Share on your Social Media

Data Science with Machine Learning Tutorial

Published On: August 9, 2024

Data Science with Machine Learning Tutorial

Computers can learn without explicit programming because of machine learning. Since it is a subset of data science, we offer you this thorough data science and machine learning tutorial to help you better comprehend the connection.

Data Science with Machine Learning Tutorial PDF

Introduction to Data Science with Machine Learning

Algorithms are used in machine learning to interpret data and train themselves to make predictions about the future without the need for human interaction. In this data science with machine learning tutorial, we cover the following:

- Overview of Data Science with Machine Learning
- Data Preprocessing
- Understanding of Data Processing
- Understanding of Data Cleaning
- Python Implementation of Data Cleaning
- Understanding of Data Transformation

Featured Articles

Related Courses at SLA

Related Posts

Tableau Developer Salary in Chennai

Published On: October 12, 2024

Introduction A Tableau Developer designs, develops, and maintains dashboards and visualizations using Tableau software. Key...

Overview of Data Science with Machine Learning

The set of guidelines, data, or observations serve as the inputs for machine learning. Facebook, Google, and other corporations employ machine learning extensively. It is important to have expertise in mathematics and statistics to become a data scientist with machine learning skills.

Data Science with Machine Learning Interview Questions

Features of Machine Learning

The following are the features of machine learning:

- Organizations can make better judgments by identifying significant links in the data using this data-driven technology.
- A machine can automatically get better by learning from its historical data.
- It analyzes the data from the given dataset and discovers various patterns.
- Large companies will find it easier to target a consumer base that is relatable if they focus on branding.
- It is comparable to data mining since both involve dealing with enormous volumes of data.

To apply machine learning algorithms for prediction or classification tasks, data is used to identify patterns and correlations between input variables and target outputs.

Usually, data is separated into two categories:

- **Labeled data:** Labeled data has a label or variable that the model is attempting to predict.
- **Unlabeled data:** Unlabeled data lacks a label or target variable.

VMware Tutorial for Cloud Computing Aspirants

Published On: October 12, 2024

VMware Tutorial for Cloud Computing Aspirants VMware software allows you to run a virtual machine...

VBA Macros Tutorial for Beginners

Published On: October 10, 2024

VBA Macros Tutorial for Beginners VBA macros are programs that automate repetitive operations in Microsoft...

VB.Net Tutorial for Beginners

Published On: October 10, 2024

VB.Net Tutorial for Beginners Visual Basic is an easy-to-learn programming language that is type-safe. Get...

Usually, machine learning uses numerical or categorical data. Age and income are two examples of quantities that can be measured and arranged numerically. Values that indicate categories, such as gender or fruit kind, are included in categorical data.

Applications of Data Science with Machine Learning

There are many uses for machine learning, making it an effective tool. These are a few of the most typical applications for machine learning:

Predictive Modeling: By using past data and machine learning, models that forecast future events can be created.

Numerous applications, including fraud detection, weather forecasting, stock market prediction, and consumer behavior prediction, can benefit from this.

Image Recognition: Models that can identify faces, objects, and other patterns in photos can be trained using machine learning.

Numerous applications, including medical image analysis, facial recognition software, and self-driving automobiles, employ this.

Natural Language Processing: Natural language is utilized in numerous applications, including chatbots, voice assistants, and sentiment analysis.

Machine learning can be used to analyze and comprehend natural language.

Recommendation Systems: Recommendation systems that make recommendations to users for goods, services, or content based on their past choices or behavior can be developed using machine learning.

Data Analysis: Big datasets can be analyzed using machine learning to find trends and insights that are hard or impossible for people to find.

Robotics: By using machine learning, robots can be trained to carry out tasks like object manipulation and spatial navigation on their own.

Data Science with Machine Learning Syllabus PDF

Data Preprocessing

An essential phase in the machine learning process is data preprocessing. This stage may involve feature engineering or selection, addressing missing values, and data cleaning and normalization.

It is possible to separate data into testing and training sets. Making sure the data is divided in a representative and random manner is crucial.

Data: Any unprocessed fact, value, text, sound, or image that isn't being decoded and examined is referred to as data.

Information: Data that has been processed, analyzed, and given users the ability to draw meaningful conclusions.

Insight: Insight or knowledge is the culmination of experiences, learning, insights, and inferred facts. leads to the development of consciousness or concepts for a person or organization.

How may data be divided in machine learning?

Training Data: The portion of the data that our model is trained on. This is the real data, both input and output, that your model sees and gains knowledge from.

Validation Data: The portion of data used to fit the model on the training dataset, perform regular evaluations of the model, and make necessary adjustments to the hyperparameters. When the model is training, this data is useful.

Testing Data: Testing data offers an objective assessment after our model has been fully trained. Our model will forecast some values when we feed in the testing data inputs.

Different Forms of Data

Here are the various forms of data:

Numerical Data: A feature is referred to as numeric if it expresses a characteristic that can be quantified.

Categorical Data: Based on some qualitative trait, an attribute can have one of the few, typically fixed, potential values. This is known as a category feature. Nominal features are another name for categorized features.

Ordinal Data: A nominal variable having categories that belong in an ordered list is represented by this. Examples are the small, medium, and large sizes of clothing or a scale from “not bad” to “very satisfied” for client happiness.

Properties of Data

Here are the properties of data:

Volume: With the world’s population expanding and technology becoming more widely used, vast amounts of data are being produced every millisecond.

Variety: Various data formats, including audio clips, videos, photos, and health information.

Velocity: The rate at which data is created and sent.

Value: The significance of the data in terms of what can be deduced by researchers.

Veracity: The assurance and accuracy of the data we are working with.

Viability: The capacity of data to be applied and incorporated into various procedures and systems.

Security: The steps are taken to guard against unauthorized access to or alteration of data.

Accessibility: The simplicity of locating and applying data to make decisions.

Integrity: The precision and comprehensiveness of facts over its whole existence.

Usability: The data's capacity to be easily used and interpreted by end users.

Understanding of Data Processing

The process of transforming data from one form to another so that it can be used more effectively and is more informative is known as data processing.

This entire process can be automated with the use of machine learning algorithms, statistical expertise, and mathematical modeling.

Generally speaking, the primary steps in data processing are as follows:

Data collection: The process of obtaining data from multiple sources, including sensors, databases, and other systems, is known as data collection. The information can be in the form of text, photos, audio, or other formats, and it can be organized or unstructured.

Data preprocessing: This stage entails preparing the data for additional analysis by cleaning, filtering, and manipulating it. This could entail transforming the data to a new format, scaling or normalizing the data, or eliminating missing numbers.

Data analysis: Several methods, including statistical analysis, machine learning algorithms, and data visualization, are used to examine the data in this step. This step's objective is to extract knowledge or insights from the data.

Data interpretation: This stage entails analyzing the data analysis's findings and making inferences from the knowledge acquired. Additionally, it might involve succinctly and clearly presenting the results using dashboards, reports, or other visualizations.

Data storage and management: Following processing and analysis, data needs to be kept safe and organized for easy access. This could entail backing up and recovering the data to prevent loss, as well as storing it in a database, cloud storage, or other systems.

Data reporting and visualization: Lastly, stakeholders are given access to the clearly comprehensible and practical outcomes of the data analysis. This could entail producing dashboards, reports, or visualizations that draw attention to important data trends and discoveries.

Data Scientist with Machine Learning Engineer Salary

Understanding of Data Cleaning

In the machine learning (ML) pipeline, data cleaning is an essential phase that entails finding and eliminating any duplicate, irrelevant, or missing data.

Ensuring that the data is reliable, consistent, and error-free is the aim of data cleaning, since inconsistent or inaccurate data can have a detrimental effect on the performance of the machine learning model.

Why is data cleaning important?

Data cleansing is a crucial step in the preparation of data that guarantees a dataset's reliability, correctness, and overall quality.

- The quality of the underlying data has a major impact on the findings' integrity while making decisions.
- The validity of analytical results can be jeopardized by errors, outliers, missing values, and inconsistencies if data is not properly cleaned.
- Additionally, clean data makes modeling and pattern detection easier because algorithms work best when fed high-quality, error-free input.

Furthermore, well-maintained datasets facilitate the interpretation of results and the development of practical insights.

Steps to Perform Data Cleanliness

Data cleaning is the methodical process of finding and fixing mistakes, inconsistencies, and inaccuracies in a dataset.

To undertake data cleaning, you must take the following crucial actions:

Elimination of Unwanted Observations

- Determine whether observations are redundant or unnecessary, then remove them from the dataset.
- Examining data entries for duplicates, unnecessary information, or data items that don't significantly add to the analysis is part of this process.
- Eliminating superfluous observations simplifies the dataset, decreasing noise and enhancing its general quality.

Correcting Structure Errors

- Take care of any structural problems with the dataset, such as inconsistent variable types, naming conventions, or data formats. Standardize formats, fix mismatched names, and guarantee consistent data representation.

- Correcting structural faults improves consistency of the data and makes correct analysis and interpretation easier.

Handling Unwanted Outliers

- Recognize and handle data points that considerably deviate from the average.
- To reduce the influence of outliers on analysis, choose whether to convert or eliminate them based on the context.
- To extract more accurate and trustworthy insights from the data, controlling outliers is essential.

Managing Missing Data

- Create plans to deal with missing data efficiently.
- This could entail using sophisticated imputation techniques, eliminating records with missing data, or imputing missing values using statistical approaches.
- Managing missing data guarantees a more comprehensive dataset, avoiding biases and preserving the analytical integrity.

Python Implementation for Database Cleaning

Let’s use the Titanic dataset to better understand each step of database cleaning. The required actions are listed below:

- Import the necessary libraries.
- Load the dataset
- Check the data information using df.info()

```
import pandas as pd

import numpy as np

df = pd.read_csv(‘titanic.csv’)

df.head()
```

Examining and Investigating Data

First, let’s examine the data structure to determine any missing values, outliers, or inconsistent patterns. Then, we can use the Python code below to check for duplicate rows.

```
df.duplicated()
```

Examine the data details with df.info().

```
df.info()
```

Examine the numerical and category columns.

```
cat_col = [col for col in df.columns if df[col].dtype == ‘object’]

print(‘Categorical columns :’,cat_col)

num_col = [col for col in df.columns if df[col].dtype != ‘object’]

print(‘Numerical columns :’,num_col)
```

Verify how many unique values there are overall in the categorized columns.

```
df[cat_col].nunique()
```

Elimination of Unwanted Observations

We are aware that the text data is not understood by our machines. Therefore, we must either remove or change the values in the category columns to numerical types.

Since the Name will always be unique and has little bearing on the target variables, we are eliminating the Name columns in this instance.

Let’s print the 50 distinct tickets for the ticket first.

```
df[‘Ticket’].unique()[:50]
```

Drop Name and Columns for Tickets

df1 = df.drop(columns=['Name','Ticket'])

df1.shape

Handling Unwanted Outliers

Let's use df.isnull() to get the percentage of missing values for each row, column by column.

It returns boolean values after verifying if the values are null or not.

The function sum() adds up all the rows with null values, divides that total by all the rows in the dataset, and then multiplies the result to get the values in percentage terms, or how many values are null out of every 100.

*round((df1.isnull().sum()/df1.shape[0])*100,2)*

We are unable to simply delete or disregard the absent observation. They need to be treated with caution since they can be a sign of something significant.

Observations with missing values can be dropped. These are the two most used methods for handling missing data.

- The absence of the value might be instructive in and of itself.
- Furthermore, even when some traits are absent from new data, predictions must frequently be made in the real world!

Use historical observations to impute the missing values.

- Once more, "missingness" is usually always informative in and of itself, and you should report any missing values to your algorithm.
- Not even creating a model to impute your values will add meaningful information. All you're doing is enhancing the patterns that

other features have already supplied.

For this scenario, we can utilize either mean or median imputations.

- When there are no severe outliers and the data is regularly distributed, *mean imputation* is appropriate.
- In cases where the data is skewed or contains outliers, *median imputation* is the better option.

```
df3 = df2.fillna(df2.Age.mean())

df3.isnull().sum()
```

Handling Unwanted Outliers

Extreme values that drastically differ from the bulk of the data are known as outliers. They might have a negative effect on the model's performance and analysis. Outliers can be dealt with using strategies like transformation, interpolation, or grouping.

```
import matplotlib.pyplot as plt

plt.boxplot(df3[‘Age’], vert=False)

plt.ylabel(‘Variable’)

plt.xlabel(‘Age’)

plt.title(‘Box Plot’)

plt.show()
```

The box and whisker graphic above illustrates that there are values in our age dataset that are outliers. Outlier values are those that are fewer than 5 and greater than 55.

```
mean = df3[‘Age’].mean()

std  = df3[‘Age’].std()

lower_bound = mean – std*2

upper_bound = mean + std*2
```

```
print(‘Lower Bound :’,lower_bound)

print(‘Upper Bound :’,upper_bound)

df4 = df3[(df3[‘Age’] >= lower_bound)

          & (df3[‘Age’] <= upper_bound)]
```

Understanding of Data Transformation

Transforming data into a format that is better suited for analysis is known as data transformation. The data can be transformed using methods including encoding, scaling, and normalizing.

Data validation and verification

Data validation and verification entail checking the data with outside sources or expert knowledge to make sure it is accurate and consistent.

First, we distinguish between independent and target features for the machine learning prediction. Here, we’ll just talk about “Sex.”

The only independent qualities that are “Age,” “SibSp,” “Parch,” “Fare,” and “Embarked” are Survived, the goal variables. Because PassengerId has no bearing on survival rates.

```
X = df3[[‘Pclass’,’Sex’,’Age’,
‘SibSp’,’Parch’,’Fare’,’Embarked’]]

Y = df3[‘Survived’]
```

Data Formating

Data formatting is the process of putting the data into a structure or format that is commonly used so that the models or algorithms used for analysis can process it with ease.

There are two techniques, such as scaling and normalization.

Scaling

The process of scaling entails converting feature values into a range. It modifies the scale without altering the original distribution’s shape.

It is helpful in situations when various features have varied scales and certain algorithms are sensitive to feature magnitude.

Standardization (Z-score scaling) and Min-Max scaling are two popular scaling techniques.

Min-Max Scaling: This technique rescales the data to fall into a given range, usually between 0 and 1. It guarantees that the maximum value maps to 1 and the minimum value maps to 0, maintaining the original distribution.

```
from sklearn.preprocessing import MinMaxScaler

scaler = MinMaxScaler(feature_range=(0, 1))

num_col_ = [col for col in X.columns if X[col].dtype != ‘object’]

x1 = X

x1[num_col_] = scaler.fit_transform(x1[num_col_])

x1.head()
```

Z-score scaling, or standardization: It changes the data so that the mean is 0 and the standard deviation is 1.

The data is scaled according to the standard deviation and is centered around the mean.

Data that has been standardized is better suited for algorithms that require characteristics to have zero mean and unit variance or that assume a Gaussian distribution.

$Z = (X - \mu) / \sigma$

X = Data

μ = Mean value of X

σ = Standard deviation of X

Normalization

In machine learning, normalization refers to the process of converting data onto the unit sphere or into the range [0, 1] (or any other range).

- Normalization is a wise strategy to use when you don't know the distribution of your data or when you know it's not Gaussian.
- When your data has different dimensions and the method you're using, like k-nearest neighbors and artificial neural networks, doesn't assume anything about how your data is distributed, normalization can be helpful.

Conclusion

We have discussed various topics in this data science with machine learning tutorial. Join SLA for the best **data science with machine learning training in Chennai.**

Softlogic Systems

KK Nagar [Corporate Office]

No.10, PT Rajan Salai, K.K. Nagar, Chennai – 600 078.
Landmark: Karnataka Bank Building
Phone: +91 86818 84318
Email: enquiry@softlogicsys.in
Map: Google Maps Link

OMR

No. E1-A10, RTS Food Street
92, Rajiv Gandhi Salai (OMR),
Navalur, Chennai - 600 130.
Landmark: Adj. to AGS Cinemas
Phone: +91 89256 88858
Email: info@softlogicsys.in
Map: Google Maps Link

Courses

- Python
- Software Testing
- Full Stack Developer
- Java
- Power BI
- Clinical SAS
- Data Science
- Embedded
- Cloud Computing
- Hardware and Networking
- VBA Macros
- Mobile App Development
- DevOps

- Corporate Training
- Hire With Us
- Job Seekers
- SLA's Recently Placed Students
- Reviews
- Sitemap

Important Links

- Disclaimer
- Privacy Policy
- Terms and Conditions

Social Media Links

f 𝕏 ◎ in ▶

Review Sources

- Google
- Trustpilot
- Glassdoor
- Mouthshut
- Sulekha
- Justdial
- Ambitionbox
- Indeed
- Software Suggest
- Sitejabber