DATA SCIENCE

NEWSLETTER VOLUME -8

Datamites
Global Institute for Data Science

What Are The 10 Statistical Techniques That Data Scientist's Need To Master?

Undoubtedly, one can say that "a Data Scientist's statistics knowledge is better than a programmer and programming knowledge is better than a statistician." So, if you are a programmer trying to switch to Data Science career then you need to gain a thorough understanding of statistical theories lying underneath. Remember that your transitions remain incomplete when you blindly utilize the machine learning framework without the base statistical knowledge. This Statistical knowledge helps you to understand the ideas lying underneath various techniques so that you know how and when to use them efficiently.

Statistical learning is a fundamental element of a Data Science training. If you look at quality training providers such as Datamites™, they have structured their Data Science course in such a way that it covers the essential knowledge of Statistics related to Data Science.

So start browsing your High school Statistics book and do not miss to cover on these below mentioned: "10 important Statistical techniques that are essential for a Data Scientist".

Linear Regression: When given two variables namely dependent and independent, Linear Regression is a method followed to predict the target variable after inserting the best linear relationship between the variables. There are two major Linear Regressions and they are Simple Linear Regression and Multiple Linear Regression. In Simple Linear Regression, a single independent variable is used to predict a dependent variable by inserting a best linear relationship whereas a Multiple Linear Regression would prefer to use many independent variables to

predict a dependent variable by inserting a best linear relationship.

Classification: It is actually a Data Mining technique wherein the categories get assigned to a collection of data in order to derive at accurate predictions and analysis. The analysis of very large datasets becomes effective when this method is used. Also called as Decision Tree, it is classified into Logistic Regression and Discriminant Analysis. The Logistic regression is actually a predictive analysis and here, the appropriate regression analysis is conducted when the dependent variable is binary. In the Discriminant analysis, the predictors X are distributed separately in each of the response classes, and then Bayes' theorem is used to turn them into estimates for the probability of the response category given the value of X. The Discriminant analysis can either be linear or quadratic.

Resampling Methods: Resampling method is a non-parametric method of statistical inference wherein the repeated samples are drawn from the original data samples. The utilization of the generic distribution tables to compute approximate p probability values does not happen here. By using experimental methods, it can generate a unique sampling distribution on the basis of the actual data. The concept of Resampling Methods can be well understood when you understand the terms Bootstrapping and Cross-Validation so you don't forget to peep into these concepts too.

Subset Selection: A subset of the p predictors that are believed to be related to the response is identified first, then a model using the least squares of the subset features is fitted. When we consider, Best Subset Selection we use a separate OLS regression for each possible combination of the p predictors and then look at the resulting model that fits. Forward Stepwise Selection starts with a model that contains no predictors, then adds predictors one by one until all of them are in the model. Backward Stepwise Selection starts with all p predictors in the model, iteratively removing the least useful predictor one by one.

Shrinkage: As the name denotes, this approach fits a model that involves all p predictors with the estimated coefficients getting shrunk towards zero relative to the least squares estimates. This Shrinkage process has the effect of reducing variance. Sometimes, depending on the type of shrinkage performed, some of the coefficients are estimated to be exactly zero. The ridge regression and the lasso are the two best-known techniques for shrinking the coefficient estimates towards zero.

Dimension Reduction: The problem of estimating p + 1 coefficients getting reduced to the simple problem of M + 1 coefficients, where M < p, then it is Dimension Reduction. This is achieved by calculating M different linear combinations of the variables which are then used as predictors to insert a linear regression model by least squares. principal component regression and partial least squares are the two approaches used for this task.

Nonlinear Models: When we talk about nonlinear regression, it is a form of regression analysis in which observational data are modeled by a function. This function is expressed as a nonlinear combination of the model parameters with one or more independent variables. The method of successive approximations is used to fit the data here. A couple of important techniques used in Nonlinear Models are a step function, piecewise function, spline and generalized additive model.

Tree-Based Methods: Both the regression and classification problems are solved with Tree-Based Methods. Here, the predictor space is segmented into a number of simple regions with a set of splitting rules which is summarised in a tree. These types of approaches are called as decision-tree methods and it grows multiple trees which are combined to get a single consensus prediction. The important approaches used here are Bagging, Boosting and the random forest algorithm.

Support Vector Machines: Being listed under Machine Learning's supervised learning models, Support Vector Machines is a classification technique that involves finding the hyperplane. In a formal way, you can say that it involves a finding of a hyperplane is a n-1 dimensional subspace of an n-dimensional space that best separates two classes of points with the maximum margin essentially a hard margin. The data points existing on either side supporting the vector is called the "support vectors."

Unsupervised Learning: Until now, we were discussing the techniques that are used on known groups and the experience provided to the algorithm is the relationship between actual entities and the group they belong to. What about the techniques that are being used when the groups are unknown? These are called Unsupervised Learning .

Artificial Intelligence and Data Science Meetup

This is a group for anyone interested in data science and big data related technologies.

If you are a Business Analysts, Data Analyst, Managers, Freshers or Developers aspiring to become data scientist ,You can join this meet up and we will explore the future and scope of data science and Pre requisite and path to become a data scientist.

EVENT NAME: Data Science Meetup

DATE : October 13, 2019

TIMINGS: 10 a.m

VENUE: Bangalore, India.

VIEW MORE

Job Listings

<u>Data Scientist</u>

RadiusAI

Years: 0 - 5 Location: Bangalore

Key Skills: Experience in SQL, Big Data.

Salary: Not Disclosed Posted On: 1 October

Apply

<u>Data Scientist</u>

India_India

Years: 0 - 5 Location: Bangalore

Key Skills: Python, R, Machine learning, Deep learning, JAVA.

Salary: Not Disclosed Posted on: 1 October

Apply

Data Scientist

Miljorens

Years: 0 - 15 Location: Bangalore

Key Skills: Experience in SQL, Big Data.

Salary: Not Disclosed Posted On: 31 September

Apply

Data Scientist

Adobe

Years: 3 - 5 Location: Bangalore

Key Skills: Python, R, Machine learning, Deep learning, JAVA.

Salary: Not Disclosed Posted on: 3 October

Apply

Data Scientist

Flipkart

Years: 0 - 15

Location: Bangalore

Key Skills: Experience in SQL, Big Data.

Salary: Not Disclosed

Posted On: 2 October

Apply

Data Scientist

FIS

Years: 3 - 5

Location: Bangalore

Key Skills: Python, R, Machine learning, Deep learning, JAVA.

Salary: Not Disclosed

Posted on: 30 September

Apply

Data Scientist

Comviva

Years: 3 - 5 Location: Bangalore

Key Skills: Experience in SQL, Big Data.

Salary: Not Disclosed Posted On: 2 October

Apply

Data Scientist

JDA Software

Years: 0 - 15 Location: Bangalore

Key Skills: Python, R, Machine learning, Deep learning, JAVA.

Salary: Not Disclosed Posted on: 1 October

Apply

Data Scientist

Intuit

Years: 4 - 10 Location: Bangalore

Key Skills: Experience in SQL, Big Data.

Salary: Not Disclosed Posted On: 1 October

Apply

Senior Data Scientist

Sabari Corporation

Years: 3 - 5 Location: Bangalore

Key Skills: Python, R, Machine learning, Deep learning, JAVA.

Salary: Not Disclosed Posted on: 31 September

Apply

Data Scientist

Micro Focus

Years: 3 - 5 Location: Bangalore

Key Skills: Experience in SQL, Big Data.

Salary: Not Disclosed Posted On: 28 September

Apply

Data Scientist

HP

Years: 3 - 5 Location: Bangalore

Key Skills: Python, R, Machine learning, Deep learning, JAVA.

Salary: Not Disclosed Posted on: 2 October

Apply

Data Scientist

Goldman Sachs

Years: 0 - 5 Location: Bangalore

Key Skills: Experience in SQL, Big Data.

Salary: Not Disclosed Posted On: 31 September

Apply

Data Scientist

Ericsson

Years: 3 - 5 Location: Bangalore

Key Skills: Python, R, Machine learning, Deep learning, JAVA.

Salary: Not Disclosed Posted on: 2 October

Apply

Datamites

Global Institute for Data Science

☎ :1800-313-3434

Email: info@datamites.com
www.datamites.com

#SwitchToDataScience