DATA-602: Introduction to Data Analysis and Machine Learning

UMBC

Instructor's Information

Instructor: Dr. Masoud Soroush
Associate Graduate Program Director
Data Science Group
Department of Computer Science and Electrical Engineering
University of Maryland Baltimore County (UMBC)

Email: masoud.soroush@umbc.edu
Office: Information Technology and Engineering (ITE) Building, Room 374

Class Information

Class Time: Wednesday 7:10 - 9:40 PM
Location: Online through Blackboard Collaborate (ULTRA)

Office Hour: Thursday 3:30 - 5:30 PM
No appointments are necessary if you walk in during the above time interval!

Prerequisite: DATA 601

TA: Nishit Vyas
TA Email: nvyas1@umbc.edu
TA Office Hour: Monday/Tuesday 5:00 - 6:00 PM

Course Description

This course provides a broad introduction to machine learning and data analysis. A wide range of both supervised and unsupervised techniques will be introduced. Topics covered include linear and polynomial regression analyses, regularization methods, logistic regression, support vector machine, naive Bayes, linear and quadratic discriminant analysis, decision trees, ensemble learning methods, neural networks, clustering methods, and dimensionality reduction algorithms. This class will cover both the theoretical and practical aspects of the algorithms. Although a great emphasis will not be placed on the details of the theoretical aspects of the algorithms, some familiarity with linear algebra, calculus and basic statistics is required to understand how these machine learning algorithms work. Necessary mathematical concepts will be reviewed as needed.

Required Software

The course will be using Python 3 with the following libraries: numpy, pandas, matplotlib, sklearn, scipy, sympy, keras, tensorflow, and pytorch. It is the student's responsibility to have a working environment. If you would like to have the environment installed locally, Anaconda is a Python distribution that has all required libraries. Alternatively, you may use Google Colab or Deepnote.

Useful Textbooks and Resources

The following books on machine learning will be useful to deepen your knowledge in this course:

1. https://www.oreilly.com/library/view/hands-on-machine-learning/9781492032632/
   Aurelien Geron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*, O'Reilly Media; 2nd edition, 2019.
2. https://www.packtpub.com/product/python-machine-learning-third-edition/9781789955750
   Sebastian Raschka and Vahid Mirjalili, *Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2*, Packt Publishing; 3rd edition, 2019.
3. https://www.packtpub.com/product/machine-learning-with-pytorch-and-scikit-learn/9781801819312
   Sebastian Raschka, Yuxi (Hayden) Liu, Vahid Mirjalili, and Dmytro Dzhulgakov, *Machine Learning with PyTorch and Scikit-Learn*, Packt Publishing; 1st edition, 2022.
4. https://www.packtpub.com/product/mastering-machine-learning-algorithms-second-edition/9781838820299
   Giuseppe Bonaccorso, *Mastering Machine Learning Algorithms*, Packt Publishing; 2nd edition, 2020.
5. https://www.packtpub.com/product/machine-learning-algorithms-second-edition/9781789347999
   Giuseppe Bonaccorso, *Machine Learning Algorithms: : Popular Algorithms for Data Science and Machine Learning*, Packt Publishing; 2nd edition, 2018.
6. https://www.oreilly.com/library/view/introduction-to-machine/9781449369880/
   Andreas C. Müller and Sarah Guido, *Introduction to Machine Learning with Python*, O'Reilly Media, Inc.; 1st edition, 2016.
7. https://scikit-learn.org/stable/
   scikit-learn Library Manual Guide
8. https://www.statlearning.com/
   Gareth James, Daniela Witten, Trevor Hastie, and Rob Tibshirani, *An Introduction to Statistical Learning*, Springer; 2nd edition, 2021.

Assignments

- **Presentation:**
  You will prepare and present a 6 to 8-minute presentation to the class on a topic of your interest that illustrates an application of machine learning to a specific problem. This can be based on your work experience or an article you have read and found interesting. Your presentation should include defining the problem, explain the methodology used and type(s) of algorithm involved to solve the problem. You can use Power Point slides for your presentation (No more than 5 slides).

- **Homework:**
  You will complete 3 homework assignments in this course. All homework assignments are completed individually and are submitted through Blackboard. For each homework assignment, you will submit one *single* fully executable jupyter notebook. Although you may consult lecture notes, notebooks, and other resources to complete the homework assignments, no collaboration with other students is allowed! Homework assignments must be submitted by the given deadlines. *Late submissions will receive no or very little credit!*

- **Group Project:**
  You will complete a group machine learning project on a topic of interest. The main goal of the project is to prepare you for a real-world data analytics task. You will work in a group formed by 4 or 5 students. You need to identify a problem or issue for your group. Your research problem should include a *supervised* component that has not been extensively discussed in the literature. You need to get the data for your research problem from official sources (*e.g.* local, state, federal, and/or international agencies) to investigate the problem. Few links to publicly available datasets in different research areas have been listed below, and you may use them to find a relevant dataset for your problem.

  - https://github.com/awesomedata/awesome-public-datasets
  - https://academictorrents.com/
  - https://datasetsearch.research.google.com/
  - https://data.gov/
  - https://www.earthdata.nasa.gov/
  - https://apps.who.int/gho/data/node.home
  - https://crime-data-explorer.fr.cloud.gov/pages/home
  - https://www.ncei.noaa.gov/weather-climate-links#loc-clim

  After carefully formulating the problem, you need to build supervised models to predict your target(s). You may then expand your study to find out if appropriately designed neural networks can improve the performance (This part is optional though). You may also employ unsupervised techniques to provide further insights into your problem. Finally, you need to state your recommendations to solve the problem. In doing that you need to carefully state the limitations of the data, study, and model(s) you dealt with. If there are already established models to address your problem, you need to compare your model with other existing models, and identify advantages and drawbacks of your model.
  Project Deliverables:
  - One single self-contained fully executable jupyter notebook. The notebook must be well documented, and it should be supported with the necessary and sufficient explanation in markdown cells.
  - A 25-minute presentation to present your research problem and your findings. The format of the presentation (Power Point, Keynote, Jupyter Notebook, etc.) is flexible and is chosen by the group, but all members should participate in presenting the results.

Grade Distribution

The weights of different components of the coursework are given in the following table.

Component	Weight
Presentation and Attendance	10%
Homework 1	15%
Homework 2	15%
Homework 3	15%
Group Project	45%
Total	100%

Final letter grades will be computed according to the following table.

Weighted Percentage	Grade
90% - 100%	A-, A, A+
80% - 89%	B-, B, B+
70% - 79%	C-, C, C+
60% - 69%	D
Below 60%	F

Course Policies

UMBC provides a range of writing assistance, which can be found at https://academicsuccess.umbc.edu/. For Research Guides Tutorials, visit https://library.umbc.edu/tutorials/index.php. Failure to follow guidelines for each assignment, including the required format, style, length, submission, etc. may result in at least one-letter-grade reduction on the assignment depending on the type and/or number of transgressions. Late assignments will not be accepted unless an extension has been agreed to in advance. Emergency situations will be handled on a case by case basis with appropriate justification and/or documentation. *Please do not ask for an extension unless you truly encounter an emergency*! Incomplete grades will not be entertained unless extenuating circumstances warrant and your request is made before the last week of class.

Academic Integrity

By enrolling in this course, each student assumes the responsibilities of an active participant in UMBC's scholarly community in which everyones academic work and behavior are held to the highest standards of honesty. Cheating, fabrication, plagiarism, and helping others to commit these acts are all forms of academic dishonesty, and they are unacceptable. Academic misconduct could result in disciplinary action that may include, but is not limited to, failure, suspension or dismissal. For more information on UMBC's Code of Conduct visit: https://catalog.umbc.edu/content.php?catoid=32&navoid=2186.

Disability Accommodations

UMBC is committed to eliminating discriminatory obstacles that may disadvantage students based on disability. Services for students with disabilities are provided for all students qualified under the Americans with Disabilities Act (ADA) of 1990, the ADAAA of 2009, and Section 504 of the Rehabilitation Act who request and are eligible for accommodations. The Office of Student Disability Services (SDS) is the UMBC department designated to coordinate reasonable accommodations that would allow students to have equal access and inclusion in all courses, programs, and activities of the University.

If you have a documented disability and need to request academic accommodations, please register with the Office of Student Disability Services (SDS) as soon as possible. To begin the registration process please visit the SDS website (https://sds.umbc.edu/accommodations/registering-with-sds/) and review the registration information, including disability documentation guidelines and how to submit the SDS registration form online using the confidential data management software called Accommodate.

Once accommodations have been approved, you and your instructors will be notified via an official accommodation letter from the SDS office. Both the SDS office and Shady Grove's Center for Academic Success (CAS) will work with you to ensure you receive the approved accommodations. If you have any questions or concerns, please contact the Office of Student Disability Services via email at disability@umbc.edu or phone at ☎: 410-455-2459. Please note that accommodations are not retroactive and begin once SDS sends an approved accommodation letter.

Title IX

UMBC Policy and Federal law (Title IX) prohibit discrimination and harassment on the basis of sex, sexual orientation, and gender identity in University programs and activities. Any student who is impacted by sexual harassment, sexual assault, domestic violence, dating violence, stalking, sexual exploitation, gender discrimination, pregnancy discrimination, gender-based harassment or retaliation should contact the University's Title IX Coordinator to make a report and/or access support and resources:

Ever Hanna, Title IX Coordinator

☎: 410-455-1026
Email: everhanna@umbc.edu

You can access support and resources even if you do not want to take any further action. You will not be forced to file a formal complaint or police report. Please be aware that the University may take action on its own if essential to protect the safety of the community.

If you are interested in or thinking about making a report, please use the online reporting at https://umbc-advocate.symplicity.com/titleix_report/index.php/pid747366?. Please note that, if you report anonymously, the University's ability to respond will be limited. For further information on this matter, please visit https://oei.umbc.edu/sample-title-ix-responsible-employee-syllabus-language/.

Course Outline

Week 1	*Basics of Machine Learning* **Objectives**: Machine learning definition and its objectives, Different types of machine learning (supervised and unsupervised), Characteristics for adopting ML algorithms, Reviewing basic linear algebra concepts, Reviewing function optimization in calculus, Reviewing hypothesis testing in statistics
Week 2	*Regression Analysis* **Objectives**: Linear and polynomial regressions, Metrics to assess regression models (RSE, R^2-score, F-statistic), Correlation Analysis (Pearson correlation, Spearman correlation, Kendall correlation), Bias-variance trade-off
Week 3	*Regularization Methods* **Objectives**: Definition of regularization and its objectives, Diagnosing overfitting and underfitting in ML, Ridge regularization, Lasso regularization, Elastic-net regularization, Lasso regularization as a feature selection method
Week 4	*Logistic Regresssion* **Objectives**: Objectives of classification algorithms, Logistic regression as a classification algorithm, Binary logistic regression, Multinomial logistic regression, Metrics to evaluate classifiers (accuracy, precision, recall, $F1$-score), Confusion matrix, ROC curve
Week 5	*Support Vector Machine*

	Objectives: Support vector classifier, Hard margin, Soft margin, Support vector machine and the kernel method, Multinomial classifications and SVM, Searching for best hyperparameters (GridSearchCV)
Week 6	*Naive Bayes, LDA, and QDA* **Objectives**: Bayes theorem, Naive Bayes classifier, Linear Discriminant Analysis, Quadratic Discriminant Analysis, A comparison of probabilistic classifiers
Week 7	*Decision Trees* **Objectives**: Basics of decision trees, Classification trees, Regression trees and regressors, Impurity measures, Feature importance, Advantages and disadvantages of trees
Week 8	*Ensemble Learning Methods* **Objectives**: Bagging methods, Random forests, Boosting methods, Gradient boosting, Voting classifier
Week 9	*Neural Networks I* **Objectives**: Artificial Neural Networks, Prominent types of deep learning architectures, Fully-connected networks, Loss function, Activation functions, Learning rate, Optimizers, Basics of pytorch, Basics of tensorflow,
Week 10	*Neural Networks II* **Objectives**: Basics of convolutional neural networks (CNN), Useful convolutions, Padding, Pooling layers
Week 11	*Neural Networks III* **Objectives**: Basics of recurrent neural networks (RNN), Feeding and training RNN networks in pytorch
Week 12	*Clustering Algorithms* **Objectives**: K-means clustering under different metrics, Hierarchical and agglomerative clustering, DBSCAN
Week 13	*No Class!* Thanksgiving break!
Week 14	*Dimensional Reduction* **Objectives**: Singular Value Decomposition (SVD), Principal Component Analysis (PCA) as a dimensionality reduction algorithm, Explained proportion of variance, Nonnegative Matrix Factorization (if time permits)
Week 15	*Group Project Presentations* **Objectives**: Each group presents a 25-minute talk summarizing their research problem and their findings.