**What is numpy?**

NumPy is mostly used for working with Numerical values as it makes it easy to apply mathematical functions.

**When is it used?**

NumPy can be used to perform a wide variety of mathematical operations on arrays.

**What is pandas?**

It has functions for analyzing, cleaning, exploring, and manipulating data.

**When is it used?**

used for data analysis and associated manipulation of tabular data in DataFrames. Pandas allows importing data from various file formats such as comma-separated values, JSON, Parquet, SQL database tables or queries, and Microsoft Excel.

**What is sci kit learn?**

sklearn is a python library to implement machine learning models and statistical modelling. Through scikit-learn, we can implement various machine learning models for regression, classification, clustering, and statistical tools for analyzing these models.

**What is matplot lib?**

Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python.

**What is seaborn?**

Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.

**What is linear regression?**

Multiple linear regression is a statistical technique used to analyze the relationship between two or more independent variables (also known as predictors) and a dependent variable. The formula for multiple linear regression can be expressed as follows:

Y = b0 + b1X1 + b2X2 + ... + bnxn

where:

Y = the dependent variable (the variable we are trying to predict)

b0 = the intercept (the point where the regression line crosses the Y-axis)

b1, b2, ..., bn = the regression coefficients (the change in Y for a unit change in X1, X2, ..., Xn)

X1, X2, ..., Xn = the independent variables (predictors)

To calculate the values of the regression coefficients (b1, b2, ..., bn), the method of least squares is used. This involves finding the line of best fit that minimizes the sum of the squared differences between the observed values of the dependent variable and the predicted values based on the independent variables.

Once the regression coefficients are calculated, the equation can be used to predict the value of the dependent variable (Y) for a given set of values of the independent variables (X1, X2, ..., Xn).

Overall, multiple linear regression is a powerful tool for modeling the relationship between multiple predictors and a response variable, and can be useful in a wide variety of fields including finance, marketing, and social sciences.

**When is it used?**

Multiple linear regression is used when you want to understand the relationship between a dependent variable and two or more independent variables. It is commonly used in many fields, including:

Economics: To analyze the impact of various economic factors, such as inflation, interest rates, and GDP, on a particular variable, such as consumer spending.

Marketing: To identify the key factors that drive consumer behavior, such as product features, pricing, and promotion, and to develop predictive models for sales and market share.

Finance: To forecast stock prices or other financial metrics based on a variety of factors, such as earnings, interest rates, and market trends.

Social sciences: To study the relationship between various social factors, such as education, income, and ethnicity, and outcomes such as health, crime, and voting behavior.

In general, multiple linear regression is useful in any situation where you want to understand the relationship between multiple variables and predict the value of a dependent variable based on the values of several independent variables.

**What is logistic regression?**

Multiple logistic regression is a statistical technique used to model the relationship between a categorical dependent variable and one or more independent variables. It is an extension of simple logistic regression, which models the relationship between a binary dependent variable and a single independent variable.

The formula for multiple logistic regression is as follows:

log(p/1-p) = β0 + β1X1 + β2X2 + ... + βkXk

where:

p is the probability of the dependent variable (Y) taking on the value of 1 (success)

X1, X2, ..., Xk are the independent variables

β0 is the intercept term (also known as the constant)

β1, β2, ..., βk are the coefficients that measure the change in the log odds of the dependent variable associated with a one-unit change in the corresponding independent variable

To calculate the predicted probability (p), we use the logistic function:

p = 1 / (1 + e^(-z))

where z = β0 + β1X1 + β2X2 + ... + βkXk

The logistic function ensures that the predicted probabilities (p) range between 0 and 1, which is necessary for modeling a binary dependent variable.

In essence, multiple logistic regression estimates the relationship between the independent variables and the dependent variable, and expresses it in terms of the log odds of the dependent variable. The log odds are then transformed into probabilities using the logistic function, which allows us to make predictions about the probability of the dependent variable taking on a particular value.

**When is it used?**

Multiple logistic regression is used when you want to analyze the relationship between a binary dependent variable and two or more independent variables. A binary dependent variable is one that can take only two values, such as "yes" or "no", "success" or "failure", or "alive" or "dead".

Multiple logistic regression is commonly used in many fields, including:

Medicine: To study the relationship between various risk factors, such as age, gender, and lifestyle, and the probability of developing a particular disease.

Marketing: To predict the probability of a consumer making a purchase based on various factors such as product features, pricing, and promotion.

Finance: To analyze the probability of default on a loan or the probability of a stock price rising or falling based on various factors such as financial ratios and market trends.

Social sciences: To study the relationship between various social factors, such as age, gender, and income, and outcomes such as voting behavior or the likelihood of participating in a particular activity.

Overall, multiple logistic regression is useful in any situation where you want to understand the relationship between multiple variables and predict the probability of a binary outcome based on the values of several independent variables.

**What is lasso?**

Multiple Lasso regression is a statistical technique used to model the relationship between a dependent variable and multiple independent variables while also performing variable selection. The Lasso regression method involves adding a penalty term to the ordinary least squares (OLS) regression equation, which helps to shrink the coefficients of the independent variables towards zero, effectively reducing the complexity of the model.

The formula for multiple Lasso regression can be expressed as:

Y = β0 + β1X1 + β2X2 + ... + βp\*Xp + ε

Where:

Y is the dependent variable

X1, X2, ... Xp are the independent variables

β0, β1, β2, ... βp are the regression coefficients

ε is the error term

The Lasso regression penalty term is added to this formula as follows:

minimize Σ(Y - β0 - Σβj\*Xj)^2 + λΣ|βj|

Where:

Σ(Y - β0 - Σβj\*Xj)^2 is the residual sum of squares

Σ|βj| is the sum of the absolute values of the regression coefficients

λ is the regularization parameter that controls the strength of the penalty term.

The Lasso regression algorithm tries to find the values of βj that minimize this formula subject to the constraint that the sum of the absolute values of the coefficients is less than or equal to a certain value, which is determined by the value of λ.

The Lasso regression algorithm helps to identify the most important independent variables that have a significant impact on the dependent variable while reducing the impact of irrelevant or redundant variables. The value of λ can be chosen using cross-validation techniques to balance between the bias and variance of the model.

**When is it used?**

Lasso regression is used when we want to model the relationship between a dependent variable and multiple independent variables while also performing variable selection. It is particularly useful when we have a large number of potential independent variables and we want to identify the most important ones that have a significant impact on the dependent variable.

For example, let's say we want to predict the price of a house based on various features such as the number of bedrooms, the size of the house, the location, and so on. We might have dozens or even hundreds of potential features to include in our model. However, not all of these features may be relevant or important in predicting the price of a house. In fact, including too many irrelevant or redundant features can lead to overfitting and poor model performance.

In this case, we can use Lasso regression to identify the most important features that have a significant impact on the price of a house, while reducing the impact of irrelevant or redundant features. Lasso regression will help us identify the optimal set of features that provide the best predictive power while also keeping the model simple and easy to interpret.

Another example where Lasso regression can be useful is in the field of genetics. Researchers may have thousands of potential genetic markers to study, but not all of them may be relevant or important in predicting a certain trait or disease. Lasso regression can be used to identify the most important genetic markers that are associated with the trait or disease, while reducing the impact of irrelevant or redundant markers. This can help researchers better understand the underlying genetic mechanisms of the trait or disease and develop more targeted and effective treatments.

**What is ridge?**

Ridge regression is a linear regression method that is used when the data suffers from multicollinearity (i.e., when the independent variables are highly correlated with each other). It adds a penalty term to the sum of squared residuals to address this issue and improve the model's stability and accuracy.

The ridge regression formula for a single independent variable is:

y = b0 + b1\*x1 + e

Where:

y is the dependent variable

x1 is the independent variable

b0 and b1 are the regression coefficients (intercept and slope, respectively)

e is the error term

The ridge regression formula for multiple independent variables is:

y = b0 + b1x1 + b2x2 + ... + bp\*xp + e

Where:

y is the dependent variable

x1, x2, ..., xp are the independent variables

b0, b1, b2, ..., bp are the regression coefficients (intercept and slopes, respectively)

e is the error term

The penalty term, also known as the regularization parameter, is added to the sum of squared residuals, and the objective is to minimize the following cost function:

Cost = RSS + λ \* ∑(bi^2)

Where:

RSS is the residual sum of squares (the sum of the squared differences between the predicted and actual values)

λ is the regularization parameter that controls the strength of the penalty term

bi is the ith regression coefficient

The penalty term, which is the sum of squared regression coefficients, shrinks the estimates of the regression coefficients towards zero, and as a result, reduces the model's complexity and variance. The value of λ is chosen through cross-validation to balance the bias-variance trade-off and to achieve the best performance of the model.

**When is it used?**

Ridge regression is typically used in situations where you have a lot of independent variables that are highly correlated with each other, which can cause instability or overfitting in a traditional linear regression model. By adding a penalty term to the regression coefficients, ridge regression can help to reduce the impact of multicollinearity and improve the model's performance.

For example, let's say you are building a model to predict the price of a house based on its various features such as size, number of rooms, location, and age. In this case, the independent variables may be highly correlated with each other, such as size and number of rooms. This could cause issues in a traditional linear regression model, as the regression coefficients may be unreliable or highly variable due to the multicollinearity.

In this case, ridge regression could be used to improve the stability and accuracy of the model. By adding a penalty term to the sum of squared regression coefficients, the model can adjust the coefficients to account for the multicollinearity, and reduce the risk of overfitting or unstable estimates.

**How are ridge lasso different?**

Ridge regression and Lasso regression are both linear regression methods that are used to address the issue of multicollinearity when building a model with multiple independent variables. However, they differ in their approach to regularization and the type of penalty term they apply.

Ridge regression adds a penalty term to the sum of squared regression coefficients, and the goal is to minimize the sum of squared residuals plus the penalty term. The penalty term is a function of the square of the coefficients, and it shrinks the estimates of the coefficients towards zero, but it does not force any of them to be exactly zero. Ridge regression is useful when there are many variables that are correlated with each other, and all of them might be important in predicting the outcome.

On the other hand, Lasso regression adds a penalty term to the sum of the absolute values of the regression coefficients, and the goal is to minimize the sum of squared residuals plus the penalty term. The penalty term is a function of the absolute value of the coefficients, and it can force some of the coefficients to be exactly zero, effectively performing variable selection. Lasso regression is useful when there are many variables, but only a subset of them are likely to be important in predicting the outcome.

So, in simple terms, the main difference between ridge and lasso regression is that ridge regression is better when you have many variables that are correlated with each other, and all of them might be important, while lasso regression is better when you have many variables, but only a subset of them are likely to be important, and the others can be removed by setting their coefficients to zero.

**What is elasticnet?**

Elastic Net is a regularization technique used in linear regression models to prevent overfitting by adding a penalty term to the loss function. The penalty term is a combination of both the L1 (Lasso) and L2 (Ridge) regularization terms.

The Elastic Net loss function is defined as follows:

L = RSS + λ[(1 - α) \* |β|\_1 + α \* |β|\_2^2]

where,

RSS: Residual Sum of Squares

λ: Regularization parameter

β: Regression coefficients

α: Elastic Net mixing parameter (0 ≤ α ≤ 1)

The Elastic Net mixing parameter α controls the balance between the L1 and L2 regularization terms. When α=0, the Elastic Net reduces to Lasso regularization, and when α=1, it reduces to Ridge regularization.

In the above formula, the L1 regularization term, represented as |β|\_1, is the absolute value of the regression coefficients summed up. This term promotes sparse solutions by shrinking the less important features to zero.

The L2 regularization term, represented as |β|\_2^2, is the squared sum of the regression coefficients. This term prevents overfitting by shrinking the regression coefficients towards zero.

Therefore, the Elastic Net regularization term balances the advantages of both L1 and L2 regularization, making it suitable for high-dimensional datasets with correlated features.

In summary, the Elastic Net regularization can be represented as:

Elastic Net = L1 regularization + L2 regularization

L1 regularization promotes sparsity, while L2 regularization prevents overfitting. By combining them, Elastic Net offers a compromise between the two regularization techniques.

**When is it used?**

Elastic Net regularization is commonly used when dealing with datasets that have a large number of features or predictors, especially when these features are correlated. It is a popular technique for feature selection and is often used in situations where there is a possibility of multicollinearity, or when it is necessary to identify a smaller set of relevant features that contribute most to the model's performance.

For example, in genetics, there may be thousands of genes that could potentially contribute to a disease. Elastic Net can be used to identify the subset of genes that are most important in predicting the disease.

Another example is in finance, where there are many predictors such as interest rates, stock prices, and economic indicators that can affect the outcome of a particular investment. Elastic Net can be used to identify the most important predictors in predicting the returns of the investment.

Overall, Elastic Net regularization is a powerful tool for building robust and accurate predictive models that can handle high-dimensional datasets with correlated features.

**Metrics?**

Notebook bagha.