Basic linear model

Linear regression models are used to show or predict the relationship between two variables or factors.

What is a basic linear model used for?

Linear regression models are used to show or predict the relationship between two variables or factors.
.

Advantages of linear basic model

- Linear regression performs exceptionally well for linearly separable data

- Easier to implement, interpret and efficient to train

- It handles overfitting pretty well using dimensionality reduction techniques, regularization, and cross-validation

- One more advantage is the extrapolation beyond a specific data set

Disadvantages of Basic linear model

- The assumption of linearity between dependent and independent variables

- It is often quite prone to noise and overfitting

- Linear regression is quite sensitive to outliers

- It is prone to multicollinearity

Ridge regression model

Ridge regression is a model tuning method that is used to analyze any data that suffers from multicollinearity.

What is ridge regression model used for?

Ridge regression is a technique used to eliminate multicollinearity in data models

Advantages

- Avoids overfitting a model.

- They do not require unbiased estimators.

- They add just enough bias to make the estimates reasonably reliable approximations to true population values.

- They still perform well in cases of a large multivariate data with the number of predictors (p) larger than the number of observations (n).

- The ridge estimator is preferably good at improving the least-squares estimate when there is multicollinearity**.**

Disadvantages

-They include all the predictors in the final model.

-They are unable to perform feature selection.

-They shrink the coefficients towards zero.

- They trade the variance for bias

Lasso regression model

The “LASSO” stands for Least Absolute Shrinkage and Selection Operator. Lasso regression is a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the resulting statistical mode.

What is LASSO used for

It is used over regression methods for a more accurate prediction. This model uses shrinkage. Shrinkage is where data values are shrunk towards a central point as the mean.

Advantages

advantages in using LASSO method, first of all it can provide a very good prediction accuracy, because shrinking and removing the coefficients can reduce variance without a substantial increase of the bias, this is especially useful when you have a small number of observations and a large number of features.

LASSO helps to increase the model interpretability by eliminating irrelevant variables that are not associated with the response variable, this way also overfitting is reduced. This is the point where we are more interested in because in this paper the focus is on the feature selection task

Disadvantage of LASSO:

LASSO selects at most n variables before it saturates. LASSO can not do group selection. If there is a group of variables among which the pairwise correlations are very high, then the LASSO tends to arbitrarily select only one variable from the group

Disadvantage of LASSO:

LASSO selects at most n variables before it saturates. LASSO can not do group selection. If there is a group of variables among which the pairwise correlations are very high, then the LASSO tends to arbitrarily select only one variable from the groupAdvantagesadvantages in using LASSO method, first of all it can provide a very good prediction accuracy, because shrinking and removing the coefficients can reduce variance without a substantial increase of the bias, this is especially useful when you have a small number of observation and a large number of features.

LASSO helps to increase the model interpretability by eliminating irrelevant variables that are not associated with the response variable, this way also overfitting is reduced. This is the point where we are more interested in because in this paper the focus is on the feature selection task

Decision Trees

Decision Trees are a non-parametric supervised learning method used for both classification and regression tasks. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features.

What are decision trees used for

-Decision trees are used for handling non-linear data sets effectively

Advantages:

- Compared to other algorithms decision trees requires less effort for data preparation during pre-processing.

- A decision tree does not require normalization of data.

- A decision tree does not require scaling of data as well.

- Missing values in the data also do NOT affect the process of building a decision tree to any considerable extent.

- A Decision tree model is very intuitive and easy to explain to technical teams as well as stakeholders.

Disadvantages:

- A small change in the data can cause a large change in the structure of the decision tree causing instability.

- For a Decision tree sometimes calculation can go far more complex compared to other algorithms.

- Decision tree often involves higher time to train the model.

- Decision tree training is relatively expensive as the complexity and time has taken are more.

- The Decision Tree algorithm is inadequate for applying regression and predicting continuous values.

Extreme Gradient boosting

XGBoost stands for Extreme Gradient Boosting;

it is a specific implementation of the Gradient Boosting method which uses more accurate approximations to find the best tree model.

What is it used for

- XGBoost is a library used for developing fast and high performance gradient boosting tree models.

Advantages:

- Often provides predictive accuracy that cannot be trumped.

- Lots of flexibility - can optimize on different loss functions and provides - several hyper parameter tuning options that make the function fit very flexible.

- No data pre-processing required - often works great with categorical and numerical values as is.

- Handles missing data - imputation not required.

Disadvantages

- Gradient Boosting Models will continue improving to minimize all errors. This can overemphasize outliers and cause overfitting.

- Computationally expensive - often require many trees (>1000) which can be time and memory

exhaustive.

- The high flexibility results in many parameters that interact and influence heavily the behavior of the approach (number of iterations, tree depth, regularization parameters, etc.). This requires a large grid search during tuning.

- Less interpretative in nature, although this is easily addressed with various tools.

Random Forest

Random forest is a technique used in modeling predictions and behavior analysis and is built on decision trees. It contains many decision trees that represent a distinct instance of the classification of data input into the random forest.

What is it used for

Random forest algorithm can be used for both classifications and regression task. It provides higher accuracy through cross validation.

Random forest classifier will handle the missing values and maintain the accuracy of a large proportion of data.

Advantages

- It reduces overfitting in decision trees and helps to improve the accuracy

- It is flexible to both classification and regression problems

- It works well with both categorical and continuous values

- It automates missing values present in the data

- Normalising of data is not required as it uses a rule-based approach.

Disadvantages

- It requires much computational power as well as resources as it builds numerous trees to combine their outputs.

- It also requires much time for training as it combines a lot of decision trees to determine the class.

- Due to the ensemble of decision trees, it also suffers interpretability and fails to determine the significance of each variable.

Catboost model

-CatBoost is an open-source gradient boosting on decision trees library with categorical features support out of the box, successor of the MatrixNet algorithm.

-CatBoost means Categorical Boosting because it is designed to work on categorical data flawlessly.

When to Use the CatBoost Algorithm?

There are two types of Data out there **Heterogeneous data** and **Homogeneous data.**

Heterogeneous data: It is any data with high variability of data types and formats. They can be ambiguous and low quality due to missing values, high data redundancy. example: Dataset to predict Credit Score

Homogeneous data: It is dataset made up of the things which are similar to each other, which means that the entire dataset is of same data type and formate. example: Dataset of Images, Video, Sound, Text.

CatBoost Works well on Heterogeneous data : if you have a classification problem with Heterogeneous data then using CatBoost will be the safest thing to do because CatBoost is able to outperform the majority of the Boosting algorithms in the first run.

What is it used for

CatBoost can be used in ranking, recommendation systems and forecasting .

How does catboost work

CatBoost converts categorical values into numbers using various statistics on combinations of categorical features and combinations of categorical and numerical features.

It reduces the need for extensive hyper-parameter tuning and lower the chances of overfitting also which leads to more generalized models.

Advantages:

- **Performance**: CatBoost provides state of the art results and it is competitive with any leading machine learning algorithm on the performance front.

- **Handling Categorical features automatically**: We can use CatBoost without any explicit pre-processing to convert categories into numbers. CatBoost converts categorical values into numbers using various statistics on combinations of categorical features and combinations of categorical and numerical features.

- **Robust:** It reduces the need for extensive hyper-parameter tuning and lower the chances of overfitting also which leads to more generalized models. Although, CatBoost has multiple parameters to tune and it contains parameters like the number of trees, learning rate, regularization, tree depth, fold size, bagging temperature and others.

- Easy-to-use: You can use CatBoost from the command line, using an user-friendly API for both Python and R.

Catboost vs Xgboost

XGBoost cannot handle categorical features by itself, it only accepts numerical values similar to Random Forest. Therefore one has to perform various encodings like label encoding, mean encoding or one-hot encoding before supplying categorical data to XGBoost.

CatBoost distinguishes itself from XGBoost by focusing on optimizing decision trees for categorical variables, or variables whose different values may have no relation with each.

Which is faster to use Catboost vs Xgboost

CatBoost means ‘ categorical ’ boosting. It is quicker to use than, say, XGBoost, because it does not require the use of pre-processing your data, which can take the most amount of time in a typical Data Science model building process.

Linear vs non-linear

A linear regression equation simply sums the terms. While the model must be linear in the parameters, you can raise an independent variable by an exponent to fit a curve. ... While both types of models can fit curvature, **nonlinear** regression is much more flexible in the shapes of the curves that it can fit.