advertising budgets (in thousands of dollars) allocated to TV, radio, and newspaper media and the corresponding sales (in thousands of units) for a specific product. Use the code given in listing 3 to load data from CSV.

2. Split the data into training and testing sets with 80% of data points for training and 20% of data points for testing.

3. Train a linear regression model and estimate the coefficient corresponds to independent variables (advertising budgets for TV, radio and newspapers).

4. Evaluate train model on testing data, calculate following statistics for testing and training data.

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

   Note that RSE is given by

$$\text{RSE} = \sqrt{\frac{\text{RSS}}{N-d}}. \tag{3}$$

   Here, N is the total number of data samples and d is the number of model parameters. As an example for a model $y = w_0 + w_1 x + \epsilon$, there are two model parameters namely w_0 and w_1.

5. Is there a relationship between advertising budgets and sales?

6. Which independent variable contributes highly on sales?

7. One may argue that possibly, allocating 25,000 dollars both television advertising and radio advertising individually (i.e., 25,000 dollars for TV and 25,000 dollars for radio) yields higher sales compared to investing 50,000 dollars in either television or radio advertising individually. Based on your trained model, comment on this argument. Here, assume that budged allocated for newspapers is zero.

```
import numpy as np
import pandas as pd

# Load data from CSV

file_path = r'path_to_your_csv_file.csv'

df = pd.read_csv(file_path)

print(df.head())
```

Listing 3: Load data from CSV.

3 Linear regression impact on outliers

1. You are given set of data points related to independent variable (x) and dependent variable (y) in Table 1.

Table 1: Data set.

i	x_i	y_i
1	0	20.26
2	1	5.61
3	2	3.14
4	3	−30.00
5	4	−40.00
6	5	−8.13
7	6	−11.73
8	7	−16.08
9	8	−19.95
10	9	−24.03

2. Use all data given in Table 1 to find a linear regression model. Plot x, y as a scatter plot and plot your linear regression model in the same scatter plot.

3. You are given two linear models as follows.

   - Model 1: $y = -4x + 12$

- Model 2: $y = -3.55x + 3.91$

Here, model 2 is your linear regression model which is learned in task 2.
A robust estimator is introduced to reduce the impact of the outliers. The robust estimator finds model parameters which minimize the following loss function

$$L(\theta, \beta) = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{(y_i - \hat{y}_i)^2}{(y_i - \hat{y}_i)^2 + \beta^2} \right). \tag{4}$$

Here, θ represents model parameters, $\beta = 1$ and number of data samples $N = 10$, respectively. Note the y_i and $\hat{y}_i$ are true and predicted $i-$th data sample, respectively.

4. For the given two models in task 3, calculate the loss function $L(\theta, \beta)$ values for all data samples using eq. (4) (you may use a computer program to calculate this).
5. Utilizing this robust estimator, determine the most suitable model from the models specified in task 3 for the provided dataset. Justify your selection.
6. How does this robust estimator reduce the impact of the outliers?
7. Plot models specified in task 3 and data point to visualize the impact of the outliers.
8. Briefly discuss the impact on β in eq. (4) to in the context of reducing the impact of the outliers.

4 Additional Resources

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

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.
- The interpretation of results and the discussion are important in the report.
- An extra penalty of 10% is applied for late submission.
- Plagiarism will be checked and in cases of plagiarism, an extra penalty of 10% will be applied.