Data Science Specialization (16 credits)

Courses: 1. Machine Learning 2. Data Mining 3. Big Data and Mining massive Datasets 4. Process Mining and 5. Data Analytics and Visualization

Course 1. Machine Learning: (4 credits)

By the time you finish this course, you'll know how to apply the most advanced machine learning algorithms to such problems as anti-spam, image recognition, clustering, building recommender systems, and many other problems. You'll also know how to select the right algorithm for the right job, as well as become expert at 'debugging' and figuring out how to improve a learning algorithm's performance.

Resources:
1. https://www.coursera.org/learn/machine-learning/ (Modules 1-16)
2. Deep Learning with Python, FRANÇOIS CHOLLET, MANNING, SHELTER ISLAND (Book) (Modules 17-20)

Module 1: Introduction to Machine Learning
Learning Objectives:

- Define What is called Machine Learning.
- Differentiate machine Learning problems in to supervised and unsupervised problems.
- Identify supervised learning problems into "regression" and "classification" problems
- Discuss how unsupervised learning approach problems with little or no idea what our results should look like

Module 2: Linear Regression with One Variable
Learning Objectives:

- Describe what is the regression problem
- Represent a problem as Linear Regression with One Variable
- Describe cost functions to measure the accuracy of a solution to regression problem
- Estimate accuracy by using a cost function of a solution to linear regression problem
- Review the parameters used in cost functions and Gradient Descent method to optimize the cost functions
- Apply Gradient Descent method for parameter learning for linear regression

Module 3: Linear Algebra Review
Learning Objectives:

- Review Linear Algebra (Vectors, Matrices, Inverse matrices, Matrix transpose, Multiplication and Matrix with a vector and with another matrix)
- Represent Linear regression as matrix multiplication and solving simultaneous equations

Module 4: Octave Programming Language
Learning Objectives:

- Setup Programming Assignment Environment
- Practice simple computing problems with octave programming (Basic operations, control statements, moving data around, computing data and plotting data, vectorization)

Module 5: Linear Regression with Multiple Variables
Learning Objectives:

- Extend Linear regression with one variable to Multivariate Linear Regression
- Modify Gradient descent method appropriately to Multivariate Linear Regression

- Describe how multiple features combined in to one
- Generalize Linear regression to a non-linear (polynomial) regression

Module 6: Logistic Regression

Learning Objectives:

- Differentiate Discrete and Continuous outcomes
- Define a classification problem
- Identify issues we get if we use linear regression to solve classification problem
- Change regression hypothesis to suit classification problem by using logistic function
- Relate Decision boundary to predict different classes
- Change cost function appropriately to suit classification
- Apply Logistic regression to classification problems

Module 7: Regularization

Learning Objectives:

- Identify overfitting and underfitting of regression
- Explain how sometimes the solution becomes either overfitting or underfitting
- Discover reasons for over fitting vs number of features
- Apply regularization to handle overfitting (Lasso and Ridge)

Module 8: Neural Networks

Learning Objectives:

- Identify reasons for linear classifiers fail in some cases
- Define a Neural Network model
- Outline how brain works and relate a Neural Network model with it
- Explain a Neural Network model mathematically
- Describe a Neural Network model as a classification model

Module 9: Neural Networks – Learning

Learning Objectives:

- Identify cost function for Neural Network
- Explain Back propagation for training Neural Network models
- Identify issues with initialization of Neural Network models
- Summarize with an algorithm how Neural Network models are built
- Discuss applications Neural Network models for autonomous driving

Module 10: Best Practices of Machine Learning

Learning Objectives:

- Evaluate a hypothesis for machine learning model
- Compare model performance on train and test datasets
- Distinguish whether **bias** or **variance** is the problem contributing to bad predictions
- Apply regularization to reduce variance
- Revise model parameters by using validation set
- Inspect what to do next to improve model performance
- Discover methods to handle skewed datasets
- Discover methods to handle large datasets

Module 11: Support Vector Machines

Learning Objectives:

- Extend optimization objective to large margin classifiers
- Explain mathematical formulation of support vector machines
- Identify non-linear kernels to be used for support vector machines
- Use SVM as classifier model

Module 12: Unsupervised Learning

Learning Objectives:

- Differentiate the input datasets for supervised and unsupervised models

- Explain K-means clustering
- Identify issues in clustering (number of clusters and initialization of cluster centers)
- Use Elbow method to choose number of clusters
- Discover applications of clustering (segmentation, text summarization etc.)

Module 13: Dimensionality Reduction
Learning Objectives:

- Explain Principal Component Analysis
- Discuss choosing the number of principal components
- Analyze How PCA can be used for Data Compression
- Demonstrate hoe PCA can be used for Data visualization
- Explain how PCA speedup learning process

Module 14: Anomaly Detection
Learning Objectives:

- Define Anomalies in a dataset
- Explain statistical concepts to identify anomalies (Gaussian distribution)
- Identify anomalies
- Compare Anomaly detection and supervised learning

Module 15: Recommender Systems
Learning Objectives:

- Discuss companies using Recommender Systems
- Explain how recommendation systems work
- Differentiate Content based and collaborative filtering recommendation systems
- Summarize implementation details of recommendation systems

Module 16: Large scale Machine Learning
Learning Objectives:

- Identify problems with large datasets
- Apply Mini-Batch Gradient Descent for large datasets
- Apply stochastic Gradient Descent for large datasets

Module 17-20: Additional Readings: Deep Learning
Learning Objectives:

- Describe Key factors behind deep learning's rising popularity and future potential
- Recognize what deep learning achieved so far
- Explain what a convolutional neural network is
- Use a pretrained convnet to do feature extraction
- Fine-tune a pretrained convnet
- Visualize what convnets learn and how they make classification decisions
- Explain Recurrent Neural Networks
- Apply Recurrent Neural Networks sequence data such as processing text, time series etc.

Course 2. Data Mining (4 credits)

By the time you finish this course, you'll know how to preprocess data for data mining, explore the data before applying data mining techniques, discover association patterns in a dataset, use advanced predictive modeling techniques to solve problems in real world, Use different advanced clustering techniques to summarize data. You will also analyze the complexity of these algorithms and use them appropriately to solve problems in the real world within available resources
Textbook: Introduction to Data Mining, PANG-NING TAN Michigan State University, MICHAEL STEINBACH University of Minnesota, VIPIN KUMAR University of Minnesota, Pearson/Addison Wesley

Module 1: Introduction
Learning Objectives:

- Explain how Data Mining has evolved
- Differentiate Information Retrieval and Data Mining
- Differentiate Data Mining tasks in to descriptive and predictive
- Discover challenges of finding knowledge from data
- Discover applications of Data Mining in different field

Module 2: Data Preprocessing
Learning Objectives:

- Discuss data-related issues that are important for successful data mining
- Describe different types of data
- Explain general characteristics of datasets
- Use sampling to select subset of data
- Explain the curse of dimensionality
- Discover methods for dimensionality reduction
- Discuss Feature subset selection, feature engineering
- Explain Discretization and Binarization
- Discuss Measures of Similarity and Dissimilarity of Attributes and data objects

Module 3: Exploring Data
Learning Objectives:

- Produce Statistical summary of the data
- Visualize data
- Explain how to visualize higher dimensional data
- Explain different multidimensional analysis technique

Module 4: Association Analysis: Basic Concepts, Apriori Algorithm
Learning Objectives:

- Define Association Analysis problem
- Explain Apriori Principle
- Use Apriori Algorithm to generate frequent Item sets and Rule generation
- Define Maximal frequent Itemset
- Define closed frequent itemset

Module 5: Association Analysis: Algorithms FP-Growth Algorithm
Learning Objectives:

- Explain FP-Growth Algorithm
- Use FP-Growth Algorithm to generate frequent Item sets and Rule generation
- Describe how to evaluate Association Patterns

Module 6: Association Analysis: Handling different types of attributes, Sequential patterns
Learning Objectives:

- Transform the categorical and symmetric binary attributes into "items"
- Describe quantitative association rule
- Explain the three types of methods to generate quantitative association rules
- Discuss concept hierarchy in Itemset
- Use DAG to represent concept hierarchy
- Formulate sequential pattern problem
- Discover sequential patterns in a temporal dataset

Module 7: Association Analysis: Subgraph patterns, Infrequent patterns
Learning Objectives:

- Describe the application of association analysis methods to more complex entities beyond item sets and sequence (chemical compounds, 3-D protein structures, network topologies, and tree structured XML document)
- Explain graph representation of entities in various applications
- Use frequent subgraph mining to derive a set of common substructures among the collection of graphs.
- Define infrequent pattern
- Compare Infrequent Patterns, Negative Patterns, and Negatively Correlated Patterns

Module 8: Classification: Basic Concepts, Decision Trees
Learning Objectives:

- Explain classification task in datamining with examples
- Explain how a classification model can serve as an explanatory tool to distinguish between objects of different classes
- Extend classification model to a predictive model
- Explain how confusion matrix can be used for model evaluation
- Describe how a decision tree works as a classification model
- Use Hunt's algorithm to build d decision tree
- Discover design issues of decision tree induction
- Explain overfitting and its causes
- Describe methods for estimating the generalizationerror of a model during training

Module 9: Classification: Model Evaluation and Comparing Classifiers
Learning Objectives:

- Explain evaluation methods of classifiers
- Evaluate the performance of a classifier by using Holdout method
- Evaluate the performance of a classifier by using Cross Validation method
- Evaluate the performance of a classifier by using Boot Strap method
- Discover Methods for Comparing Classifiers

Module 10: Classification: Nearest-Neighbor classifiers
Learning Objectives:

- Differentiate eager learners and lazy learners
- Explain the steps of Nearest-Neighbor classifier model building
- Describe characteristics of Nearest-Neighbor classifier

Module 11: Classification: Bayesian Classifiers
Learning Objectives:

- Explain the probabilistic relationships between the attribute set and the class variable
- State Bayes' theorem
- Explain how Bayes' theorem can be used for classification
- Describe how Naive Bayes Classifier works
- Describe how Bayesian Belief Network works

Module 12: Classification: Ensemble methods
Learning Objectives:

- Describe techniques ensemble or classifier combination methods for classification.
- Explain rationale for ensemble method
- Describe Methods for Constructing an Ensemble Classifier
- Explain Bagging and Boosting techniques
- Describe Random Forest classifier

Module 13: Classification: Class imbalance problem
Learning Objectives:

- Define Class Imbalance problem
- Discover alternating classifier evaluating methods such as ROC curve and cost sensitive learning
- Explain Sampling-Based Approaches to handle class imbalance problem

Module 14: Cluster Analysis: Basic concepts and K-means algorithm
Learning Objectives:

- Explain how cluster analysis helps for data summarization, data compression, efficiently finding nearest neighbors
- Describe cluster analysis applications to into different fields study
- Define what is cluster analysis
- Explain different types of clustering
- Explain methods of measure quality of clustering
- Explain issues in K-means clustering and how to overcome them
- Discuss strengths and weaknesses of k-means clustering
- Use K-means to find clusters in a dataset

Module 15: Cluster Analysis: Agglomerative Hierarchical Clustering
Learning Objectives:

- Explain Hierarchical Clustering
- Describe different types of Hierarchical Clustering
- Use Hierarchical Clustering to find clusters in a dataset
- Discuss strengths and weaknesses of Hierarchical Clustering

Module 16: Cluster Analysis: DB Scan
Learning Objectives:

- Explain what Density-Based clustering is
- Describe center-based approach for DB Scan
- Explain DB Scan algorithm
- Discuss strengths and weaknesses of DB Scan

Module 17: Cluster Analysis: Cluster evaluation
Learning Objectives:

- Differentiate evaluation methods for classification and cluster models
- Explain cohesion and separation to evaluate supervised clusters
- Define the Silhouette Coefficient
- Explain different methods to evaluate unsupervised clusters
- Compare K-means and DB Scan

Module 18: Self Organizing maps
Learning Objectives:

- Discuss the characteristics of data that can strongly affect cluster analysis
- Explain self-organizing maps (SOM)
- Describe SOM algorithm

Module 19: Graph based clustering
Learning Objectives:

- Explain Graph Based clustering
- Describe MST clustering algorithm
- Describe Chameleon Algorithm

Module 20: Anomaly detection
Learning Objectives:

- Define anomalies in a dataset
- Illustrate some applications for which anomalies are of considerable interest
- Discuss causes of anomalies
- Discuss different techniques of finding anomalies
- Discuss important issues that need to be addressed when dealing with anomalies

Course 3. Mining Massive Datasets (4 credits)

By the time you finish this course, you will apply many of the interesting algorithms that have been developed for efficient processing of large amounts of data in order to extract simple and useful models of that data. You will use these techniques to predict properties of future instances of the same sort of data, or simply to make sense of the data already available. You can explain how some techniques for processing large datasets used in building machine learning. You can also distinguish some algorithms and ideas for dealing with big data that are not usually classified as machine learning.

Resource: https://lagunita.stanford.edu/courses/course-v1:ComputerScience+MMDS+SelfPaced/about

Module 1: MapReduce: Computational Model
Learning Objectives:

- Explain challenges dealing with Big Data and parallel computing
- Describe what is distributed file system
- Explain what cluster computing is
- Describe the value of MapReduce that it allows to write code that can exploit massively parallel computing without one has to think about the parallelism itself.

Module 2: MapReduce: Scheduling, Dataflow, Combiners and Partition functions
Learning Objectives:

- Explain how MapReduce takes over the responsibility for coping with the possibility of hardware or other failures as a long-running job executes.
- Describe MapReduce Environment
- Implement Word Frequency count problem on Hadoop platform by writing MapReduce framework
- Explain how combiners and partition functions refine MapReduce algorithms efficient

Module 3: Page Rank
Learning Objectives:

- Discuss the technique invented by Larry Page for estimating the importance of a Web page
- Explain how Page Rank used in Google to help choose the most useful pages to show in response to a search query
- Represent web as directed graph
- Explain web search challenges
- Describe the Link Analysis approaches for computing importance of nodes in a graph

Module 4: Page Rank – Google Formulation
Learning Objectives:

- Explain flow formulation in directed web graph
- Interpret the flow formulation as Matrix formulation
- Explain issues in ranking nodes in web graph
- Discuss google formulation of page rank
- Explain page rank computation algorithm

Module 5: Locality-Sensitive Hashing

Learning Objectives:

- Explain challenges to find related pairs of items, even in relatively small datasets
- Discuss techniques for dealing with situations where we need to find related pairs of items
- Discuss applications of set-similarity
- Explain three essential techniques to find similar documents
- Explain implementation of Jaccard similarity measure to construct signatures and min-hashing to find similar documents
- Explain LSH for min-hashing signatures to find similar large set of documents

Module 6: Applications of LSH

Learning Objectives:

- Use LSH to Entity Resolution
- Use LSH to Fingerprints comparison
- Use LSH to find Similar News Articles

Module 7: Distance Measures, Nearest-Neighbor Learning

Learning Objectives:

- Differentiate Euclidean and non-Euclidean distance characteristics
- Discuss different distance measures between different data objects
- Use LSH for Nearest-Neighbor Learning for large scale datasets

Module 8: Frequent Item sets

Learning Objectives:

- Discuss applications of Frequent Item sets
- Discuss challenges of finding Frequent Item sets in large datasets
- Extend Apriori algorithm to find Frequent Item sets in large datasets
- Explain PCY algorithm to find Frequent Item sets in large datasets
- Explain SON algorithm to find Frequent Item sets in large datasets
- Explain Toivonen's algorithm to find Frequent Item sets in large datasets

Module 9: Communities in Social Networks

Learning Objectives:

- Discuss communities in a network
- Explain Community Affiliation Graph Model and AGM generative process
- Extend AGM to BigCLAM

Module 10: Communities in Social Networks: Graph Partitioning

Learning Objectives:

- Detect clusters in Large Graphs

- Find densely linked clusters
- Explain what makes a good cluster
- Explain spectral partitioning algorithms

Module 11: Stream Algorithms

Learning Objectives:

- Discuss challenges with streaming data
- Differentiate static data management and stream data management
- Differentiate static queries and ad-hoc queries
- Discuss applications of streaming algorithms
- Explain sliding windows to answer queries
- Describe the solution to counting bits problem

Module 12: Stream Algorithms: Counting distinct elements

Learning Objectives:

- Explain what a bloom filter and its application is to find distinct elements
- Discuss Applications of find distinct elements in different fields
- Discuss different methods to find distinct elements when there is no space to store the complete set?

Module 13: Recommendation Systems: Content Based

Learning Objectives:

- Explain Long Tail problem
- Discuss types of recommendations
- Explain what utility matrix is
- Discuss approaches to Recommender Systems
- Explain content-based Recommender Systems
- Discuss pros and cons of content-based Recommender Systems

Module 14: Recommendation Systems: Collaboration Based

Learning Objectives:

- Explain collaborative filtering-based Recommender Systems
- Discuss pros and cons of collaborative filtering-based Recommender Systems
- Discuss evaluating procedures for Recommender Systems

Module 15: Dimensionality Reduction: SVD

Learning Objectives:

- Explain need of reducing dimensions
- Applications of dimensionality reduction

- Describe SVD for dimensionality reduction

Module 16: Dimensionality Reduction: CUR

Learning Objectives:

- Explain what CUR dimension reduction is
- Discuss how CUR is good approximation to SVD
- Explain how to compute CUR
- Discuss pros and cons of CUR

Module 17: Dimensionality Reduction: Latent Factor Models for recommender systems
Learning Objectives:

- Explain how to model global and local effects for recommender systems
- Discuss how to make good recommendations applying Latent Factor models
- Discuss how to deal with missing entries in utility matrix
- Explain how to build a recommender system by using an optimization function
- Explain how to address overfitting by regularization

Module 18: Clustering

Learning Objectives:

- Explain why clustering is hard
- Discuss methods of clustering
- Explain BFR algorithm for k-means clustering for large datasets
- Discuss limitations of BFR algorithm
- Describe CURE (Clustering Using REpresenta3ves)

Module 19: Computational Advertising

Learning Objectives:

- Explain the problem of computational advertisements
- Explain the algorithm matching
- Discuss solutions to online matching algorithm and their limitations
- Explain what performance-based advertising is
- Explain how to estimate CTR Clickthrough rate
- Describe the BALANCE algorithm
- Analyze the performance of BALANCE algorithm

Module 20: Decision Trees
Learning Objectives:

- Explain Decision Tree construction
- Use Map Reduce to construct Decision Tree

Course 4: Process Mining (2 credits)

By the time you finish this course you should: - Explain Business Process Intelligence techniques (in particular process mining), Relate process mining techniques to other analysis techniques such as simulation, business intelligence, data mining, machine learning, and verification, apply basic process discovery techniques to learn a process model from an event log (both manually and using tools), apply basic conformance checking techniques to compare event logs and process models (both manually and using tools), Extend a process model with information extracted from the event log (e.g., show bottlenecks), Determine the data needed to start a process mining project, characterize the questions that can be answered based on such event data, explain how process mining can also be used for operational support (prediction and recommendation.
Resource: https://www.coursera.org/learn/process-mining

Module 1: Data and Process Mining, Real-life Process Mining Session

Learning Objectives:

- Explain how process mining is broader data science discipline
- Explain analyzing processes
- Illustrate different types of process mining
- Demonstrate how process mining can solve process problems you might encounter in your own organization.

Module 2: Event Logs and Process Models, Petri nets

Learning Objectives:
- Explain how events are partially ordered and refer to process instances
- Discover process models from event data.
- Use Petri nets to explain the basics of process modelling

Module 3: Alpha Algorithm

Learning Objectives:
- Use Alpha algorithm in process mining, to reconstruct causality from a set of sequences of events
- Discuss limitations of Alpha algorithm

Module 4: Intro to ProM and Disco

Learning Objectives:

- Practice ProM and Disco to explore the strength of these tools to create different process mining models
- Use ProM and Disco to create different process mining models

Module 5: Quality and Representational Bias

Learning Objectives:
- Illustrate that there are very different approaches to automatically discover processes from raw data

- Explain scalability issues and probable solutions to them

Module 6: Dependency Graphs and Causal Nets, Transition Systems and Concurrency

Learning Objectives:
- Explain heuristic mining algorithm.
- Explain how dependency graphs are very important to get the causal structure of a process model
- Construct a transition system based on an event log
- Convert the transition system to a Petri net by detecting concurrency

Module 7: Process Discovery and Conformance Checking

Learning Objectives:
- Explain how it is valuable to seek the confrontation between modeled behavior (i.e., process models) and real behavior (i.e., event logs).
- Use Conformance checking techniques to quantify the different deviations

Module 8: Exploring Event Data

Learning Objectives:
- Explain how dotted charts can provide valuable insights in the early stages of a process mining project
- Use dotted charts to get valuable insights in the early stages of a process mining project

Module 9: Decision Point Analysis

Learning Objectives:
- Discuss challenges of process mining
- Explain holistic view of process mining
- Integrate different perspectives of process mining.

Module 10: Operational Support

Learning Objectives:
- Explain how process mining can be applied on running processes
- Discover methods to get the (right) event data, process mining software, and how to get from data to results

Course 5: Data Analytics and Data Visualization (2 credits)

By the time you finish this course you should: Use scraping techniques to download data, process data for analysis, analyze data and visualize data. You should also scrape data from IMDB and create a report of ratings of movies by genre as a media firm analyst. As trade analyst you should scrape data from BSE and use historical data to find the best performing stocks, and the best predictors

Resource:
https://docs.google.com/document/d/1B3pPbRqQNol9HKVThdvN4Cy1NdzZUxP4AOTG2qxfIUU/edit?usp=sharing

Module 1: Data Repositories and resources

Learning Objectives:
- Identify different data repositories, data resources
- Identify different formats of data
- Discover different methods to acquire data from data resources

Module 2: Data Scraping Tools

Learning Objectives:
- Practice Data Scraping Tools Beautiful Soup, Scrapy or Selenium to download static and dynamic web data

Module 3: Data Parsing and Transformation

Learning Objectives:
- Download IMDB movie rating data
- Parse and transform data to find top rated movies
- Download BSE historical stock prices data
- Parse and transform it to find top gainers for the last one year

Module 4: Data Analysis: Summarization by Segment

Learning Objectives:
- Summarize the data by segment
- Discover groups performing far better and far below

Module 5: Data Analysis: Prediction using Linear Regression

Learning Objectives:
- Analyze relationships between items in the dataset
- Predict top rated items for future

Module 6: Data Visualization: Vector Graphics SVG

Learning Objectives:
- Explain image files suffer from what problems
- Explain how SVG address problems image files face
- Create SVG graphs

Module 7: Data Visualization: Design

Learning Objectives:
- Use code to generate visuals from data
- Use a template that will convert data into a HTML file -- typically embedding SVG within that. (Tornado’s template)

- Convert image files into PDF without using a browser

Module 8: Data Visualization: Templates

Learning Objectives:

- Illustrate how aesthetic sense is key to creating visualisations
- Discover how Color, Typography and Layout can be used to great effect in increasing the impact the design

Module 9: Interactive Visualizations

Learning Objectives:

- Use JavaScript, to create visualizations dynamic (changing over time) as well as interactive
- Structure templates that produce interactive visualizations, and design them to bring out the full meaning of the data -- without overwhelming the user.

Module 10: Data Visualization Tools: Sci2, Tablo

Learning Objectives:

- Create model visualizations by using tools Sci2 and Tablo