Data Science Curriculum for Analysts

Recent research by Johns Hopkins covered by The Washington Post has shown that when learning a new skill, switching up what you're doing can help you learn twice as fast! If you're learning to play baseball - change the size and weight of the bat, if you're playing tennis - try different racquets. But how could you use this principle to help people learn to code, learn to do data science?

Data Society developed a data science curriculum where students learn by applying data science methods to different real world data sets from a variety of domains, which speeds up retention and makes the students more versatile data scientists. At completion of this curriculum, students receive a certification from Data Society certifying them as Data Analysts. By the end of this curriculum, students are able to:

1. Think about data as data scientists, clean it and manipulate it
2. Code and format data in R
3. Create intricate and interactive visualizations with R
4. Perform and validate clustering and data mining analyses
5. Structure network problems and analyze network structures to extract insights
6. Mine text data to extract summaries, key topics, opinions and sentiment
7. Build and evaluate ensemble classification models
8. Perform regression and time series analyses to create forecasts

Materials provided:

1. Engaging, animated videos
2. Downloadable PDFs to use as reference materials
3. R code templates from the instructional videos and exercises
4. Data sets used in the instructional videos and exercises

Introduction to Data Science, R and Visualization

By the end of this course, students will be able to:

1. Explain what data science is, how it is used commercially, and which skills a data scientist needs to have
2. Import, export, clean and manipulate data in R
3. Create varied static visualizations with R
4. Create dynamic and appealing visualizations with R

Course curriculum:

1. Getting started with R:
    a) Installing R and RStudio
    b) Introduction to RStudio
2. Overview of data science and R:
    a) What is data science?
    b) A data scientist's approach
    c) Performing basic calculations in R
    d) Loading data into R
3. Working with data in R:
    a) Understanding data types, how and when to use them
    b) Transforming and cleaning data
    c) Selecting and subsetting data
    d) Summarizing and aggregating data
4. Basic visualizations:
    a) Basic plotting in R
    b) Basic plotting in ggplot2
    c) Customizing graphs and adjusting formats
    d) Creating heat maps

Introduction to Data Science, R and Visualization (cont'd)

5. Advanced visualizations:

    a) Advanced plotting in ggplot2, incorporating many variables
    b) 3D visualizations

6. Visualizing data with the Google API:

    a) Accessing the Google API to plot data on maps
    b) Geocoding cities, addresses and landmarks
    c) Creating reusable functions to automate analyses and visualizations
    d) Tips and additional resources

Interactive Visualizations

By the end of this course, students will be able to:

1. Create interactive visualizations, such as maps, charts, and graphs as well as interactive web applications that allow users to manipulate data

2. Publish dynamic visualizations and applications to websites

3. Scrape (collect) and manipulate data from almost any website

1. Dynamic graphs with rCharts:

    a) Introduction to interactive visualization
    b) Why interactive graphs?
    c) Introduction to charts and graphs with rCharts
    d) Publishing your interactive visualizations to the web

2. Interactive Shiny web applications:

    a) Building applications with Shiny – front end (UI)
    b) Building applications with Shiny – back end (server)
    c) Adding pages and functionality to your Shiny application

3. Web scraping, networks, and mapping:

    a) Web scraping in R, collect data from almost any website
    b) Interactive network visualization
    c) Creating interactive maps and varying display aesthetics to bring out topography, buildings, roads and other features

Clustering and Data Mining

By the end of this course, students will be able to:

1. Mine data to find latent patterns and groups in different types of data
2. Evaluate the accuracy and effectiveness of clustering analyses
3. Understand the purpose and implications of what clustering methods can and cannot achieve
4. Identify use cases where clustering analyses are relevant, and where they are not applicable

1. Introduction to clustering:
    a) Commercial applications of data mining
    b) Introduction to clustering, the k-means algorithm used on voting data
2. Implementation of clustering:
    a) k-means clustering on multi-dimensional data
    b) Evaluating the quality of clustering
    c) Determining the right number of clusters to use
3. Clustering multivariate data:
    a) Working with binary data - cosine distance
    b) Clustering binary data - spherical k-means
    c) Assessing quality of spherical k-means clustering
    d) Interpreting clusters of binary data and making recommendations
    e) Pitfalls of clustering
    f) Additional tips and resources

Network Analysis

By the end of this course, students will be able to:

1. Frame a problem from the standpoint of networks and relationships
2. Use R to calculate network measurements and understand their implications
3. Build network diffusion simulations to analyze the spread of messages, computer viruses and disease
4. Measure trust, predict connections and identify communities based on structural features as well as node similarities

1. Introduction to networks:

   a) Commercial applications of network analysis
   b) Geographic networks
   c) Interest networks
   d) Human communities
   e) Graph components: nodes and edges

2. Measuring and visualizing networks:

   a) Centrality measures: degree, closeness, betweenness, eigenvector and PageRank
   b) Geolocation and visualizing geographic networks
   c) Visualizing non-geographic networks
   d) Identifying network weaknesses

3. Network propagation and message diffusion:

   a) SIRS: infection diffusion and cascading failures
   b) Mining Twitter using its API
   c) Interactive network diffusion simulations: animation package

4. Measuring trust and community detection:

   a) Political data and The Sunlight Foundation data
   b) Node similarity - Jaccard distance
   c) Hierarchical clustering for community detection
   d) Modularity-based methods for community detection
   e) Additional tips and resources

Text Mining

By the end of this course, students will be able to:

1. Import, clean and parse various types of text data into R
2. Identify key elements and topics in text data and visualize it
3. Measure sentiment and extract context around key topics in text
4. Classify documents into distinct groups
5. Automatically summarize large bodies of text

1. Introduction to text analysis:
    a) Commercial applications of text mining
    b) Sentence structure and parts of speech
    c) Bag of words, n-grams and word clouds
    d) Cleaning text: capitalization, punctuation and stemming
2. Entity extraction:
    a) Word dictionaries
    b) Principal Component Analysis
    c) Singular Value Decomposition
3. Sentiment analysis and topic models:
    a) Positive vs. negative: degree of sentiment
    b) Item Response Theory
    c) Latent Semantic Allocation
    d) Latent Dirichlet Allocation
4. Document classification:
    a) Jaccard and cosine similarity
    b) TFIDF
    c) Hierarchical clustering
5. Text summarization:
    a) Sentence detection and Luhn's method for text summarization
    b) Additional tips and resources

Classification

By the end of this course, students will be able to:

1. Identify opportunities and use cases for predictive analytics
2. Build classification models to anticipate events and behaviors
3. Evaluate accuracy of predictive algorithms
4. Build ensemble models

1. Introduction to classification and supervised machine learning:
    a) Commercial applications of classification models and predictive analytics
2. Classification algorithms:
    a) k-Nearest Neighbors
    b) Decision trees: gini coefficient and information gain
    c) Random forests
    d) Support vector machines
    e) Logistic regression
    f) Multivariate logistic regression
    g) Penalized logistic regression
    h) Naïve Bayes
    i) Linear discriminant analysis
3. How good is your model:
    a) Confusion matrices and misclassification rates
    b) Base line errors
    c) ROC curves
    d) AUC values
4. Improving your model:
    a) Bagging
    b) Boosting
    c) Ensemble models
    d) Additional tips and resources

Regression and Time-Series Analysis

By the end of this course, students will be able to:

1. Identify opportunities and use cases for regression and time-series models
2. Build single and multivariate regression models
3. Assess statistical significance and validate models for explanatory power and bias
4. Use time-series models to identify seasonality and create forecasts

1. Introduction to regression and time-series analysis:

   a) Commercial applications of forecasting and time-series analysis
   b) Linear relationships: slope, y-intercept, variable interactions
   c) Variance and standard deviation
   d) Covariance and correlation
   e) Normal distribution and bell curves

2. Regression modeling:

   a) Distribution of errors: Q-Q plot, heteroscedasticity
   b) Multivariate regression
   c) R^2 and adjusted R^2
   d) p-values and t-test
   e) F-test and F-distribution
   f) Multicollinearity test
   g) Heteroscedasticity test
   h) Model selection: Akaike Information Criterion
   i) Polynomial regression
   j) Confidence intervals

3. Time-series analysis and seasonality:

   a) Moving averages
   b) Seasonality detection: autocorrelation
   c) Seasonality: additive vs. multiplicative
   d) Decomposing seasonal data: trend, level and seasonality
   e) Multiplicative Holt-Winters exponential smoothing
   f) Forecasting seasonal trends
   g) LOcal regrESSion: LOESS
   h) Additional tips and resources