**Machine Learning Model with Linear Regression with L1 and L2 Regularization using Python**

**Introduction**

Linear regression is one of the famous and well understood algorithms in statistics and are hence widely deployed in machine learning. Linear models make predictions using linear function of input features and are mathematically expressed as follows:

ŷ=ω[0]x[0]+w[1]x[1]+…. ω [p]x[p]+β

where x[0] to x[p] are features representing a single data point with p number of features, ω and β are parameters of the model that are learned while ŷ is the predicted output. For a dataset with a single feature the relationship between the target y and feature x can be expressed using the equation of a line , *y=*mx+c, where m and c are the slopes and intercept on the y axis. Respectively. However, for more features, ω contains the slopes along each feature axis then the ŷ can then be expressed as the weighted sum of the individual input features, with the weights given by the entries of ω.

Linear regression is often referred to as Ordinary Least Square defining the smallest values of ω and β that minimizes the mean square error between predictions ŷ and the true regression targets yi in the training set. The mean squared error is thus the sum f the squared differences between the predictions and the true values. The fitting parameters ω and β can be determined using various methods such as gradient-descent method including linear algebra methods that essentially attempts to minimize the error function:

where ŷ is the fitted function and yi is the actual data

**Ridge Regression (L2 Regularization)**

Ridgw regression is a type of linear regression model, as it also uses the same formula as the ordinary least squares. In ridge regression however, the coefficients ω are chosen to also ensure that they are as small as possible in essence all entries of ω should be selected such that they are closest to 0. This means that the features should have has little effect on the target y and still predict well, this phenomenon is what is referred to as regularization. Regularization means explicitly restricting a model to avoid overfitting. Mathematically, the ridge regression penalizes L2 norm of the coefficients ω by adjusting the L2 regularization parameter as shown in the equation below:

when = 0 the equation becomes ordinary least squares for simple linear regression, and at , then tends to 0. In essence, the magnitude of controls the weight, usually between zero and one, of the different parts of the objective. When the L2 regularization is implemented on a dataset, the training score R2 is usually lower that that of a LinearRegression but the test score are higher, this is because high values of reduces overfitting the data. The Ridge model makes a trade-off between the simplicity of the model (near zero coefficients) and its performance on the training set. How much importance the model places on simplicity versus training set performance can be specified by the user, using the alpha parameter.

In general, as alpha increases the coefficients are pushed to zero, which decreases the score of the training set by avoiding overfitting of the datasets but helps to improve generalization, on the other hand decreasing alpha allows the coefficient to be well restricted and becomes similar to the behavior of a simple linear regression.

**Lasso Regression (L1 Regularization**)

Lasso Regularization it also restricts the coefficients and tries to force them to zero. The penalizes the L1 norm of the coefficient vector, or the sum of the absolute values of the coefficients, mathematically the cost or error function is expressed as:

The consequence of L1 regularization is that when using the Lasso, some coefficients are exactly zero. This means some features are entirely ignored by the model. This can be seen as a form of automatic feature selection. Having some coefficients be exactly zero often makes a model easier to interpret and can reveal the most important features of your model.

**Illustration of Linear Regression with L1 and L2 Regularization using Boston Dataset from mglearn Library**

DATA DESCRIPTION

FEATURE NAMES:

['CRIM' 'ZN' 'INDUS' 'CHAS' 'NOX' 'RM' 'AGE' 'DIS' 'RAD' 'TAX' 'PTRATIO'

'B' 'LSTAT']

BOSTON HOUSE PRICES DATASET

Data Set Characteristics:

Number of Instances: 506

Number of Attributes: 13 numeric/categorical predictive. Median Value (attribute 14) is usually the target.

Attribute Information (in order):

* CRIM per capita crime rate by town
* ZN proportion of residential land zoned for lots over 25,000 sq.ft.
* INDUS proportion of non-retail business acres per town
* CHAS Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
* NOX nitric oxides concentration (parts per 10 million)
* RM average number of rooms per dwelling
* AGE proportion of owner-occupied units built prior to 1940
* DIS weighted distances to five Boston employment centres
* RAD index of accessibility to radial highways
* TAX full-value property-tax rate per $10,000

PTRATIO pupil-teacher ratio by town

* B 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
* LSTAT % lower status of the population
* MEDV Median value of owner-occupied homes in $1000's

Missing Attribute Values: None

**The steps involved in analysing the data are as follows:**

1. Understanding the question. This dataset is to predict the median value of homes in several Boston neighborhoods in the 1970s, using the information about the neighborhoods such as crime rate, proximity to the Charles River, highway accessibility and so on. The datasets contains 506 data points, described by 13 features.
2. Evaluation of the data. The data is explored and the data and target variables are obtained from the boston dataset. Using the LinearRegression class from the sklearn.linear\_model library the training data set is fitted with an overall mean score of 0.74 and test set mean score of 0.71, after the same data set is fitted using a the Ridge class for L2 regularization on the data, a training set and test set mean score are 0.73 and 0.75 respectively. The result is expected , the regularization shows a less overfitted training set and an improvement in the test test compared to the linear regression.
3. Step 3: After 100 trails of different combination of the dataset:

* Top predictor which has the most influence on the result and accuracy of the test is the nitric oxides concentration.
* The test accuracy after implementing the ridge regression is around 69%, which shows that the model doesn’t really fit the data and also means that the 13 features may not be sufficient in making reliable predictions. Suggestions may be that the features may need to be engineered for more accurate predictions.
* As the alpha parameter increases the testing accuracy decreases, the highest test accuracy (regression) of 0.687081 is at alpha of 0.1 and less.
* Compared to the linear regression the testing and training accuracy using the Ridge analysis is quite close to the Linear regression, which means that when the data is not too much the prediction using ridge analysis might be clos to a linear regression

**References**

*Belsley, Kuh & Welsch, 'Regression diagnostics: Identifying Influential Data and Sources of Collinearity', Wiley, 1980. 244-261.*

*Quinlan,R. (1993). Combining Instance-Based and Model-Based Learning. In Proceedings on the Tenth International Conference of Machine Learning, 236-243, University of Massachusetts, Amherst. Morgan Kaufmann.*