University of California, Merced
SPRK 001: The Machine Learning Age
Spring 2021

Instructor: Adam M. Croom, Ph.D. acroom@ucmerced.edu

Office Hours: By appointment on Zoom
https://ucmerced.zoom.us/my/adamcroom

Lectures: Tuesdays and Thursdays, 11:30 am - 1:20 pm

Course Description: This is an introductory course on machine learning that introduces students to key concepts in machine learning, ethical implications of machine learning, and practical coding skills in Python. Throughout this course we will examine how machine learning is used for a wide variety of practical applications, including object and facial recognition, classifying songs and movies into different genres, creating recommender systems to help users find items that are similar to others that they like, determining the most liked and shared posts on social media sites like Twitter and Facebook, and much more. In addition to exploring a variety of practical applications for machine learning in this course, we will also review key concepts in machine learning, reflect on the ethical implications of machine learning, and develop the practical coding skills required to complete a full machine learning project workflow from beginning to end.

Topics and techniques that will be covered in this course include: (1) **Machine learning concepts**: we will begin by learning about key concepts in machine learning such as *machine learning*, *artificial intelligence*, *data science*, *training data*, *validation data*, *testing data*, *neural networks*, *decision trees*, *deep learning*, *supervised learning*, *unsupervised learning*, *image classification*, and *natural language processing*; (2) **Python programming fundamentals**: we will learn how to code with Python and practice using Jupyter Notebooks to work with scripts, functions, lists, dictionaries, data frames, Boolean operators, loops, and random number generators; (3) **Python packages**: we will practice importing and using a variety of useful packages for Python including Pandas (for data manipulation and analysis), NumPy (for arrays and mathematical functions), Matplotlib (for producing visualizations), Seaborn (for producing visualizations), Requests (for making HTTP requests), BeautifulSoup (for parsing HTML and XML documents), Scikit-learn (for machine learning), NLTK (for natural language processing), Gensim (for natural language processing), Polyglot (for natural language processing), and SpaCy (for natural language processing); (4) **Data collection**: we will learn about data collection methods and practice using programming packages and tools such as Requests, BeautifulSoup, and Octoparse to collected data from a variety of online resources including Amazon, Facebook, Google Maps, Indeed, Twitter, Wikipedia, Yelp, and YouTube; (5) **Data cleaning**: we will learn about the importance of data cleaning and practice techniques to address common issues with collected data, such as fixing whitespace and capitalization inconsistencies in category labels, collapsing multiple categories into one, removing duplicates, and reformatting strings for consistency; (6) **Image classification**: we will learn about image classification and practice detecting object shapes with edge detection filters, finding elements in images by their contours, improving image quality with contrast enhancement, and restoring images by removing objects and text from pictures; (7) **Natural language processing**: we will learn about natural language processing (NLP) techniques

and practice accessing text corpora and lexical resources, processing raw text, writing structured programs, categorizing and tagging words, classifying text, extracting information from text, analyzing sentence structure, building feature-based grammars, analyzing the meaning of text, and analyzing the sentiment of text; (8) **Calculating statistics**: we will review and practice fundamental statistical techniques for working with data and finding results, including how to calculate the mean, median, percentile, standard deviation, variance, covariance, and Pearson correlation coefficient; (9) **Producing visualizations**: we will learn about visualization techniques and practice producing a variety of different kinds of visualizations including box-and-whisker plots, dendrograms, histograms, scatter plots, and word clouds; (10) **Ethics of machine learning**: we will learn about major ethical theories, including deontological ethics, utilitarian ethics, and virtue ethics, as well as significant ethical issues in machine learning, including fairness and bias, privacy and transparency, and standards for auditing and accountability; and (11) **Machine learning project workflow**: we will review the basic workflow for machine learning projects, including data collection, data cleaning, data analysis, training and testing models, producing visualizations, calculating summary statistics, and reporting the main results. Throughout this course we will practice completing at least 7 different kinds of machine learning projects including (a) using data from Google Maps to determine the best rated gyms in California, (b) using data from Facebook to determine the most liked and reshared posts by the UFC, (c) using data from Twitter to determine the most liked and retweeted posts by Science Magazine, (d) using data from NeurIPS to determine the most popular topics in AI and machine learning, (e) using data from IMDb and Wikipedia to determine movie similarities based on plot summaries, (f) using data from MNIST to build a classifier for images of letters (A, B, C, etc.) in American Sign Language (ASL); and (g) using data from Echonest to build a classifier for genres of songs (Hip-Hop, Rock, etc.) on Spotify. By completing a variety of different kinds of machine learning projects together in this course, students will gain the real-world knowledge and practical skills required to complete their own original projects by the end of the semester.

Outcomes: In this course students will gain (a) the theoretical knowledge to discuss key concepts and ethical implications of machine learning, and (b) the practical skills to complete a full machine learning project workflow from beginning to end. Students completing this course will develop a data-driven approach to answering practical questions and gain coding proficiency with the Python programming language. Each student will demonstrate the practical machine learning skills that they have developed in this course with their own unique final project, where they raise a practical question (such as, "which movie is most similar to *The Dark Knight*?") and aim to answer that question by collecting relevant data (for example, movie plot summaries from IMDb and Wikipedia), cleaning and manipulating the data, training and testing models, calculating summary statistics, producing visualizations, and discussing results.

Required Reading: The readings will be available for you on CatCourses.

Assignment Submissions: Submit your assignments by uploading them directly onto CatCourses.

Grading Procedures: Your grade for this course will be based on your performance on weekly assignments (70%) and a final project (30%). Assignments will include a combination of multiple choice questions, short essays, and coding exercises. Instructions and a grading rubric will be provided for each assignment.

Academic Integrity: Each student must abide by the Academic Honesty Policy at UC Merced. You must do all of your own work on all assignments and copying is never allowed. Violations of academic integrity will result in disciplinary action.

Accommodations for Students with Disabilities: The University of California is committed to ensuring equal opportunities and inclusion for students with disabilities based on the principles of independent living, accessible universal design, and diversity. The University of California requests for academic accommodations to be made during the first three weeks of the semester, except for unusual circumstances, and students are encouraged to register with the Disability Services Center to verify their eligibility for appropriate accommodations. I am available to discuss appropriate academic accommodations that may be required for students with disabilities, so if you have any questions about this please feel free to ask.

Additional Remarks: This syllabus is tentative and subject to change so stay tuned for updates. If you have any questions or want to talk more about the course, majoring in philosophy, or your future career, I encourage you to visit me during office hours for a chat. I value your contributions to the course and I look forward to seeing you develop this semester.

Readings

Bird, S., Klein, E. & Loper, E. (2009). **Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit**. Sebastopol: O'Reilly Media. http://www.nltk.org/book

Caplan, R., Donovan, J., Hanson, L. & Matthews, J. (2019). **Algorithmic Accountability: A Primer**. New York: Data and Society. https://datasociety.net/wp-content/uploads/2019/09/DandS_Algorithmic_Accountability.pdf

Jurafsky, D. & Martin, J. H. (2020). **Speech and Language Processing** (3rd Edition). London: Pearson Education. https://web.stanford.edu/~jurafsky/slp3

Muller, V. C. (2020). Ethics of artificial intelligence and robotics. **Stanford Encyclopedia of Philosophy**. https://plato.stanford.edu/entries/ethics-ai