ACTIVIDAD #2

Tipo actividad: Reading comprehension: "5 Regression Algorithms you should know – Introductory Guide!" , multiple choice exercise about the previous reading

6) Socialization key vocabulary: "5 Regression Algorithms you should know – Introductory Guide!"

- Regression: Meaning: A machine learning technique where the model predicts the output as a continuous numerical value.
- Linear Regression: Meaning: An ML algorithm used for supervised learning that predicts a dependent variable based on given independent variables, establishing a linear relationship.
- Decision Tree: Meaning: A model applicable to data with numerical and categorical features, capturing non-linear interactions and making decisions based on a tree structure.
- Support Vector Regression (SVR): Meaning: A regression technique using hyperplanes and a kernel trick to predict real values, known for robustness to outliers and high prediction accuracy.

- Lasso Regression: Meaning: A regression analysis employing shrinkage and variable selection, constraining regression coefficients to shrink towards zero, avoiding overfitting.
- Random Forest Regressor: Meaning: An ensemble of decision trees used for classification and regression, executing by constructing multiple trees and outputting the mean prediction of individual trees for regression tasks.
- Pros and Cons: Meaning: Advantages and disadvantages associated with machine learning algorithms, influencing factors like simplicity, interpretability, robustness, and performance characteristics.

7) Reading comprehension activity #2: "5 Regression Algorithms you should know – Introductory Guide!"

Reading: "5 Regression Algorithms you should know – Introductory Guide!"

In Machine Learning, we use various kinds of algorithms to allow machines to learn the relationships within the data provided and make predictions based on patterns or rules identified from the dataset. So, regression is a machine learning technique where the model predicts the output as a continuous numerical value.

Regression analysis is often used in finance, investing, and others, and finds out the relationship between a single dependent variable(target variable) dependent on several independent ones. For example, predicting house price, stock market or salary of an employee, etc are the most common regression problems.

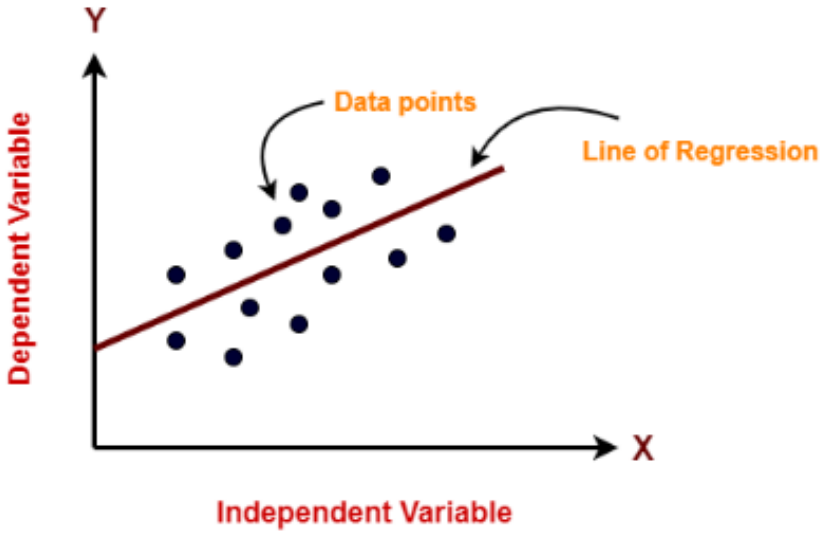
List of Top 5 Regression Algorithms

Here is a list of top 5 regression algorithms

- Linear Regression
- Decision Tree
- Support Vector Regression
- Lasso Regression
- Random Forest

Linear regression

Linear Regression is an ML algorithm used for supervised learning. Linear regression performs the task to predict a dependent variable(target) based on the given independent variable(s). So, this regression technique finds out a linear relationship between a dependent variable and the other given independent variables. Hence, the name of this algorithm is Linear Regression.



In the figure above, on X-axis is the independent variable and on Y-axis is the output. The regression line is the best fit line for a model. And our main objective in this algorithm is to find this best fit line.

Pros:

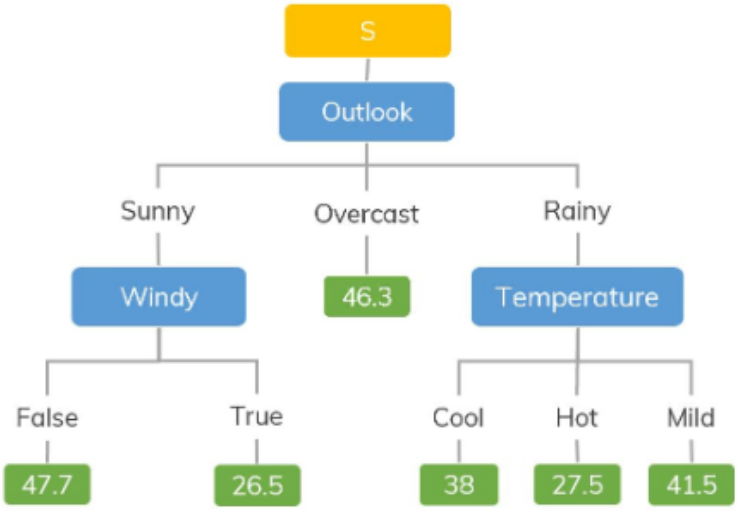
- Linear Regression is simple to implement.
- Less complexity compared to other algorithms.
- Linear Regression may lead to overfitting but it can be avoided using some dimensionality reduction techniques, regularization techniques, and cross-validation.

Cons:

- Outliers affect this algorithm badly.
- It over-simplifies real-world problems by assuming a linear relationship among the variables, hence not recommended for practical use-cases.

Decision Tree

The decision tree models can be applied to all those data which contains numerical features and categorical features. Decision trees are good at capturing non-linear interaction between the features and the target variable. Decision trees somewhat match human-level thinking so it's very intuitive to understand the data.



For example, if we are classifying how many hours a kid plays in particular weather then the decision tree looks somewhat like this above in the image.

So, in short, a decision tree is a tree where each node represents a feature, each branch represents a decision, and each leaf represents an outcome(numerical value for regression).

Pros:

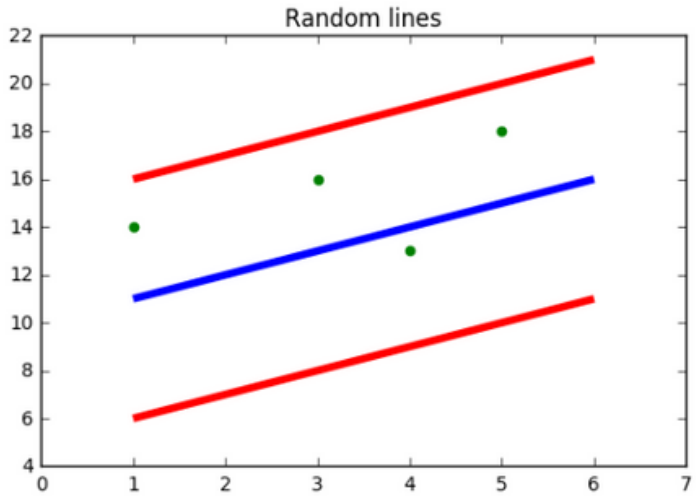
- Easy to understand and interpret, visually intuitive.
- It can work with numerical and categorical features.
- Requires little data preprocessing: no need for one-hot encoding, dummy variables, etc.

Cons:

- It tends to overfit.
- A small change in the data tends to cause a big difference in the tree structure, which causes instability.

Support Vector Regression

You must have heard about SVM i.e., Support Vector Machine. SVR also uses the same idea of SVM but here it tries to predict the real values. This algorithm uses hyperplanes to segregate the data. In case this separation is not possible then it uses a kernel trick where the dimension is increased and then the data points become separable by a hyperplane.



In the figure above, the Blue line is the HyperPlane; Red Line is the Boundary Line

All the data points are within the boundary line(Red Line). The main objective of SVR is to basically consider the points that are within the boundary line.

Pros:

- Robust to outliers.
- Excellent generalization capability
- High prediction accuracy.

Cons:

- Not suitable for large datasets.
- They do not perform very well when the data set has more noise.

Lasso Regression

LASSO stands for Least Absolute Selection Shrinkage Operator. Shrinkage is basically defined as a constraint on attributes or parameters.
The algorithm operates by finding and applying a constraint on the model attributes that cause regression coefficients for some variables to shrink toward a zero.
Variables with a regression coefficient of zero are excluded from the model.
So, lasso regression analysis is basically a shrinkage and variable selection method and it helps to determine which of the predictors are most important.

Pros:

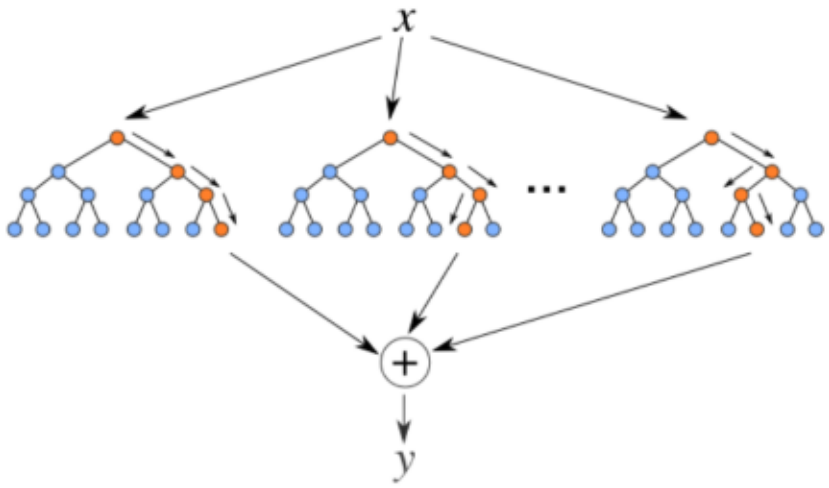
- It avoids overfitting

Cons:

- LASSO will select only one feature from a group of correlated features
- Selected features can be highly biased.

Random Forest Regressor

Random Forests are an ensemble(combination) of decision trees. It is a Supervised Learning algorithm used for classification and regression. The input data is passed through multiple decision trees. It executes by constructing a different number of decision trees at training time and outputting the class that is the mode of the classes (for classification) or mean prediction (for regression) of the individual trees.



Pros:

- Good at learning complex and non-linear relationships.
- Very easy to interpret and understand

Cons:

- They are prone to overfitting
- Using larger random forest ensembles to achieve higher performance slows down their speed and then they also need more memory.

Adapted from: *https://www.analyticsvidhya.com/blog/2021/05/5-regression-algorithms-you-should-know-introductory-guide/*

8) Multiple choice activity.

1.What is the primary goal of regression problems?
A. To predict discrete class labels
B. To estimate a mapping function based on input and output variable
C. To identify outliers in the data
D. To create decision trees

2.Which algorithm is known for its simplicity, ease of implementation, but may lead to overfitting?

A. Decision Tree
B. Support Vector Regression
C. Linear Regression
D. Random Forest

3.What does Lasso Regression aim to achieve?

A. Maximizing overfitting
B. Shrinkage and variable selection
C. Creating decision trees
D. Predicting discrete class labels

4.What is the main objective of Support Vector Regression (SVR)?

A. Identifying outliers
B. Predicting discrete class labels
C. Considering within the boundary line
D. Creating an ensemble of decision trees

5.In the context of Decision Trees, what does each leaf represent?

A. Feature
B. Decisión
C. Outcome (numerical value for regression)
D. Test case for an attribute

6.What is the purpose of Random Forest Regressor in ensemble learning?

A. Predicting discrete class labels
B. Creating linear relationships
C. Constructing a single decision tree
D. Aggregating predictions from multiple decision trees