Assignment 2: Logistic Regression

Seoyoung Yoo
Matrikel-Nr. 1904668
06 November 2022

1 Dataset Statistics

Explore and preprocess the dataset.

a) Look at the kernel density plot (code provided) of all features and discuss what you see (or don't see).

As a result of observing the data through Exploratory Data Analysis (EDA) through the kernel density plot, it was found that most of the feature values had a small value such as 0. Looking at all features through Figure 1.1, it is not possible to obtain meaningful results because the characteristics of the data are not well revealed. As shown in Figure 1.2, because of changing the range of the x-axis to [-0.3,0.3], the characteristics of the features are well revealed, and meaningful results can be obtained.

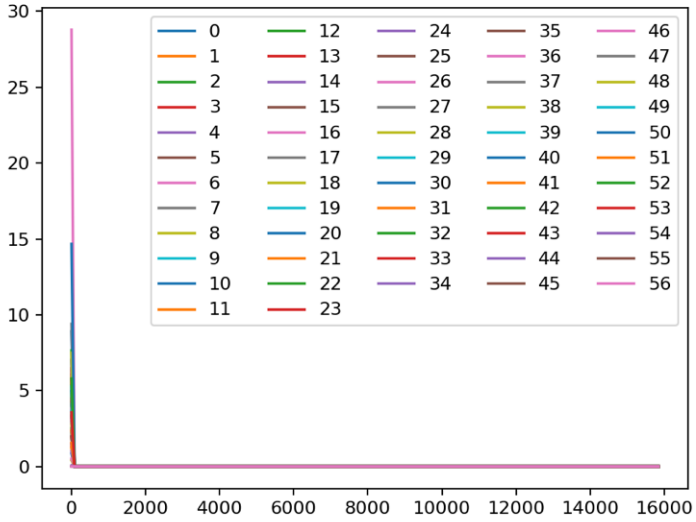


Figure 1.1 Kernel density plot with all features

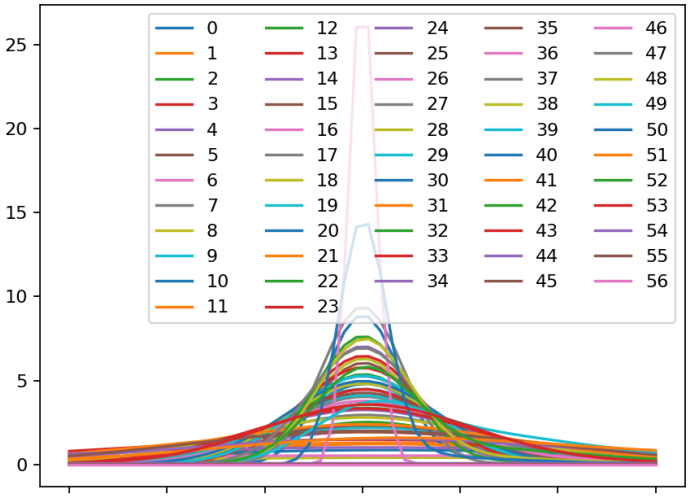


Figure 1.2 Kernel density plot with features range from 0.3 to 0.3

b) Normalize the data using z-scores, i.e., normalize each feature to mean 0 and variance 1. Normalize both training and test data. In particular, think about how test data should be normalized.

Using the formula for calculating a z-score is z = (x-μ)/σ. From the code, Xz refers to the score of train data's z-score and Xtestz refers to the score test data's z-score.

$$Z = \frac{X - \mu}{\sigma}$$

where X is the raw score, μ is the population mean, and σ is the population standard deviation

c) Redo the kernel density plot on the normalized data. What changed? Is there anything that "sticks out"?

From Figure 1.3, the data is distributed on both sides with the features 0 as the center in the x-axis range [-5,5]. By z-score standardization, each data is converted into a value indicating how far apart it is from the mean. The converted data has the characteristic of being flattened and the amplitude is reduced. The interval between each data is decreased due to the decrease in the amplitude.

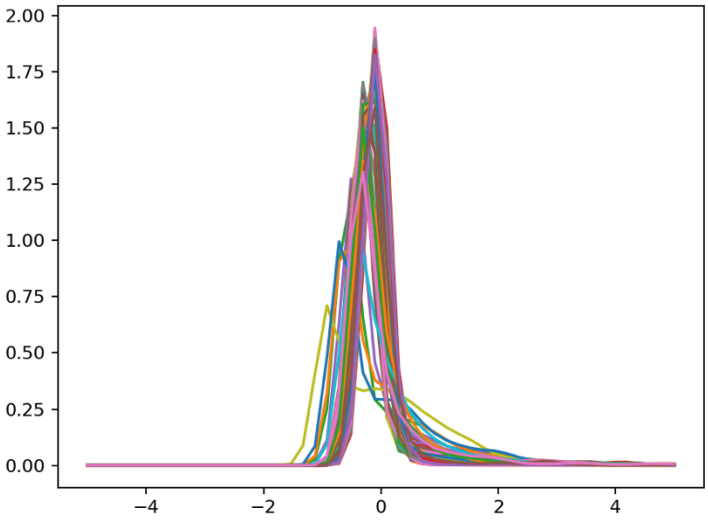


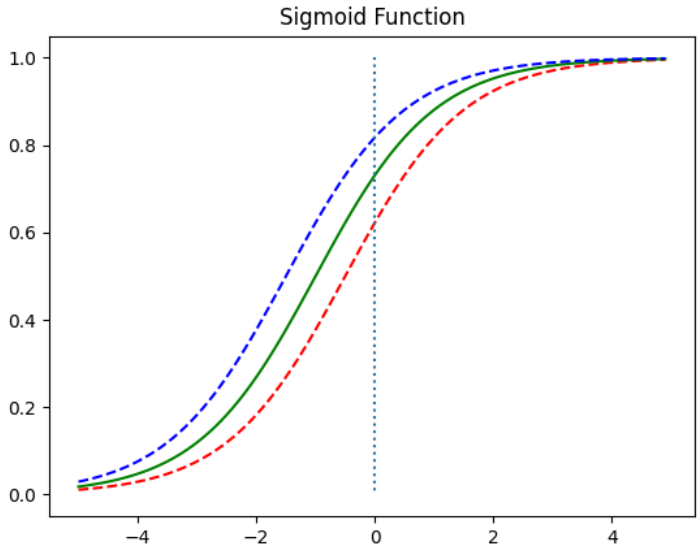
Figure 1.3 Kernel density plot with the normalized data

2 Maximum Likelihood Estimation

a) Show analytically that if we use a bias term, rescaling (multiply by constant) and shifting (add a constant) features leads to ML estimates with the same likelihood. Why do you think we computed z-scores then?

The sigmoid function used for logistic regression is a monotonic function, and it can be seen that the same likelihood function value is the same even if a bias term is added.

The graph (Figure 1.4) below shows how the graph moves according to the bias term value. (b=1.5, b=1, b=0.5 respectively from left to right) The sigmoid function converges to 1 as the input increases



and converges to 0 as the input decreases. It has a value from 0 to 1, and if the output value is 0.5 or more, it is 1 (True), and if it is 0.5 or less. As the bias increased, the graph shifted to the left in parallel.

$$\mathbf{H}(\boldsymbol{x}) = \frac{\mathbf{1}}{\mathbf{1} + \boldsymbol{e}^{-(\boldsymbol{\omega x}+\boldsymbol{b})}} = \boldsymbol{sigmoid}(\boldsymbol{\omega x} + \boldsymbol{b}) = \boldsymbol{\sigma}(\boldsymbol{\omega x} + \boldsymbol{b})$$

Figure 2.1 Sigmoid Function: b=1.5, b=1, b=0.5 respectively

In general, the reasons for z-score normalization are: 1) it can reduce error variance and increase statistical power, 2) increase normality because outliers are randomly distributed, and 3) reduce bias in the estimate of practical interest. This is because it has less influence.

In the case of the MLE calculation, the task is classified as the same value as before even when a bias term is added, whereas in the case of MAP (Maximum Posterior Estimation), the z-score normalization is performed because the classification operation is not properly affected by the posterior value. will be.

e) Explore the behavior of both methods for the parameters provided to you. Discuss!

Figures 2.2 and 2.3 show that convergence is faster when using stochastic gradient descent than when using gradient descent. Gradient Descent initializes the parameter theta, which means that all values are set to 0.

The gradient of J (theta) is obtained from theta set as the initial value, the value obtained by subtracting the gradient from the initial value is used as the next theta value, and this operation is repeated until theta converges. For this reason, there is a disadvantage that a lot of time and memory are required for model training because the entire data set is considered when updating parameters. In stochastic gradient descent, the process of updating theta by finding the gradient of the loss function on one data is alternately repeated for all data sets. It can be seen that convergence is faster by improving the disadvantage of viewing all m data sets per single iteration.

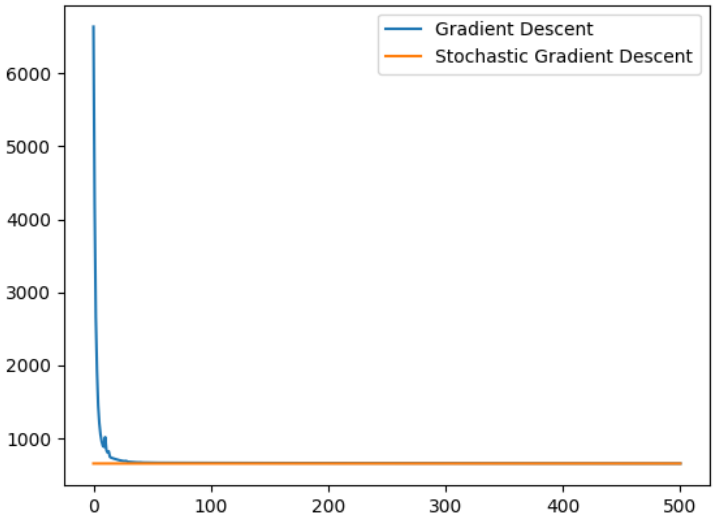


Figure 2.2

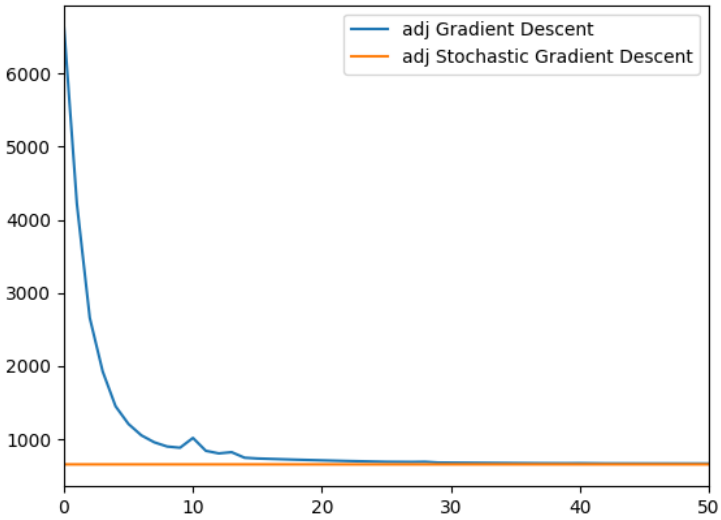


Figure 2.3

3 Prediction

Explore the models that you fit in the previous task and discuss. Study the composition of the weight vector: which features are important, which are not? Is this intuitive?

From Figure 2.4, it can be seen that the fourth feature value has the largest value with a weight of 8.086 and is the most important. On the other hand, it can be seen that the 25th feature is the one with the most -2.0766 value.

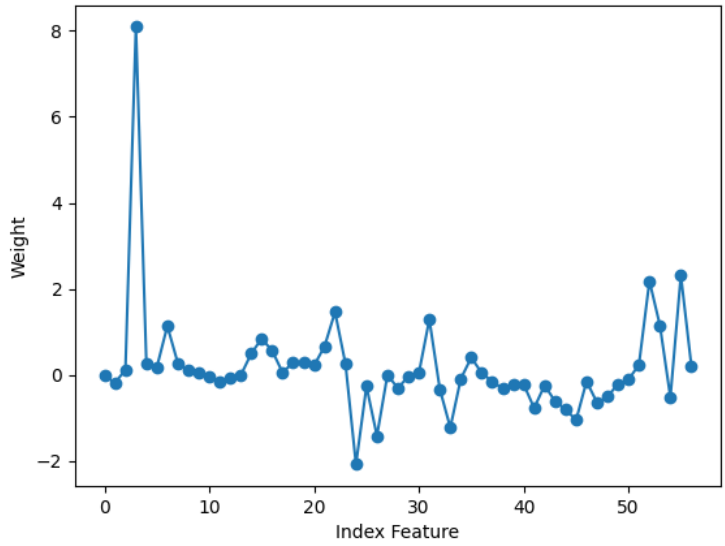


Figure 2.4

4 Maximum Aposteriori Estimation

a) **Implement gradient descent for MAP estimation of logistic regression with a Gaussian prior / L2 regularization with hyperparameter λ. You can reuse the methods of your solution for MLE.**

b) **Study the effect of the prior on the result by varying the value of λ. Consider at least the training data log-likelihood, the test data log-likelihood, and the prediction accuracy. Are these results surprising to you?**

c) **Study the composition of the weight vector for varying choices of λ (try very large values). Try to explain what you saw in the task above.**