SACRAMENTO
STATE

CSC 177- Data Analytics and Mining

Project 2: Linear Regression Project & Classification Tree Homework

Team Challengers (23):
1. Srujay Reddy Vangoor
2. Vaibhav Jain
3. Bashar Allwza
4. Varun Bailapudi
5. Uddayankith Chodagam

Linear Regression

Linear Regression

Linear regression is a type of statistical analysis used to predict the relationship between two variables. It assumes a linear relationship between the independent variable and the dependent variable, and aims to find the best-fitting line that describes the relationship.

Linear Regression (Contd.)

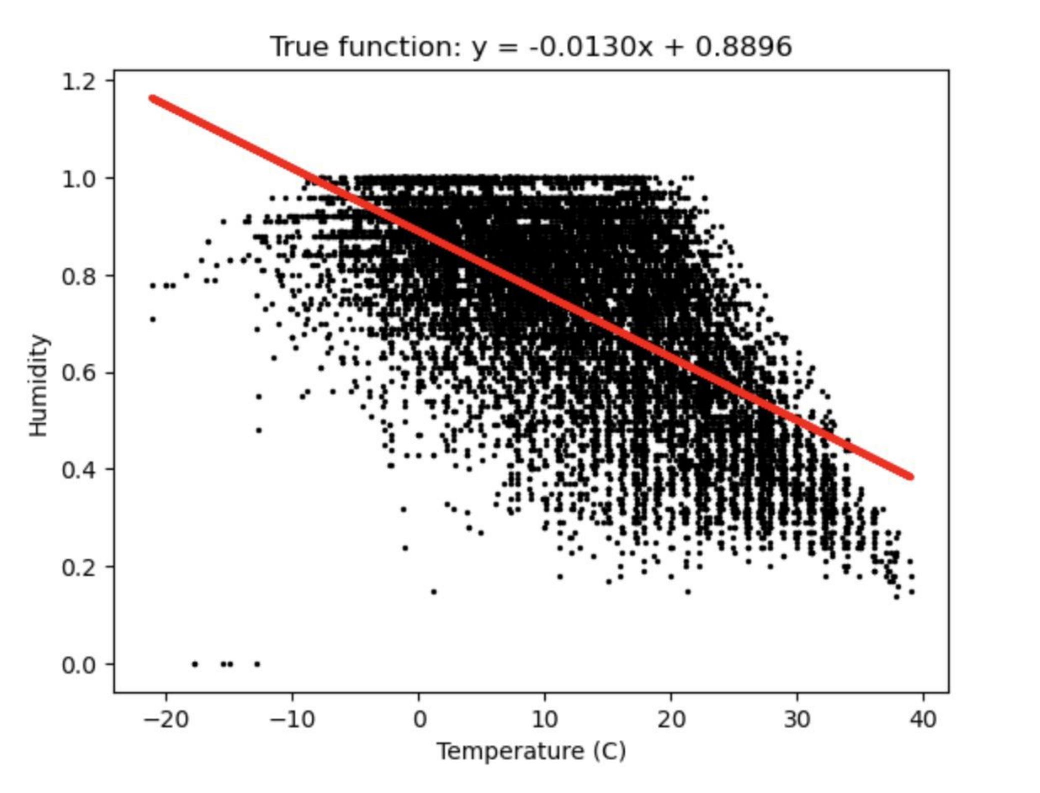
To predict the relationship between two variables, we’ll use a simple linear regression model. In a simple linear regression model, we’ll predict the outcome of a variable known as the dependent variable using only one independent variable.

Steps in linear regression model

To build a linear regression model in python, we’ll follow five steps:

- Reading and understanding the data
- Visualizing the data
- Performing simple linear regression
- Residual analysis
- Predictions on the test set

Linear Regression (Contd.)



Multiple Linear Regression

Multiple Linear Regression is a statistical method used to study the linear relationship between a dependent variable and multiple independent variables. Categorical variables can be handled in multiple linear regression using one-hot encoding or label encoding. The steps to perform multiple linear Regression are almost similar to that of simple linear Regression. The Difference Lies in the evaluation. We can use it to find out which factor has the highest impact on the predicted output and how different variables relate to each other.

Steps in Multiple Linear Regression Model

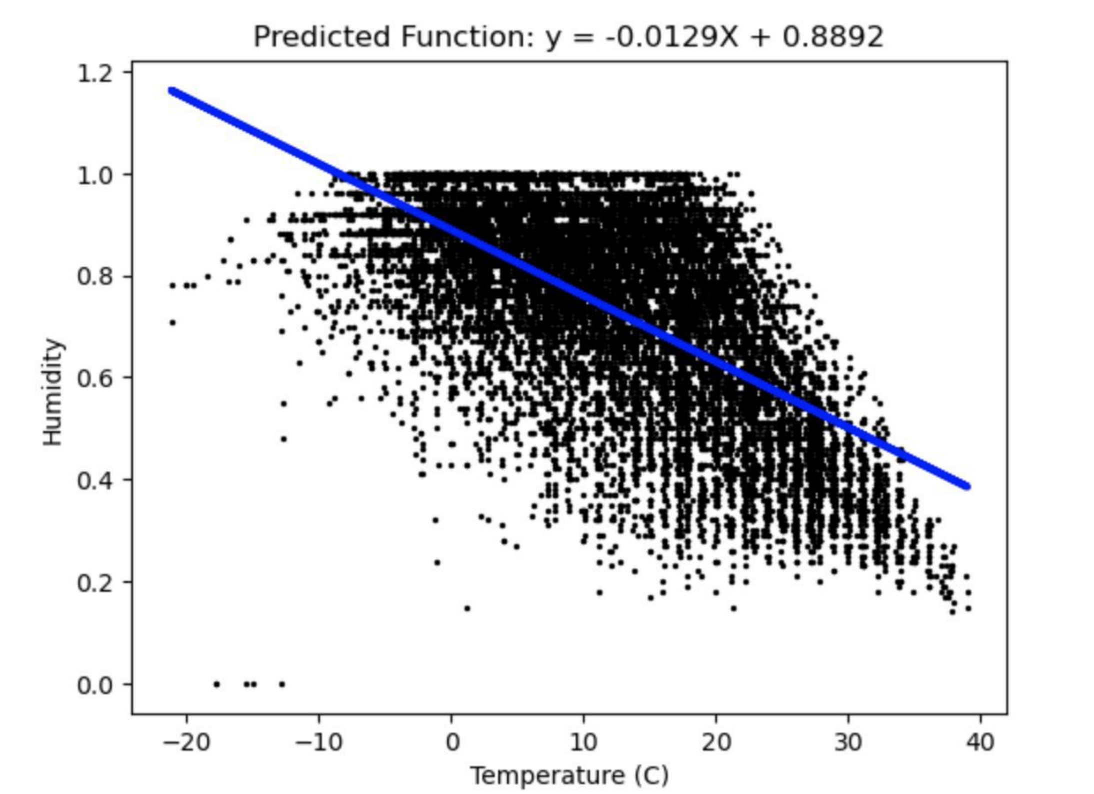
Step 1: Data Pre Processing

1. Importing The Libraries.
2. Importing the Data Set.
3. Encoding the Categorical Data.
4. Avoiding the Dummy Variable Trap.
5. Splitting the Data set into Training Set and Test Set.

Step 2: Fitting Multiple Linear Regression to the Training set

Step 3: Predict the Test set results.

Multiple Linear Regression (Contd.)



Analysis

We find that there is a strong relationship between the temperature and humidity. Based on my findings, As the temperature increases, the humidity of will decrease. Theoretically, this makes sense because as the temperature rises, the amount of moisture in the air will decrease due to evaporation. This is an inverse relationship.

Classification

Classification Tree

A classification tree is a flowchart-like tree structure where an internal node represents a feature(or attribute), the branch represents a decision rule, and each leaf node represents the outcome.

The topmost node in a decision tree is known as the root node. It learns to partition on the basis of the attribute value. It partitions the tree in a recursive manner called recursive partitioning. This flowchart-like structure helps you in decision-making. It's visualization like a flowchart diagram which easily mimics the human level thinking. That is why decision trees are easy to understand and interpret.

Working of a Decision Tree

The basic idea behind any decision tree algorithm is as follows:

1. Select the best attribute using Attribute Selection Measures (ASM) to split the records.

2. Make that attribute a decision node and breaks the dataset into smaller subsets.

3. Start tree building by repeating this process recursively for each child until one of the conditions will match:

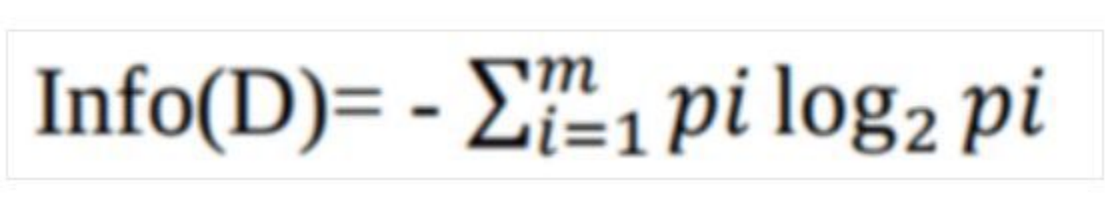
    - All the tuples belong to the same attribute value.
    - There are no more remaining attributes.
    - There are no more instances.

Attribute Selection Measures

Attribute selection measure is a heuristic for selecting the splitting criterion that partitions data in the best possible manner. It is also known as splitting rules because it helps us to determine breakpoints for tuples on a given node. ASM provides a rank to each feature (or attribute) by explaining the given dataset. The best score attribute will be selected as a splitting attribute (Source). In the case of a continuous-valued attribute, split points for branches also need to define. The most popular selection measures are Information Gain, Gain Ratio, and Gini Index.

Information Gain

Claude Shannon invented the concept of entropy, which measures the impurity of the input set. In physics and mathematics, entropy is referred to as the randomness or the impurity in a system. In information theory, it refers to the impurity in a group of examples. Information gain is the decrease in entropy. Information gain computes the difference between entropy before the split and average entropy after the split of the dataset based on given attribute values. ID3 (Iterative Dichotomiser) decision tree algorithm uses information gain.



$$\text{Info}(D) = -\sum_{i=1}^{m} pi \log_2 pi$$

Where Pi is the probability that an arbitrary tuple in D belongs to class Ci.

Gain Ratio

Information gain is biased for the attribute with many outcomes. It means it prefers the attribute with a large number of distinct values. For instance, consider an attribute with a unique identifier, such as customer_ID, that has zero info(D) because of pure partition. This maximizes the information gain and creates useless partitioning.

Thank you!!