SR Cloud Technologies

☎ +91 8897896600 / +91 8897896622

www.srcloudtech.com

sairaghav2013@gmail.com

Course Description

Data Science is the study of the generalizable extraction of knowledge from data. Being a data scientist requires an integrated skill set spanning mathematics, statistics, machine learning, databases and other branches of computer science along with a good understanding of the craft of problem formulation to engineer effective solutions. This course will introduce students to this rapidly growing field and equip them with some of its basic principles and tools as well as its general mindset. Students will learn concepts, techniques and tools they need to deal with various facets of data science practice, including data collection and integration, exploratory data analysis, predictive modeling, descriptive modeling, data product creation, evaluation, and effective communication. The focus in the treatment of these topics will be on breadth, rather than depth, and emphasis will be placed on integration and synthesis of concepts and their application to solving problems. To make the learning contextual, real datasets from a variety of disciplines will be used.

Learning Outcomes

At the end of the course, peers should be able to:

- Describe what Data Science is and the skill sets needed to be a data scientist
- Explain the significance of exploratory data analysis (EDA) in data science. Apply basic tools
- (plots, graphs, summary statistics) to carry out EDA
- Describe the Data Science Process and how its components interact
- Use APIs and other tools to scrap the Web and collect data
- Apply EDA and the Data Science process in a case study
- Use Python to carry out basic statistical modeling and analysis
- Explain in basic terms what Statistical Inference means. Identify probability distributions commonly used as foundations for statistical modeling. Fit a model to data
- Apply basic machine learning algorithms (Linear Regression, k-Nearest Neighbors (k-NN), k-means, Naive Bayes, etc.) for predictive modeling
- Identify common approaches used for Feature Generation. Identify basic Feature Selection algorithms (Decision Trees, Random Forests) and use in applications

- Identify and explain fundamental mathematical and algorithmic ingredients that constitute a Recommendation Engine (dimensionality reduction, singular value decomposition, principal component analysis). Build their own recommendation system using existing components
- Create effective visualization of given data (to communicate or persuade)
- Work effectively in teams on data science projects
- Reason around ethical and privacy issues in data science conduct and apply ethical practices

Audience

The course is suitable for upper-level undergraduate (or graduate) students in computer science, computer engineering, electrical engineering, applied mathematics, business, computational sciences and related analytic fields.

Prerequisites

Students are expected to have basic knowledge of algorithms and reasonable programming experience and some familiarity with basic linear algebra (e.g. solution of linear systems and eigenvalue/vector computation) and basic probability and statistics. If you are interested in taking the course but are not sure if you have the right background, talk to the instructor. You may still be allowed to take the course if you are willing to put in the extra effort to fill in any gaps.

Course Syllabus

1. Introduction to Data Science

2. Introduction to Python Language

3. Basic Concepts (supervised learning, unsupervised learning etc.)

4. Linear Algebra, Probability

5. Naive Bayes

6. Principal Component Analysis

7. Linear Discriminant Analysis

8. Support Vector Machines

9. Linear Regression, Logistic Regression

10. Clustering

11. Feature Selection

12. Recommendation Systems

13. Sentiment Analysis

14. Data Visualization

15. Deep Learning (Neural Networks)

Data Science PROGRAM CURRICULUM

- Introduction to programming in Python
- Python for Data Science (Data Cleansing, Manipulation, and EDA)
- Machine Learning
  - Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it learn for themselves
  - The process of learning begins with observations or data, such as examples, direct experience, or instruction, in order to look for patterns in data and make better decisions in the future based on the examples that we provide. The primary aim is to allow the computers learn automatically without human intervention or assistance and adjust actions accordingly
  - But, using the classic algorithms of machine learning, text is considered as a sequence of keywords; instead, an approach based on semantic analysis mimics the human ability to understand the meaning of a text
- Supervised Learning
  - Supervised learning as the name indicates the presence of a supervisor as a teacher. Basically, supervised learning is a learning in which we teach or train the machine using data which is well labelled that means some data is already tagged with the correct answer. After that, the machine is provided with a new set of examples(data) so that supervised learning algorithm analyses the training data (set of training examples) and produces a correct outcome from labelled data

- Unsupervised Learning
  - Unsupervised learning is the training of machine using information that is neither classified nor labelled and allowing the algorithm to act on that information without guidance. Here the task of machine is to group unsorted information according to similarities, patterns and differences without any prior training of data
  - Unlike supervised learning, no teacher is provided that means no training will be given to the machine. Therefore, machine is restricted to find the hidden structure in unlabelled data by our-self
- Regression
  - A regression problem is when the output variable is a real or continuous value, such as "salary" or "weight". Many different models can be used, the simplest is the linear regression. It tries to fit data with the best hyper-plane which goes through the points
  - For Examples:
    - Which of the following is a regression task?
    - Predicting age of a person
    - Predicting nationality of a person
    - Predicting whether stock price of a company will increase tomorrow
- Classification
  - A classification problem is when the output variable is a category, such as "red" or "blue" or "disease" and "no disease". A classification model attempts to draw some conclusion from observed values. Given one or more inputs a classification model will try to predict the value of one or more outcomes
  - For example, when filtering emails "spam" or "not spam", when looking at transaction data, "fraudulent", or "authorized". In short Classification either predicts categorical class labels or classifies data (construct a model) based on the training set and the values (class labels) in classifying attributes and uses it in classifying new data. There are a number of classification models. Classification models include logistic regression, decision tree, random forest, gradient-boosted tree, multilayer perceptron, one-vs-rest, and Naive Bayes
  - For example:
    - Which of the following is/are classification problem(s)?
    - Predicting the gender of a person by his/her handwriting style
    - Predicting house price based on area
    - Predicting whether monsoon will be normal next year
    - Predict the number of copies a music album will be sold next month
- Clustering
  - It is basically a type of unsupervised learning method. An unsupervised learning method is a method in which we draw references from datasets consisting of input data without labelled responses. Generally, it is used as a process to find meaningful

structure, explanatory underlying processes, generative features, and groupings inherent in a set of examples

   - Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups. It is basically a collection of objects on the basis of similarity and dissimilarity between them

- PCA
   - Principal Component Analysis (PCA) is an unsupervised, non-parametric statistical technique primarily used for dimensionality reduction in machine learning
   - High dimensionality means that the dataset has a large number of features. The primary problem associated with high-dimensionality in the machine learning field is model overfitting, which reduces the ability to generalize beyond the examples in the training set. Richard Bellman described this phenomenon in 1961 as the Curse of Dimensionality where “Many algorithms that work fine in low dimensions become intractable when the input is high-dimensional”
- Decision Trees
   - A decision tree is a flowchart-like structure in which each internal node represents a test on a feature (e.g. whether a coin flip comes up heads or tails), each leaf node represents a class label (decision taken after computing all features) and branches represent conjunctions of features that lead to those class labels. The paths from root to leaf represent classification rules
- Ensemble techniques
   - Ensemble learning is the process by which multiple models, such as classifiers or experts, are strategically generated and combined to solve a particular computational intelligence problem. Ensemble learning is primarily used to improve the (classification, prediction, function approximation, etc.) performance of a model, or reduce the likelihood of an unfortunate selection of a poor one. Other applications of ensemble learning include assigning a confidence to the decision made by the model, selecting optimal (or near optimal) features, data fusion, incremental learning, nonstationary learning and error-correcting
   - **Bagging:** that often considers homogeneous weak learners, learns them independently from each other in parallel and combines them following some kind of deterministic averaging process
   - **Boosting**: that often considers homogeneous weak learners, learns them sequentially in a very adaptative way (a base model depends on the previous ones) and combines them following a deterministic strategy
- Time Series Forecasting
   - Time series forecasting is an important area of machine learning that is often neglected

  - It is important because there are so many prediction problems that involve a time component. These problems are neglected because it is this time component that makes time series problems more difficult to handle

- Text Mining & Sentiment Analysis
  - Text Mining and Sentiment Analysis can provide interesting insights when used to analyse free form text like social media posts, customer reviews, feedback comments, and survey responses. Key phrases extracted from these text sources are useful to identify trends and popular topics and themes. Sentiment scores provide a way to perform quantitative analysis on text data. Combining these techniques, using visually engaging dashboards will help unlock the value of your text data