M462 – Theoretical Big Data Analytics
M562 – Advanced Theoretical Data Analytics

Instructor Information:

Instructor: Javier Pérez-Álvaro
Office: M301
Email: javier.perez-alvaro@mso.umt.edu
Office hours: See http://www.umt.edu/people/perezalvaro for up-to-date OH.
Time and place**:** Monday, Wednesday, Friday 2:00-2:50 p.m., Math 103.

Course description: Offered spring. Prereq., M 221 and two other Mathematics / Statistics classes at the 200-level or above, or consent of instr. The main goal of this course is to provide students with the opportunity to acquire conceptual knowledge and theoretical understanding of mathematical methods applicable to data analytics and real-time computations.

Learning Outcomes:

1. Understand the theory and foundations of predictive analytics and dimension reduction.
2. Develop understanding and practical experience in optimizing objective functions.
3. Ability to formulate and implement algorithms for predictive analytics.
4. Understand the mathematical basis for artificial neural networks.

Topics: We will cover a large number of Machine Learning/Data Science techniques:

1. Linear Regression
    a. Direct methods (Normal equations, QR factorization, SVD-PCA factorization)
    b. Iterative methods (Gradient descent, stochastic gradient descent, gradient descent with momentum)

2. Overfitting and Regularization

3. Classification Algorithms
    a. Logistic regression
    b. K-nearest neighbors

4. Dimensionality Reduction Algorithms:
    a. projection onto lower-dimensional spaces: Singular Value Decomposition – Principal Component Analysis (SVD-PCA), t-Distributed Stochastic Neighbor Embedding (t-SNE
    b. Visualization of high-dimensional data
    c. Anomaly detection

5. Clustering Algorithms
    a. K-Means
    b. Hierarchical clustering
    c. DBSCAN
    d. Spectral clustering

6. Recommendation Systems
7. Neural Networks

Getting Python: You can download Python from Python.org. But if you don't already have Python, I recommend instead installing the Anaconda distribution (www.anaconda.com/download), which already includes most of the libraries that you need to do data science.

Text book: There is no required textbook for this course

Homework: Course grade is based on homework. Homework assignments often consist of a set of problems, both mathematical and Python programming, from which students choose and complete/solve. M 462/562 students sometimes differ with respect to mathematical preparation and domain knowledge and so the assignments will contain problems that are appropriate to preparation. Approximately 10 homework assignments will be assigned and the point value will vary between 10 and 30. Homework will be collected approximately every 10 days (depending on the difficult of the task). Students are encouraged to work together.

Final Project: In lieu of a final, students will be graded on their final project. Students must work on projects in groups of one or two individuals. Students are responsible for a written paper (75% of the project grade) and oral presentation (25% of the project grade).

Grading policy: Your course grade will be based on homework and a final (take-home) exam

Item	Percentage
Homework	80%
Project	20%

Recommended books/readings:

1. Python and Python's main scientific libraries: pandas, NumPy and Matplotlib
    a. If you don't know Python yet, *http://learnpython.org/* is a great place to start. The official tutorial on Python.org, *https://docs.python.org/3/tutorial*, is also quite good.
    b. *Python Pocket Reference,* by Mark Lutz
    c. *Python for Data Analysis: Data Wrangling with Pandas, NumPy and Ipython,* by Wes McKinney.
    d. *Python Data Science Handbook: Essential Tools for Working with Data*, by Jake VanderPlas.

2. Linear Algebra
    a. Introduction to Linear Algebra, by Gilbert Strang.

3. Machine Learning/Data Science
    a. *Pattern Recognition and Machine Learning*, Christopher M. Bishop, Springer.

b. *Hands-On Unsupervised Learning using Python,* by Ankur A. Patel.
c. *Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow,* by Aurelien Geron.
d. *Data Science from Scratch,* by Joel Grus.
e. *Deep Learning from Scratch,* by Seth Weidman.
f. *Algorithms for Data Science,* by Brian Steele, John Chandler, Swarna Reddy.
g. *Doing Data Science: Straight Talk from the FrontLine,* by Cathy O'Neil and Rachel Schutt.

4. Mathematics of Machine Learning/Data Science
    a. *Linear Algebra and Learning from Data*, by Gilbert Strang.
    b. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, by Trevor Hastie, Robert Tibshirani, and Jerome Friedman. Book website
    c. *Matrix Methods in Data Mining and Pattern Recognition*, by Lars Elden.
    d. *Convex Optimization*, by Stephen Boyd, and Lieven Vandenberghe. Available here
    e. *Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers*, Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Available here
    f. *Recommender Systems,* by Charu C. Aggarwal.

Student Conduct: All students need to be familiar with the Student Conduct Code. You can find in the "A to Z Index" on the UM home page. All students must practice academic honesty. Academic misconduct is subject to an academic penalty by the course instructor and/or a disciplinary sanction by the University.

Accommodation: The University of Montana assures equal access to instruction through collaboration between students with disabilities, instructors and Disability Services for Students (DSS). If you think that you may have a disability adversely affecting your academic performance, and you have not already registered with DSS, please contact DSS in Lommasson Center 154 or call 406.243.2243. I will work with you and DSS to provide an appropriate accommodation.

Important note: Announcements made in class are considered addenda to this syllabus. Make sure you stay informed as the progress of the class.