COMP 4448 Data Science Tools 2

Course Overview

Data science is a cross-disciplinary field that requires an integrated skill set spanning statistics, machine learning, data mining, and more. The Data Science Tools 2 course explores various tools needed to perform data analysis at different phases of the data science life cycle. This course covers the workflow of the data science process, including data preprocessing, data partitioning, model construction, and model evaluation. The course focuses on supervised (classification and regression) models, including rule-based classification, decision trees, k-nearest neighbors, naïve Bayes, linear regression, and logistic regression, as well as unsupervised models such as clustering. Students will learn how the algorithms used for constructing these models work. The Python programming language will be used to explore the concepts in this course. This course assumes some familiarity with Python programming fundamentals, including basic Python data structure, control flow statements, and functions. Students are also expected to be familiar with key data analysis packages in Python such as NumPy, Pandas, statsmodels, scikit-learn, SciPy, matplotlib, and seaborn.

Objectives

Students will be able to

- Prepare, clean, and transform data as well as perform feature selection and extraction.
- Partition data and construct and evaluate models.
- Implement supervised learning models, including decision trees, k-nearest neighbors, naïve Bayes, linear regression with gradient descent, and logistic regression with gradient ascent.
- Implement unsupervised learning models such as clustering.

Textbooks and Materials

There is no required textbook for this course, but the following book is recommended:

- Larose, C. D., & Larose, D. T. (2019). *Data science using Python and R*. Retrieved from https://ebookcentral.proquest.com
    - You can access this book through DU's library using this link: https://ebookcentral.proquest.com/lib/du/reader.action?docID=5741211&ppg=172

Grading

Assignment/Assessment	Points	Weight on Final Grade
Quizzes (Weeks 2 and 3)	100 (50 points each)	12.50%
Assignments (Weeks 1, 4-10)	400 (50 points each)	50.00%
Participation (Attend class on time, leave at the end of class time, participate in in-class activities, ask and answer questions, do all asynchronous discussions)	100 (10 points each week)	12.50%
Final Project (Written)	100	12.50%
Final Project (Presentation)	100	12.50%
Total	800	100.00%

Overall course percent	Grade
93.0% - 100%	A
90.0% - 92.99%	A-
86.0% - 89.99%	B+
83.0% - 85.99%	B
80.0% - 82.99%	B-
76.0% - 79.99%	C+
73.0% - 75.99%	C
70.0% - 72.99%	C-
66.0% - 69.99%	D+
63.0% - 65.99%	D
60.0% - 62.99%	D-
<60%	F

Assignment and Assessment Information

Quizzes

Students will complete two quizzes on the data science process during Week 2 and Week 3 of the course. The first quiz is based on the content covered during Week 2 and is due before the Week 2 live session. The second quiz is based on the content covered during Week 3 and is due before the Week 3 live session.

Assignments

There are seven weekly assignments to be completed from Week 3 to Week 9. Weekly assignments typically involve data analysis. Each week's assignment is based on the content (algorithms, methods, and models) covered for that week. These assignments will provide students with hands-on experience in implementing algorithms and applying various methods and models to diverse datasets in order to solve real-world problems. The assignment for each week is due 48 hours after that week's live session. It is recommended that students use

Jupyter Notebook to complete the assignments. The description and requirements of each assignment are in the assignment Word document. For each question inside the Word document, students need to screenshot their code and output of the analysis and paste inside the Word document whenever required. Students are required to submit both the assignment Word document and their Jupyter Notebook file (.ipynb) to the course site.

Final Project (Written)

Find a dataset that could be analyzed using a specific algorithm or method of interest covered in this course, to answer a research question. Your written project should address the following:

- Describe the dataset and the source. Your dataset should not be related to any dataset already used in class examples. Avoid using some popular datasets such as the iris dataset. Find some dataset that you think is unique and interesting to you. Though you can use data from the internet, avoid copying analysis examples directly from the internet. Instead, apply the knowledge you have learned from this course to your work.
- Describe the variables, both input and output variables, that would be used for your analysis. You can include descriptive statistics for numerical variables, frequencies of categorical variables, and appropriate graphs such as bar charts for categorical variables and histograms of numerical variables.
- State the research question to be answered through your analysis. Your research question should focus on a supervised or unsupervised learning problem.
- Describe what you did for data preprocessing and how you partitioned the data. Also include rationale for why you preprocessed and partitioned the data the way you did.
- You can code an algorithm from scratch to address your research question or use algorithms already built into the Python package, such as scikit-learn and statsmodels, to build a model that addresses your research question. Build different models (of the same functional form or different functional forms) that can address the research question; then compare the model to select which one has a better performance using a validation dataset. Write a brief report on how you implemented the algorithm(s) and which model you selected and why. Include validation results or graphs that supported your decision for model selection.
- Evaluate your final (selected) model using the test dataset to understand how well your model generalizes to new examples or input data. Report the overall accuracy of your model.

Final Project (Presentation)

This part of the final project involves presenting what you did for your final project. This presentation will be done during the Week 10 live session. Every student will present their work. The presentation should address the following:

- Slide 1: Title page
- Slide 2: Purpose of your analysis and research question
- Slide 3: Description of your dataset and source
- Slide 4: Description of input (or input and output) variables in your dataset
- Slide 5: Data preprocessing

- Slides 6–7: Relevant descriptive statistics, frequency tables, and/or graphs
- Slide 8: Data splitting or partitioning
- Slide 9: Code snippet for your algorithm or a screenshot of code for building the model
- Slide 10: Model selection and model evaluation—performance measures, graphs, and so forth.
- Slide 11: Conclusion

Participation

Your participation grade each week is determined by

- Completion of the asynchronous discussions
- Attendance and active participation in each week's live session

Weekly Schedule

The weekly schedule includes asynchronous discussions in addition to the weekly assignments. Please complete readings prior to beginning this week's asynchronous content for the indicated week. Most of the readings are from the text *Data Science Using Python and R* by Larose and Larose. It is recommended that you skip sections of the books where R code is used for illustrations and instead focus on examples that use Python code.

Week 1. Python Quick Review and Introduction to Data Science

Readings:

- Quick Python Review Notes
- NumPy and Pandas Basics Notes
- Data Manipulation, Cleaning, and Transformation Notes

Assignment (due 48 hours after Week 1 live session):

- Week 1 Assignment (Quick Python Review)

Week 2. The Data Science Process, Part 1

Readings:

- Larose, C. D., & Larose, D. T. (2019). *Data science using Python and R*
    - *Chapter 1: Introduction to Data Science*
    - *Chapter 3: Data Preparation*
    - *Chapter 4: Exploratory Data Analysis*

Discussion Prompt (due before Week 2 live session):

- It is usually said that data scientists spend 80% of their time on cleaning, preparing, and managing data for analysis. Using at least two paragraphs, explain why you think data cleaning and preparation is a vital aspect of data analysis.

Quiz 1 (due before Week 2 live session)

Week 3. The Data Science Process, Part 2
Readings:

- Larose, C. D., & Larose, D. T. (2019). *Data science using Python and R*
    - *Chapter 5: Preparing to Model the Data*
    - *Chapter 7: Model Evaluation*

Discussion Prompt (due before Week 3 live session):

- Why is it not a good idea to use only a single dataset such as the training set to evaluate the performance of a model?
- In order to evaluate the performance of a model, the data is usually split into training set, validation set, and test set. In at least two paragraphs, explain the goal of the following:
    a) Training set
    b) Validation set
    c) Test set

Quiz 2 (due before Week 3 live session):

Week 4. Rule-Based Classification

Readings:

- Han, Jiawei, et al. (2011). *Data mining: Concepts and techniques*. Elsevier Science & Technology. ProQuest Ebook Central, https://ebookcentral.proquest.com/lib/du/detail.action?docID=729031.
    - *Chapter 8.4: Rule Based Classification*

Discussion Prompt (due before Week 4 live session):

- IF–THEN rules can be used for classification. What are some performance measures that can be used to evaluate the accuracy of the rule, and how do these measures differ? What are the advantages and limitations of each performance measure? You should compare at least two performance measures for evaluating the performance of an IF–THEN rule.

Assignment (due 48 hours after Week 4 live session):

- Week 4 Assignment (Rule-Based Classification)

Week 5. Decision Trees and Random Forest
Readings:

- Larose, C. D., & Larose, D. T. (2019). *Data science using Python and R*
    - *Chapter 6: Decision Trees*

Discussion Prompt (due before Week 5 live session):

- A decision tree is used for both classification and regression. Sometimes, to improve classification accuracy and avoid overfitting, a combination of decision tree models can be used instead of a single decision tree model.
- The concept of using several decision tree models for a classification or regression task is known as random forest. In at least two paragraphs, explain in your own words how random forest works.

Assignment (due 48 hours after Week 5 live session):

- Week 5 Assignment (Decision Trees)

Week 6. K-Nearest Neighbors

Readings:

- Han, Jiawei, et al. (2011). *Data mining: Concepts and techniques*. Elsevier Science & Technology. ProQuest Ebook Central, https://ebookcentral.proquest.com/lib/du/detail.action?docID=729031.
    - *Chapter 9.5: Lazy Learners (or Learning From Your Neighbors)*

Discussion Prompt (due before Week 6 live session):

- K-nearest neighbor is a classification algorithm. In at least two paragraphs, explain how k-nearest neighbor is unique or different from other classification algorithms, such as decision tree algorithms.

Assignment (due 48 hours after Week 6 live session):

- Week 6 Assignment (K-Nearest Neighbor)

Week 7. Naïve Bayes and Probabilistic Modeling
Readings:

- Larose, C. D., & Larose, D. T. (2019). *Data science using Python and R*

- *Chapter 8: Naïve Bayes Classification*

Discussion Prompt (due before Week 7 live session):

- The naïve Bayes classifier is a special type of Bayesian classifier. What makes the naïve Bayes classifier different from the general Bayesian classifier? What is the difficulty in using a Bayesian classifier, and how does naïve Bayes solve this problem?

Assignment (due 48 hours after Week 7 live session):

- Week 7 Assignment (Naïve Bayes)

Week 8. Linear and Polynomial Regression

Readings:

- Larose, C. D., & Larose, D. T. (2019). *Data science using Python and R*
  - *Chapter 11: Regression Modeling*

Discussion Prompt (due before Week 8 live session):

- What are the four major assumptions of a linear regression model, and how is each of these assumptions tested or checked?
- What are the consequences of violating any of these assumptions?

Assignment (due 48 hours after Week 8 live session):

- Week 8 Assignment (Linear Regression)

Week 9. Logistic Regression

Readings:

- Larose, C. D., & Larose, D. T. (2019). *Data science using Python and R*
  - *Chapter 13: Generalized Linear Modeling*

Discussion Prompt (due before Week 9 live session):

- A logistic regression is a classification model. Present three or more reasons why is it not a good idea to use a linear regression for classification.
- How is a logistic regression different from a classification model such as a decision tree? That is, what are other unique advantages of using a logistic regression over a decision tree?

Assignment (due 48 hours after Week 9 live session):

- Week 9 Assignment (Logistic Regression)

Week 10. K-Means and Hierarchical Clustering
Readings:

- Larose, C. D., & Larose, D. T. (2019). *Data science using Python and R*
  - *Chapter 10: Clustering*

Discussion Prompt (due before Week 10 live session):

- How is clustering different from regression or classification models?
- Clustering requires a distance metric such as Euclidean distance, squared Euclidean distance, or Mahalanobis distance. However, such distances require that the values of the variables be numerical. If your dataset has both numerical and categorical variables, how would you prepare the data for clustering, or what appropriate distance metric would be more suitable?

Assignment (due 48 hours after Week 10 live session):

- Week 10 Assignment (K-Means and Hierarchical Clustering)

Attendance Policy

Attendance at all live session meetings is mandatory.

Program Mission

Our MS in data science provides students with a broad course of study in programming, algorithms, statistics, and data management, as well as a depth of understanding in specific fields such as data mining, machine learning, and parallel systems. Graduates of the data science program go on to work in a wide variety of careers, including business, government, education, and the natural sciences.

Honor Code and Academic Integrity

All students are expected to abide by the University of Denver Honor Code. These expectations include the application of academic integrity and honesty in your class participation and assignments. Violations of these policies include, but are not limited to,

- Plagiarism, including any representation of another's work or ideas as one's own in academic and educational submissions.

- Cheating, including any actual or attempted use of resources not authorized by the instructor(s) for academic submissions.
- Fabrication, including any falsification or creation of data, research, or resources to support academic submissions.

Violations of the Honor Code may have serious consequences including, but not limited to, a zero for an assignment or exam, a failing grade in the course, and reporting of violations to the Office of Student Conduct.

Diversity, Inclusiveness, Respect

DU has a core commitment to fostering a diverse learning community that is inclusive and respectful. Our diversity is reflected by differences in race, culture, age, religion, sexual orientation, socioeconomic background, and myriad other social identities and life experiences. The goal of inclusiveness, in a diverse community, encourages and appreciates expressions of different ideas, opinions, and beliefs, so that conversations and interactions that could potentially be divisive turn instead into opportunities for intellectual and personal enrichment.

A dedication to inclusiveness requires respecting what others say, their right to say it, and the thoughtful consideration of others' communication. Both speaking up AND listening are valuable tools for furthering thoughtful, enlightening dialogue. Respecting one another's individual differences is critical in transforming a collection of diverse individuals into an inclusive, collaborative, and excellent learning community. Our core commitment shapes our core expectation for behavior inside and outside of the classroom.