Joseph Mulingwa Kithome

Jmulingwakithome.jmk@gmail.com

Data Science

Practice Questions

**1.What is data science? List the differences between supervised and unsupervised learning.**

Data science is the process of extracting knowledge and insights from data by using scientific methods.

**What Is Supervised Learning?**

If you’re learning a task under supervision, someone is present judging whether you’re getting the right answer. Similarly, in supervised learning, that means having a full set of labeled data while training an algorithm.

**What Is Unsupervised Learning?**

Clean, perfectly labeled datasets aren’t easy to come by. And sometimes, researchers are asking the algorithm questions they don’t know the answer to. That’s where unsupervised learning comes in.In unsupervised learning, a deep learning model is handed a dataset without explicit instructions on what to do with it. The training dataset is a collection of examples without a specific desired outcome or correct answer. The neural network then attempts to automatically find structure in the data by extracting useful features and analyzing its structure.

**Differences**

|  |  |
| --- | --- |
| Supervised Learning | Unsupervised Learning |
| In a supervised learning model, input and output variables will be given. | In unsupervised learning model, only input data will be given |
| Algorithms are trained using labeled data. | Algorithms are used against data which is not labeled |
| Uses Algorithms like Support vector machine, Neural network, Linear and logistics regression, random forest, and Classification trees | Unsupervised algorithms can be divided into different categories: like Cluster algorithms, K-means, Hierarchical clustering, etc. |
| Supervised learning model uses training data to learn a link between the input and the outputs. | Unsupervised learning does not use output data. |
| Learning method takes place offline. | Learning method takes place in real time. |

**2.What is selection Bias?**

**Selection bias** is the bias introduced by the selection of groups, individuals or data for analysis in such a way such that there is no proper randomization achieved. This ensures that the sample obtained is not a representative of the population intended to be analyzed.

**3.What is bias-variance trade-off?**

The term bias-variance tradeoff is commonly used in statistics and machine learning. It is said to be the property of a model that the variance of the parameter estimates across samples can be reduced by increasing the bias in the estimated parameters.

**4.What is confusion matrix?**

confusion matrix is a table that is often used to describe the performance of a classification model mor other a classifier on a set of test data for which the true values are known.

TP+TN/TP+TN+FP+FN

**5.What is the difference between “***Long***” and “***Wide***” format data?**

In wide format, categorical data is always grouped. You can think of it as a summary of long data. It is easier to read and interpret as compared to long format. In long vertical format, every row represents an observation belonging to a particular category.

**6.What is the difference between point estimates and confidence Interval?**

* A point estimate is a single number. Whereas, a confidence interval, naturally, is an interval.
* The point estimate is located exactly in the middle of the confidence interval. However, confidence intervals provide much more information and are preferred when making inferences.