Data Science at Chico State

Topic overview of 3 propsed courses

10/22/2017

Course 1 : Introduction to Data Science (MATH/CSCI 385)

Description: *Data Science is the science of learning from data in order to gain useful predictions and insights. The course will provide an overview of the wide area of data science, with a particular focus on to the tools required to store, clean, manipulate, visualize, model, and ultimately extract information from various sources of data. Topics include: The analytics life cycle, data extraction, integration and modeling in R/Python, data visualization, relational databases and SQL, text processing and sentiment analysis. Emphasis is placed on reproducible research, code sharing, version control, and communicating results to a non-technical audience. Sample application areas: Bioinformatics, Health Informatics, Institutional Research, Business Analytics, Environmental Science and Renewable Energies*

Prerequisite: Math maturity at the level of Survey of Calculus (MATH 109), introduction to programming (CINS 100/CSC I111/MATH 130)

Topics

What is Data Science?

- Data Analytics Lifecycle
  - CRISP-DM https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining
- *The many faces of a data scientist.* Why DS is an interdisciplinary field.
  - https://www.datacamp.com/community/tutorials/data-science-industry-infographic
  - http://ucanalytics.com/blogs/career-data-science-analytics-play-strengths/
  - Data Analyst / Data Architech / Data Engineer / Statistician / Database Administrator / Product Engineer / Business Analyist / Data Science Team Leader
- *Application Areas* With examples and faculty links. Bioinformatics, Business Analytics, Health informatics, Environmental Sciences,
  - Invited speakers (20 minutes) to talk about their application area.

How to be a Successfull Data Scientist

- Overview and Challenges
  - Hard skills vs. Soft skills - Finding and respecting the balance.
  - Working with a Data Science Team
- Project Management and Team collaboration tools
- Automating the Data Science Project Workflow
  - Version control using Git/Github

Getting Data

- Reading and writing to files from various types
- Scraping the web for data
- Connecting to API's

- Relational Databases with SQL

Preparing Data for Analysis

- Data types & summary statistics
- Quality control
  - Data editing
  - Identify and deal with missing data
- Transformations
- Data aggregation
- Concepts of Tidy data
  - When and how to reshape
- Dates and times

Data Exploration: Uncovering Valuable Information

- Question Formulation
- Data Visualization & Interpretation
- Unstructured data analysis
  - string manipulation
  - regular expressions
  - word frequency & clouds
  - sentiment analysis
- Cluster identification
  - knn, concept of distance
  - Dendograms, Heatmaps

Sharing the Insights: Reporting and Dissemination

- Reproducible reports using Markdown and LaTeX
- Web-based interactive apps using Shiny
- Creating slide decks (Beamer/Slidify)
- Presenting results to a non-technical audience

Course 2: Advanced Topics in Data Science (MATH/CSCI 485)

Description: *How to be a successful Data Scientist: Overview and challenges, project management and effective team collaboration tools and methods, workflow automation. Ethics of predictive analytics and privacy and open data. Reporting and dissemination of research using interactive dashboards and web-publishing. Identifying and applying appropriate methods to answer a business or research need. Introduction to current scalable technologies to handle Big Data. Introduction to advanced statistical analysis for Data Science: Sample topics include: Non-Parametric techniques, Simulation methods, Network analysis, Experimental Design, Predictive modeling using Supervised and Unsupervised Machine Learning techniques, Modeling and mapping of Geospatial data.*

Prerequisites: Introduction to Data science (MATH/CSCI 385), Corequisite class: MATH 456 (Applied Statistics II)

Topics

How to be a Successfull Data Scientist, continued. . .

- Overview and Challenges
  - Hard skills vs. Soft skills - Finding and respecting the balance.
  - Working with a Data Science Team
- Project Management and Team collaboration tools
- Automating the Data Science Project Workflow
  - Automation via Makefiles

Privacy, Ethics and Open Data

- Ethics of building and using predictive models
- Open data in a hostile world

Advanced Analytical Methods

- Choosing the right analytic tool for the job. - Hammer vs saw
- Non-Parametric Analysis
  - An overview of alternative tests for when assumptions just aren't met.
- Mindful and Targeted data collection
  - Identifying target metrics, Designing data collection tools, Analyzing results, transforming results to action.
- Predictive Analytics
  - Cross Validation
  - Navigating the Machine Learning Soup of Algorithms
    - Classification & Regression
    - Ensemble Models
  - Quantifying model fit, generalization, predictive ability, uncertainty
    - Regularization: Ridge, Lasso, Bias-variance trade-off
    - Feature Selection & Dimension Reduction

Big Data Processing and Analysis

- Different architecture needed: HDFS / Hadoop / Apache Spark
- Methods of connecting and using that architecture: `sparklyr`, NoSQL
- Assessing scalability of models and methods
- Amazon Cloud Services
- Writing parallizable code

Advanced Reporting and Dissemination

- Authoring websites using Rmarkdown/Github/Jekyll
- Deploying your code with confidence using Travis CI
- Creating and deploying Interactive Dashboards
- JavaScript and D3.js, plotly

Specialized Topics

The following topics could be introduced as short, introductory modules. This is not an exhaustive list. Topics covered are up to the instructor's discresion. Not all topics on this list can be covered in a single semester in great depth. This is about breadth and exposure to different analytical tools and methods, not about depth and theory. Many topics are covered in greater theoretical depth in other classes such as AI and Graph Theory.

- Time Series forecasting
- Modeling of Spatially correlated data.
  - Using GIS shapefiles
- Simulation Techniques
  - Bootstrapping and Beyond!
- Bioinformatics
  - Gene processing, frequency heat mapping
  - Hierarchical clustering
- Network Analysis
  - Introduction to Graph Theory
  - Social Networks, Images, protein interactions

Course 3 : Capstone Project in Data Science

Description: *Students will work independently to provide a service in the form of a data product to a local business, researcher, or community member. Students provide status reports at weekly meetings, and present their finished project to a group of peers at the end of the semester in an appropriate venue such as at an undergraduate seminar series or poster symposium. Existing capstone courses such as CSCI 490 or CINS 490 or Independent study projects through CSCI 499 or MATH 499 can substitute providing an approved Data Science project.*

The final project report must be submitted to and approved by the independent study faculty mentor, with copy provided to the Data Science advisor at the completion of the study. Student will work with the Data Science advisor to display their finished work in a public setting.