Data Science Fundamentals – Course Outline

Prerequisites

- School level probability and statistics
- Basic math and linear algebra
- Basic algorithms
- Python programming (candidates should take a MOOC if needed before training – audit at least this course: https://www.edx.org/course/introduction-python-data-science-microsoft-dat208x-2)

1. Overview of Data science [5 Hours]
    a. What is Data science and its history
    b. Data, its potential value, and data usage in different fields for various purposes
    c. Business cases and economic potential and examples
    d. The impact of data size
    e. Introduction to Big data analytics
    f. The role of a data scientist and their impact on this field
    g. Data Science Life Cycle
        i. Planning and logistics (goals and management)
        ii. Data acquisition, preparation and exploration
        iii. Analysis/modeling and Production (presentation/automation)

2. Data Storage and Retrieval [3 Hours]
    a. Storage and retrieval of data
    b. Datasets and features/predictors
    c. Data types and formats
    d. Data sources and data structures – (E.g., data frames, databases – relational and NoSQL) – *NoSQL is wide. A brief introduction might suffice.*

3. Getting started with programming languages, packages, and frameworks for data science [6 Hours]
    a. limit details to Python 3.x : Examples, such as in EDA, will be given using Python
    b. A quick review of Jupyter: setup and packages – ex. Anaconda
    c. A quick review of a well-known data science tools (see also: https://speakerdeck.com/jakevdp/pythons-data-science-stack-jsm-2016)

4. Exploratory Data Analysis [8 Hours]
    a. Types of Data (nominal/categorical, numeric, …)
    b. Summary Statistics – Quantitative
        - Review of statistical and mathematical foundations for data scientists (mean, median, variance, mode, correlations…)

- Review of probability/statistical distributions
- Review of mathematical foundations: multivariable calculus, linear algebra and algorithms (mention briefly)

c. Data visualization and summarization I (2-d charts, maps, infographics, static and dynamic, part II in second course)
- Scatter plots
- Histograms
- Pie charts
- Box-and-whisker plots
- Multi-dimensional graphs

d. Similarity and Dissimilarity

5. Data Engineering [10 Hours]
   a. Data Acquisition
   b. What are different types of data organization (Transactional, relational, structured, unstructured, Graphs, Web data, textual, document based, multimedia, spatial and spatiotemporal, stream and time series data etc.…)
   c. Data integration: Integration of multiple databases, data cubes, or files
   d. Data cleaning: Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies
   e. Data Transformation and Data Discretization: Normalization and Concept of hierarchy generation
   f. Data reduction: Dimensionality reduction, Numerosity reduction, and Data compression

6. Algorithms for applied machine learning and predictive analytics [Predictive analytics examples to be given while illustrating theories] [6 Hours]
   a. Types of Learning
   b. Introduction to Supervised learning
      i. Linear Regression
      ii. Decision Trees (simple visualization and explanation)
   c. Introduction to Unsupervised learning
      i. K-Means
   d. Introduction to neural networks (very basic, promo to part II)
   e. Introduction to Model evaluation
   f. The problem of overfitting

7. Data science ethics I (issues with privacy, safety, security, data ownership, algorithm validity & fairness, legal considerations, more in part II). [2 Hours]
   a. Data privacy
   b. Ownership of data
   c. Security
   d. Legal considerations

8. Simple case studies (more elaborate in part II, indicate business value) [5 Hours]
    a. Recommender systems (such as Netflix)
    b. Marketing (Target)
    c. Social network analysis
    d. Medical data analysis
    e. Financial stock price and inflation
    f. Communication (e.g. Paltel).

9. Projects and presentations [3hours]
    a. Form groups of 3-5 people after 2nd meeting
    b. Help them pick cases and find datasets or provide datasets and ask them to analyze)
    c. Include tasks throughout the course for evaluation
    d. Trainees should apply data engineering fundamentals and apply EDA concepts then create a model and evaluate it then present as a group