07/20/17
Subject to changes and revisions

University of California, Davis
Graduate School of Management
MSBA 454: Machine Learning
Winter 2018
(Three-Unit)

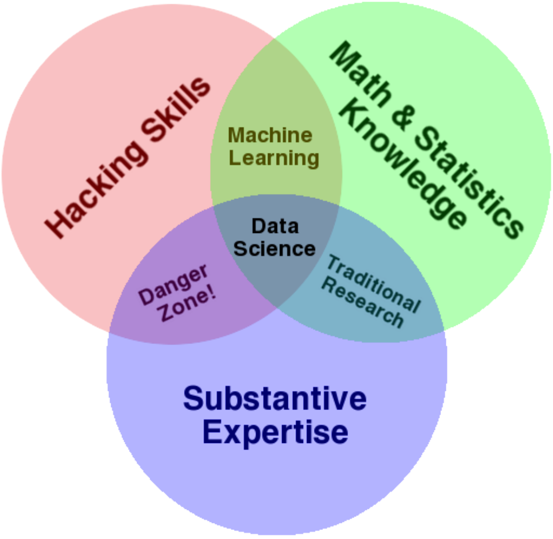

Diagram: Drew Conway

Syllabus

Instructor: Noah Gift
Lecturer
Phone: 415-300-7069 (cell)
E-mail: ngift@ucdavis.edu
No formal office hours. By appointment.
(I can usually see you before or after class if you let me know in advance)

Course Description:

During this course, every student will have developed a rich portfolio of work in Machine Learning stored in Github and Juypter Notebook. As a team, you will have developed a production Machine Learning Model. The expectation is that for a 10-week period students will immerse themselves in Data Science, ML and coding. At the end of the course, a student will be able to demonstrate their proficiency in Machine Learning theory and practice by developing several Jupyter Notebooks. They Jupyter Notebooks can serve as a portfolio in applying for your next position, or a way of demonstrating your skills to your current employer.

Guest Lecturers:

Guest Lecturer 1: Mario Izquierdo, Core Member Twitch API Team, 1/19
Guest Lecturer 2: Gary Knight,

Course Objectives:

- Learn how to promote your brand as a Data Scientist using Github. Kaggle, Juypter Notebook and Bokeh, Plotly,Shiny, Rmarkdown.
- Learn to apply Machine Learning techniques to both well-known data sets and new datasets: Unsupervised learning, Supervised learning, Deep Learning, Distributed Machine Learning with Spark
- Learn to be self-sufficient on the AWS Cloud as a Machine Learning Engineer: collect and clean your own data, and create production ML models.
- Understand how to do machine learning in both R and Python (and potentially any language).
- Apply your knowledge in course to create a reproducible portfolio project that highlights your ability to perform as a Machine Learning practitioner.
- Make Machine Learning Contacts Who May Help You for Your Entire Career

Required readings and online resources:

- Sebastian Raschka & Vahid Mirjalili, "Python Machine Learning", Second Edition, Packt, 2017
- Norman Matloff, "Statistical Regression and Classification", CRC Press, 2017
- Jake VanderPlas, https://jakevdp.github.io/blog/2017/03/03/reproducible-data-analysis-in-jupyter/
- Jake VanderPlas, "Python Data Science Handbook: Essential Tools for Working with Data", O'Reilly, 2016: https://github.com/jakevdp/PythonDataScienceHandbook
- Hadley Wickham & Garrett Grolemund, "R for Data Science", http://r4ds.had.co.nz/, O'Reilly, 2016.
- Ben Lorica , Mike Loukides, "What are machine learning engineers?"https://www.oreilly.com/ideas/what-are-machine-learning-engineers, O'Reilly
- Omar El Gabry, "A Journey through Titanic",https://www.kaggle.com/omarelgabry/a-journey-through-titanic, Kaggle

Optional but Recommended Readings, Media and Notebooks

- Yaser S. Abu-Mostafa, Malik Magdon-Ismail, Hsuan-Tien Lin, "Learning From Data", AMLbook.com, 2012
- https://work.caltech.edu/telecourse.html
- Max Kuhn, Kjell Johnson, "Applied Predictive Modeling", Springer, *2013*

- Andreas Mueller, “Introduction to Machine Learning With Python”, https://github.com/amueller/introduction_to_ml_with_python
- https://github.com/jupyter/jupyter/wiki/A-gallery-of-interesting-Jupyter-Notebooks#introductory-tutorials
- Talking Machines: Getting a Start in ML and Applied AI at Facebook, http://pca.st/84iv, July 13th, 2017.
- Joaquin Quiñonero Candela, University of Cambridge, Machine Learning: http://mlg.eng.cam.ac.uk/teaching/4f13/1112/
- David MacKay, http://www.inference.org.uk/itprnn/book.pdf, Cambridge University Press, 2003
- Andrew Ng: https://www.coursera.org/learn/machine-learning, Coursera, 2011
- Tyler Cowen, “Average is Over”, Dutton, 2013
- https://software-carpentry.org/lessons/
- John Markoff, “Machines of Loving Grace”, Harper Collins, 2015

Machine Learning **Regression** Group Project (40% of Grade)

Teams of four to five people will be formed. Your team will act like the Data Science team at a startup and deliver a machine learning **regression** model. I would recommend you begin work on this project during your first week (first day of class), have weekly sprints and target an MVP that fits all requirements by Week 5. If your group opts to go for Bonus Opportunity you could work weekly on bonus opportunity features during Week 6-10.

A sample machine learning regression project is in the course Github repository that fulfills these requirements. Please study this to ensure you are meeting expectations.

To receive full credit for the Group Project you must fulfill these requirements:

Code

- All members of the team must commit code on the project

Machine Learning

- Must use a **Regression** technique
- Must deploy model you created (not just calling out to Azure, AWS, or Google AI API).
- Must use a real data set (not fake data set, i.e. randomly generated numbers)

Juypter Notebook Documentation

- Must explain in detail accuracy of model
- Explain the tradeoffs you made
- Do exploratory data exploration

Bonus Opportunity (Up to 10% of Grade)

For groups that want to go above and beyond ***(truly exceptional effort)*** on the project you may receive bonus credit. Here are some ideas of what may be considered partial or full Bonus Credit.

- Build a Data Visualization Front End to Your API using Bokeh, Plotly, Shiny, or something similar.
- Build a true MVP you could pitch to a startup accelerator or Angel Investor in its current state.
- Incredible software engineering craftsmanship: monitoring, logging, load-testing, etc.
- Incredible Machine Learning craftsmanship and incorporation of class learnings and beyond.
- Exceptional showmanship in presentation
- You have an API deployed using Chalice/Flask or even R and the class can query it

Machine Learning **Classification** Individual Project and Presentation (40% of Grade):

Every student will give a five-minute presentation in class on a machine learning **classification** project. The format will be as follows:

- Recommended: Juypter Notebook, R Markdown. Optional: (Bokeh or Shiny)

To receive full credit for the Individual Project you must fulfill these requirements:

Machine Learning

- Must use a **classification** technique
- Must deploy model you created (not just calling out to Azure, AWS, or Google AI API).
- Must use a real data set (not fake data set, i.e. randomly generated numbers)

Juypter Notebook Documentation

- Must explain in detail accuracy of model
- Explain the tradeoffs you made
- Do exploratory data exploration

Both Individual and Group Project will be judged according to criteria (additional points available for Group, see above):

- Reproducibility of the Project: Is every step reproducible in Juypter Notebook or R Markdown Studio?
- Creativity: Did you create new insights that you shared to the class?
- Craftsmanship: Did you take care in crafting your presentation and put thought into little details like thoughtful data visualization?
- Difficulty: Was this a challenging problem you tackled?

- ML Technique: How did you approach the problem?
- Software Engineering Carpentry: Did you apply software engineering best practices?
- Resume Worthiness: Will this project get you a job?

Individual Evaluation:

- Consistent with the Graduate School of Management's policy for three-unit courses, this is a graded, for-letter course.
- Attendance for all ten sessions is required.
- The course will be graded on a curve.
- Individual evaluation is broken down as follows:

Grading

- Homework: 10%
    - Due before class
    - Checked into Github
    - Full Credit or No Credit (If Incomplete or Late)
    - Is Individual
- Quiz: 10%
    - Lowest Quiz thrown out
    - Beginning of every class
    - On pre-class readings and last lecture
    - Pre-class Assessment can replace a quiz as well
- Group Project: 40%
- Individual Project: 40%
- Bonus: 10%

A, A+:	Performed above expectations
B+, B:	Fully met expectations of course
C+ and lower:	Some aspects may not meet expectations

About the Instructor:

Noah Gift is a consulting CTO, Cloud Architect and Data Scientist who worked at companies such as: Linden Lab, Loggly, AT&T Interactive, Weta Digital, Turner Studios, Sony Imageworks, Disney Feature Animation, Caltech and ABC Network News. Noah has close to 25 years professional experience in Media and Technology starting at 18 at ABC Network News as a National News editor and most recently as consultant who builds teams, technology, products and revenue. He has built production machine learning models in R and Python that predicted successfully with millions of dollars on the line.

Noah has MBA from UC Davis and a MS in Computer Information Systems from CSULA. He is the co-author of the O'Reilly book, "Python for Unix for and Linux Systems Administration", as well as Pragmatic AI. He has published over 50 articles for many top publications including IBM, Red Hat and O'Reilly. Noah is a member of the Python Software Foundation and has been using Python professionally since 2000.

COURSE SCHEDULE

Session One (January 5th, 2018: 5:45PM-8:30PM):

Pre-class readings:

- Jake VanderPlas, "Python Data Science Handbook: Essential Tools for Working with Data", O'Reilly, 2016: https://jakevdp.github.io/PythonDataScienceHandbook/ Chapter 1,2
- Study Kaggle Notebook and Be Prepared to Work Through In Class: https://www.kaggle.com/omarelgabry/a-journey-through-titanic

In Class:

- Class Introductions
- Course, Projects and Teams Explained
  - Project Teams Finalized
- Lecture: **Machine Learning Software Carpentry**
  - **IPython/Juypter Notebook**
  - **Numpy**
  - **Pandas**
  - **Github**
  - **Flask/Chalice Hello World (bonus Material)**
- Project Team Case Study:
  - **Titanic: Machine Learning from Disaster: https://www.kaggle.com/c/titanic**
- Quiz 1:

HW Assignment #1: *Create Your Own Analysis of Titanic Dataset and submit prediction to Kaggle. Store a copy of your Juypter notebook in Github.*

Session Two (January 12th, 2018: 5:45PM-8:30PM):

Pre-class readings:

- https://www.ibm.com/developerworks/library/ba-social-influence-python-pandas-machine-learning-r-1/
- https://www.ibm.com/developerworks/analytics/library/ba-social-influence-python-pandas-machine-learning-r-2/index.html
- https://www.kaggle.com/noahgift/nba-team-valuation-exploration
- https://www.kaggle.com/noahgift/nba-player-power-influence-and-performance

- Sebastian Raschka, “Python Machine Learning”, Packt, 2017, Chapter 1
- Jake VanderPlas, “Python Data Science Handbook: Essential Tools for Working with Data”, O’Reilly, 2016: https://jakevdp.github.io/PythonDataScienceHandbook/ Chapter 4,5

In Class:

- Quiz 2
- Review HW Assignment #1
- Lecture: **Machine Learning Overview**
  - **What is ML?**
  - **Supervised ML**
  - **Unsupervised ML**
  - **Classification**
  - **Regression**
- Lecture: **Breakdown of a Production Machine Learning Project- Social Media Influence, Salary and Performance in the NBA (API and Juypter Notebook)**
  - This project will be an example project to refer to in building Group Project and Individual Projects

HW Assignment #2: Fork notebooks on Kaggle and create your own data exploration and submit your notebooks to Kaggle.

Session Three (January 19th, 2018: 5:45PM-8:30PM):

Pre-class readings:

- Hadley Wickham & Garrett Grolemund, “R for Data Science”, http://r4ds.had.co.nz/, O’Reilly, 2016. Chapter 1,3, 27, 28, 29, 30
- Sebastian Raschka, “Python Machine Learning”, Packt, 2017, Chapter 4
- Jake VanderPlas, “Python Data Science Handbook: Essential Tools for Working with Data”, O’Reilly, 2016: https://jakevdp.github.io/PythonDataScienceHandbook/ Chapter 3

In Class:

- Quiz 3
- Review HW Assignment #2
- Lecture: **Data Manipulation and Transformation**
  - **Pandas Operations**
  - **Cleaning Data**
- Lecture: **Visualization**
  - **Matplotlib**
  - **Seaborn**
  - **ggplot in R and Python**
- Project Team Case Study: Creating test and training sets with Wine dataset
- Mario Izquierdo***, Core Member Twitch API Team:*** Guest Lecture: 8-8:30

HW Assignment #3:

- Create a faceted plot in ggplot of a public data set use size, shape and color as well as facts.
- Create a Correlation Heatmap in Seaborn using a public dataset.
- Create your own Test and Training sets using a public dataset.

Session Four (January 26th, 2018: 5:45PM-8:30PM):

Pre-class readings:

- Norman Matloff, "Statistical Regression and Classification", CRC Press, 2017, Chapters 1-4.
- Sebastian Raschka, "Python Machine Learning", Packt, 2017, Chapter 10

In Class:

- Quiz 4
- Review HW Assignment #3
- Lecture: **Regression**
- Project Team Case Study: Predicting Home Prices

HW Assignment #4: Statistical Regression and Classification", CRC Press, 2017: pg. 61, do 1.22 Exercises 1.22: 1-4, pg. 120: 2.14: 1-4. (You can use Python or R)

Session Five (Feb 2th, 2018: 5:45PM-8:30PM):

Pre-class readings:

- Sebastian Raschka, "Python Machine Learning", Packt, 2017, Chapter 2,3
- Norman Matloff, "Statistical Regression and Classification", CRC Press, 2017, Chapters 12

In Class:

- Quiz 5
- Review HW Assignment #4
- Lecture: **Classification**
- Project Team Case Study: Classification of Sex of Abalone

HW Assignment #5:

- "Statistical Regression and Classification", CRC Press, 2017: pg. 450, 12.5: 1-5. (You can use Python or R)
- Create your own classification model of the Iris data set in Juypter notebook. Include a visualization of the decision boundries.

Session Six (Feb 9th, 2018: 5:45PM-8:30PM):

Pre-class readings:

- Sebastian Raschka, "Python Machine Learning", Packt, 2017, Chapter 11

In Class:

- Quiz 6
- Review HW Assignment #5
- Lecture: **Unsupervised Learning and Clustering**
- Individual Project Presentations: 1st 25%.
- Gary Knight, ***Digital Strategy Consultant***: Guest Lecture: 8-8:30pm

HW Assignment #6:

- Create kNN clustering Juypter Notebook of a public dataset.
- Create a faceted ggplot of a kNN cluster of a public dataset

Session Seven (Feb 16th, 2018: 5:45PM-8:30PM):

Pre-class readings:

- Read through source code of Surprise Recommendation Framework: https://github.com/NicolasHug/Surprise

In Class:

- Quiz 7
- Review HW Assignment #6
- Lecture: **Recommendations**
- Individual Project Presentations: 2nd 25%.

HW Assignment #7:

- Implement your own recommendation system on a public data set in a Juypter notebook using Surprise

Session Eight (Feb 23th, 2018: 5:45PM-8:30PM):

Pre-class readings:

- Sebastian Raschka, “Python Machine Learning”, Packt, 2017, Chapter 12,13
- Norman Matloff, “Statistical Regression and Classification”, CRC Press, 2017, Chapters 12
- Tensor Flow Image Recognition Tutorial: https://www.tensorflow.org/tutorials/image_recognition#usage_with_python_api

In Class:

- Quiz 8
- Review HW Assignment #7
- Lecture: **Neural Networks**
    - **Tensorflow**
    - **Pytorch**

- Individual Project Presentations: 3rd 25%

HW Assignment #8:

- Implement your own Tensorflow based Image Recognition Juypter Notebook

Session Nine (March 2nd, 2018: 5:45PM-8:30PM):

Pre-class readings:

- Sebastian Raschka, “Python Machine Learning”, Packt, 2017, Chapter 8

In Class:

- Quiz 9
- Review HW Assignment #8
- Lecture: **Sentiment Analysis**
- Individual Project Presentations: 4th 25%.

HW Assignment #9:

- Implement your own sentiment analysis of a public dataset Jupyter Notebook

Session Ten (March 9th, 2018: 5:45PM-8:30PM):

Pre-class readings:

- Explore Notebooks on PySpark: https://github.com/jadianes/spark-py-notebooks

In Class:

- Quiz 10
- Review HW Assignment #9
- Lecture: **Distributed Machine Learning**
- Demo Day: ~ 10-12 Group Project 5 Minute Demos