

DATA SCIENCE EAST AFRICA

Mabel Karani



November 17, 2020

mabelkarani55@gmail.com

PRACTICE QUESTIONS

PRACTICE QUESTIONS

1. What is Data Science?

Data science is an inter-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from many structural and unstructured data. Data science is related to data mining, machine learning and big data.

1. Differences between Supervised and Unsupervised Learning

|  |  |  |
| --- | --- | --- |
| **Parameters** | **Supervised machine learning technique** | **Unsupervised machine learning technique** |
| Process | In a supervised learning model, input and output variables will be given. | In unsupervised learning model, only input data will be given |
| Input Data | Algorithms are trained using labeled data. | Algorithms are used against data which is not labeled |
| Algorithms Used | 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. |
| Computational Complexity | Supervised learning is a simpler method. | Unsupervised learning is computationally complex |
| Use of Data | Supervised learning model uses training data to learn a link between the input and the outputs. | Unsupervised learning does not use output data |

1. What is Selection Bias?

Selection bias is the bias introduced by the selection of individuals, groups or data for analysis in such a way that proper randomization is not achieved, thereby ensuring that the sample obtained is not representative of the population intended to be analyzed. It is sometimes referred to as the selection effect.

1. What is Bias- variance trade-off?

Bias is the simplifying assumptions made by the model to make the target function easier to approximate. Variance is the amount that the estimate of the target function will change given different training data. Trade-off is tension between the error introduced by the bias and the variance

1. What is confusion Matrix?

A confusion matrix is a technique for summarizing the performance of a classification algorithm. Classification accuracy alone can be misleading if you have an unequal number of observations in each class or if you have more than two classes in your dataset.

1. 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.

1. 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.

1. How to combat overfitting and underfitting?

**Overfitting** is a modeling error which occurs when a function is too closely fit to a limited set of data points. It is the result of an overly complex model with an excessive number of training points. A model that is overfitted is inaccurate because the model has effectively memorized existing data points.

**Underfitting** is a modeling error which occurs when a function does not fit the data points well enough. It is the result of a simple model with an insufficient number of training points. A model that is under fitted is inaccurate because the trend does not reflect the reality of the data.

Handling Overfitting

* Handling Overfitting:
* Cross-validation
* Regularization
* Early stopping
* Pruning
* Dropout

Handling Underfitting:

* Get more training data.
* Increase the size or number of parameters in the model.
* Increase the complexity of the model.
* Increasing the training time, until cost function is minimised.