IBM Data Analysis Certification

Aug 2020

Lesson Summaries

**Part 1: What is Data Science?**

Week 1

* Defining Data Science
  + Data science is the study of large quantities of data, which can reveal insights that help organizations make strategic choices.
  + There are many paths to a career in data science; most, but not all, involve a little math, a little science, and a lot of curiosity about data.
  + New data scientists need to be curious, judgmental and argumentative.
  + Why data science is considered the sexiest job in the 20th century, paying high salaries for skilled workers in a field with low labor supply and high demand
* What Do Data Scientists Do?
  + The typical work day for a Data Scientist varies depending on what type of project they are working on.
  + Many algorithms are used to bring out insights from data.
  + Accessing algorithms, tools, and data through the Cloud enables Data Scientists to stay up-to-date and collaborate easily.

Week 2

* Big Data and Data Mining
  + Always important to clarify the question that needs to be answered using the data being analyzed
  + How Big Data is defined by the Vs: Velocity, Volume, Variety, Veracity, and Value.
    - Velocity – speed at which data accumulates
    - Volume – scale of data, or increase in amount of data stored
    - Variety – diversity of the data
    - Veracity – quality and origin of data, and its conformity to facts and accuracy
      * Attributes include Consistency, Completeness, Integrity and Ambiguity
    - Value – our ability and need to turn data into value (business/social/medical)
  + How Hadoop and other tools, combined with distributed computing power, are used to handle the demands of Big Data.
  + What skills are required to analyze Big Data.
  + About the process of Data Mining, and how it produces results.
    - Data Mining is the process of automatically searching and analyzing data, discovering previously unrevealed patterns
* Deep Learning and Machine Learning
  + The differences between some common Data Science terms, including Deep Learning and Machine Learning.
  + Machine Learning is a subset of AI that uses computer algorithms to analyze data and make intelligent decisions based on what is learned w/o being programmed
  + Deep Learning is a type of Machine Learning that simulates human decision-making using neural networks.
    - Deep Learning algorithms can label and categorize info and identify patterns
    - Enables AI systems to continuously learn on the job and improve the quality and accuracy of results by determining whether decisions were correct
  + Machine Learning has many applications, from recommender systems that provide relevant choices for customers on commercial websites, to detailed analysis of financial markets.
  + Artificial Neural Networks in AI are a collection of small computing units called neurons that take incoming data and learn to make decisions over time
    - Often layer-deep
    - Main reason that deep learning algorithms become more efficient as the data sets increase in volume
  + Differentiation between AI and Data Science
    - Data science is process of extracting knowledge and insights from large volumes of data. Broad term encompassing entire data processing methodology
    - AI includes everything that allows computers to learn how to solve problems and make intelligent decisions
  + How to use regression to analyze data.

Week 3

* Data Science in Business
  + Data Science helps physicians provide the best treatment for their patients, and helps meteorologists predict the extent of local weather events, and can even help predict natural disasters like earthquakes and tornadoes.
  + That companies can start on their data science journey by capturing data. Once they have data, they can begin analyzing it.
  + Some ways that data is generated by consumers.
  + How businesses like Netflix, Amazon, UPs, Google, and Apple use the data generated by their consumers and employees.
  + The purpose of the final deliverable of a Data Science project is to communicate new information and insights from the data analysis to key decision-makers.
* The Report Structure
  + The length and content of the final report will vary depending on the needs of the project.
  + The structure of the final report for a Data Science project should include a cover page, table of contents, executive summary, detailed contents, acknowledgements, references and appendices.
  + The report should present a thorough analysis of the data and communicate the project findings.

**Part 2: Tools for Data Science**

Week 1

Languages of Data Science

* Introduction to PYTHON
  + Most popular programming language
  + 75% of data scientists have experience
* Introduction to R
  + Most often used by statisticians, mathematicians, etc.
  + Has become the world’s largest repository of statistical knowledge
  + Has stronger object-oriented programming facilities than most statistical computing languages
* Introduction to SQL
  + “Structured Query Language”
  + Much older than Python and R
  + Useful in handling structured data (charts, rows, tables)
    - Designed for use with relational databases
* Other Languages
  + Java – widely adopted general-purpose object-oriented language
  + Scala – general-purpose language that provides support for functional programming
    - Many designs decisions were in criticism to Java
  + Julia
    - Circa 2012
    - Designed for high-performance numerical analysis and computational science
    - COMPILED – meaning the code is executed directly on the processor as executable code, calling on all other data libraries

Data Science Tools

* Data Management – process of persisting and retrieving data
  + Tools. – MySQL, PostgreSQL
* Extract, Transform, and Load (ETL) – process of retrieving data from remote data management systems
* Data Visualization – part of initial data exploration as well as final deliverable
* Model Building – process of creating a machine learning or deep learning model using an appropriate algorithm with a lot of data
* Model Deployment – makes machine learning or deep learning model available to third-party applications
* Model Monitoring and Assessment – ensures continuous performance quality checks on deployed models
* Code Asset Management – uses versioning and other collaborative features to facilitate teamwork
* Integrated Development Environments – tools that help data scientist to implement, execute, test, and deploy their work
* Data Integration and Transformation Tools
  + Apache Kafka, Apache Nifi, Apache AirFlow
* Most widely used open source data management tools are relational databases
* 2 types of visualization tools:
  + Programming Libraries – need to use code
  + Tools – contain a user interface
* Key properties of Jupyter
  + Ability to unify documentation, code, output from the code, shell commands, and visualizations into a single doc
  + Can run R and Python code in addition to other programming languages
* Key property of Apache Spark
  + Linear Scalability – if double the number of servers in a cluster, you’ll also roughly double its performance
* In data management, most of an enterprise’s relevant data is stored in an Oracle Database, Microsoft SQL Server, or IBM Db2
* rStudio – primary choice for development in the R programming language
* SAAS – software as a service
  + Cloud provider operates the product by backing up your data and configuration and installing updates
* IBM Watson Studio and OpenScale
  + Together, they cover the complete development life cycle for all data science, machine learning, and AI tasks
  + Available as a Cloud offering as well as a package running on top of Kubernaetes/RedHat OpenShift in a local data center called IBM Cloud Pak for Data