Pace University

Seindenberg School of Computer Science and Information Systems

CS632v: Intro to Big Data Analytics/Data Science (CRN: 23125/22439)

Spring 2018 [Meet every Thu 5:40-8:30pm at 163WM Rm. 1520 (NYC) & via VC (PLV)]

Instructor: Prof. Tassos H. Sarbanes - TA: tba

tsarbanes@pace.edu
Cell: 516-469-8359
Office Hours: As needed, Thursday's

Prerequisites

Math: Linear Algebra, Statistics/Probabilities courses
CS/Programming Languages: BigO/Functional or Object Oriented Programming course, Python, R
Visualization Tools: Jupyter Notebooks, Zeppelin Notebooks

Credits

3

Description

This course introduces the students to Big Data Technologies, Data Analytics at scale, and Data-driven Science systems in order to extract insights data from in various forms. These scientific processes will include various phases and techniques such as Data Preparation, Model Building, and Prediction, Clustering, Association, Regression (Linear and Logistic), Classification, Decision Trees, Textual Data Analysis and Data Presentation. The basic concepts will be covered with examples which can be tried on R or Python by using RStudio and/or Jupyter Notebooks (aka IPython Notebooks). These miniaturized examples of real-world problems are designed in such way that the student will gain a clear understanding and get firm foundation of the methods covered in the course. In addition, the course gives an introduction to R Statistical Language, Apache (Databricks) Spark, and Anaconda Analytics platforms.

Objectives

After completing this course, the students will have a clear understanding and of data mining and predictive modeling concepts and tools. The student will be able to identify the goals of an analytical project and define and outline the detailed steps for implementation. After mastering the examples given in the course, the student will be in a very good position to apply his/her knowledge to do real-world projects in analytics. Having this introductory exposure to modern big data tools and architecture, such as Hadoop and Spark; students will know when these tools are necessary and will be poised to quickly train up and utilize them in a big data project. Both R and Python have become the de facto languages for Data Science projects. R was built with statistics and data analysis in mind. Although Python is a general purpose language many libraries have been added through packages, primarily for Machine Learning applications. Should you choose R or Python for Data Science? Thanks to recent advances made in both languages, you really can't go wrong with either. I personally favor Python, especially its interactive mode (Jypyter Notebooks, previously known as IPython Notebooks).

Course Outline

Week 01 (01/25/2018)

- Introduction to Big Data -- Examples of Big Data Analytics (ML & DL)
- Sign-up with Anaconda-Cloud. Download Anaconda 4.2.x (Python 3.x version)

Week 02 (02/01/2018)

- Data Analytics Lifecycle, Data Science Workflows
- Introduction to MapReduce (MR1) and Hadoop
- Introduction to Python, PyPI (Python Package Index)

Week 03 (02/08/2018)

- Introduction to Data Types, Time Series Data, Longitudinal Data, Unstructured Text Data
- Data Exploration in Python (Data Wrangling): NumPy, SciPy, Pandas
- Introduction to Python's Visualization: Matplotlib, Plotly, Seaborn, Bokeh

Week 04 (02/15/2018)

- Introduction to R (RStudio IDE)
- Semiology of Graphics (James Bertin) -- R Graphics for Data Display (ggplot2 library)
- Introduction to Anaconda Open Data Science platform

Week 05 (02/22/2018)

- Data Analytics at Scale / Streaming Analytics
- Introduction to Apache Spark (MR2), Scala (Functional) Programming Language
- Introduction to Spark's MLlib and Sparklyr (dplyr backend)

Week 06 (03/01/2018)

- Machine Learning, Supervised Learning Techniques
- Naïve Bayes Algorithm
- Introduction to scikit-learn Python library

Week 07 (03/08/2018)

- Midterm Exam

Week 7-1 (03/15/2018)

- Spring Break. No Class!

Week 08 (03/22/2018)

- Introduction to Linear Regression (LR) & Prediction using LR
- Machine Learning Model Validation, Regularization, Overfitting, H-Parameter Tuning

Week 09 (03/29/2018)

- Introduction to Logistic Regression & Scoring with Logistic Regression
- Risk Models / Response Models
- Introduction to Support Vector Machine (SVM) and Singular Value Decomposition (SVD)

Week 10 (04/05/2018)

- Advanced Analytical Theory and Methods: Classification
- Introduction to Decision Trees Modeling (CART, Boosted Trees)
- Decision Tree Algorithms, Bagging Predictors, the XGBoost Algorithm

Week 11 (04/12/2018)

- Unsupervised Learning -- Cluster Analysis (k-Means)
- The Curse of Dimensionality
- Principal Component Analysis (PCA) : Eigenvectors, Eigenvalues and Dimension Reduction

Week 12 (04/19/2018)

- Natural Language Processing (NLP)
- Introduction to TF-IDF (Term Frequency - Inverse Document Frequency) statistical model
- Introduction to Word2vec and GloVe (Python) libraries/models

Week 13 (04/26/2018)

- Introduction to Deep Learning
- Biological Neural Networks vs Artificial Neural Networks
- Neural Networks – Perceptron Learning
- Introduction to TensorFlow, Keras (Python)

Week 14 (05/03/2018)

- Computer Vision – Convolutional Neural Networks (CNNs)
- Recurrent Neural Networks (RNNs) and LSTNs (Long Short-Term Networks)
- Introduction to TensorFlow, Keras (Python)

Week 15 (05/10/2018)

- Final Exam

Textbook

All lectures' material in PPTs (**attendance highly recommended**)
"Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data", EMC Education Services, John Wiley & Sons, Indianapolis, Indiana 2015
http://seidenberg.pace.edu/~ctappert/dps/d860-16/assign/books/2015DataScience&BigDataAnalytics.pdf

Grading Policy

- ✓ 30% for mid-term
- ✓ 30% final
- ✓ 40% for assignments and projects

Accommodations for Students with Disabilities:

The University's commitment to equal educational opportunities for students with disabilities includes providing reasonable accommodations for the needs of students with disabilities. To request a reasonable accommodation for a qualified disability a student with a disability must self-identify and register with the Office of Disability Services for his or her campus. No one, including faculty, is authorized to evaluate the need for or grant a request for an accommodation except the Office of Disability Services. Moreover, no one, including faculty, is authorized to contact the Office of Disability Services on behalf of a student. For further information, please see Resources for Students with Disabilities at www.pace.edu/counseling/resources-and-support-services-for-students-with-disabilities.

Academic Integrity: (From the Student Handbook)

The Academic Integrity Code supports Pace University's commitment to academic honesty and creates a culture at the University that emphasizes high standards of academic

integrity, ethical behavior, and responsible conduct. The purpose of the Code is to educate students about what constitutes academic misconduct, to deter cheating and plagiarism, and to create a fair process and a set of procedures to handle cases of academic misconduct including documentation and application of sanctions. Academic integrity is defined as honesty and ethical conduct in learning and the educational process. The educational environment is enhanced when students believe that their academic competence is being judged fairly and that they will not be at a disadvantage because of the dishonesty of another. All members of the University community are expected to uphold the highest standards of academic integrity.

Students are required to be honest and ethical in satisfying their academic assignments and requirements. Academic integrity requires that, except as may be authorized by the instructor, a student must demonstrate independent intellectual and academic achievements. Therefore, when a student uses or relies upon an idea or material obtained from another source, proper credit or attribution must be given. A failure to give credit or attribution to ideas or material obtained from an outside source is plagiarism. Plagiarism is strictly forbidden. Every student is responsible for giving the proper credit or attribution for any quotation, idea, data, or other material obtained from another source that is presented (whether orally or in writing) in the student's papers, reports, submissions, examinations, presentations and the like.

A student who fails to comply with the standards of academic integrity is subject to disciplinary actions such as, but not limited to, a reduction in the grade for the assignment or the course, a failing grade in the assignment or the course, suspension and/or dismissal.
Please read more at the link provided below (Student Handbook):
http://www.pace.edu/sites/default/files/files/student-handbook/pace-university-academic-integrity-code.pdf

Textbooks --- References

Learn Python The Hard Way, 3rd Edition (pdf)
https://github.com/chris-void/pyway

R for Data Science: O'Reilly --- Garret Grolemund, Hadley Wickman

Many courses in MOOCs (Massive Online Courses) : Coursera, EdX, Udacity

Books-related-to-R
https://www.r-project.org/doc/bib/R-books.html

Statistics e-book (free download)
https://www.r-statistics.com/2009/10/free-statistics-e-books-for-download/

Machine Learning Books
Pattern Recognition and Machine Learning
Christopher Bishop, Laboratory Director, Microsoft Research Cambridge

Bayesian Reasoning and Machine Learning
David Barber (Hidden Markov Models, Graphical Models - the Bayesian stuff)

Machine Learning by Tom M. Mitchell of CMU

Deep Learning Books
Deep Learning
Ian Goodfellow, research scientist at OpenAI
Yoshua Bengio of MILA (Montreal Institute for Learning Algorithms)
https://github.com/HFTrader/DeepLearningBook/blob/master/DeepLearningBook.pdf