Ridge regression

Ridge regression is a type of linear regression that is used to deal with multicollinearity in data. Multicollinearity occurs when two or more independent variables in a dataset are highly correlated with each other. This can lead to unstable and unreliable estimates of the coefficients in a linear regression model.

Ridge regression works by adding a penalty term to the standard linear regression cost function, which is designed to shrink the coefficients of the independent variables towards zero. This penalty term is controlled by a hyperparameter called lambda, which determines the amount of regularization applied to the model. As lambda increases, the coefficients of the independent variables are shrunk towards zero, which helps to reduce the impact of multicollinearity on the model.

Ridge regression is particularly useful in situations where there are many independent variables in the data, or when the independent variables are highly correlated with each other. It is often used in fields such as finance, economics, and social sciences, where multicollinearity is common.

Let's take a practical example to understand how Ridge regression works. Consider a dataset of housing prices, where the independent variables are the square footage of the house, the number of bedrooms, and the number of bathrooms. If these variables are highly correlated with each other, the standard linear regression model may produce unstable and unreliable estimates of the coefficients. This is where Ridge regression comes in.

Using Ridge regression, we can add a penalty term to the cost function, which will help to shrink the coefficients towards zero. This will help to reduce the impact of multicollinearity on the model, and produce more reliable estimates of the coefficients.

In the code, we first import the necessary libraries, including Pandas for creating the dataframe, NumPy for working with numerical data, Matplotlib for visualizing the data, and Scikit-learn's Ridge class for creating the Ridge regression model.

We then create a dataframe with the data for the independent variables (square footage, number of bedrooms, and number of bathrooms) and the dependent variable (housing prices).

Next, we split the data into training and testing sets, using 80% of the data for training and 20% for testing. We then create the Ridge regression model using the Ridge class and fit the model to the training data.

Finally, we use the model to predict the housing prices for the testing data, and calculate the mean squared error (MSE) to evaluate the performance of the model. A lower MSE indicates better performance.