

DATA 301

Keith Perkins

Outline

- Where is Course Content?
- Syllabus
- Evaluation
- Who is this course for?
- Topics (a tentative list)

Course Content

<https://cnuclclasses.github.io/DATA301/>

Bookmark it, almost all content will be here

Syllabus

<https://cnuclases.github.io/DATA301/>

Go to Syllabus tab

Required textbook: None

Suggested Texts:

Python for Data Analysis, Wes McKinney

Machine Learning with Python Cookbook, Chris Albon – common task recipes

Evaluation

- As per syllabus:
 - 2 exams
 - 1 final exam
- Multiple projects
 - (up to 5)

Who is this course for?

- Anyone who works with data (social sciences, biology, psychology, law, medical etc.)
- This course will go over some techniques for using data to:
 - Identify data clusters
 - Make predictions
 - Find anomalies (maybe)
 - Determine what part of the data has greatest influence on output
- This course is an introduction. There is a LOT more to the field.

Topics?

- Let's see what others suggest
 - search 'data science syllabus'
- Seems a bit much
- How about if we just do an introduction?

Topics

- Week 1
 - Course outline, pre-reqs
 - goals
 - general project workflow
- Week 2-3
 - Data preprocessing, cleaning, EDA, leakage
- Week 4-5
 - Unsupervised algorithms
 - Clustering
- Week 6
 - Splitting the dataset
 - Dataset imbalance
 - Leakage
- Week 7-9
 - Decision Trees
 - Random Forest
 - Algorithm evaluation

Topics

- Week 10
 - Gradient Boosted Trees
 - What tree based algorithms cannot do that regressions and Neural Networks can do
- Week 11
 - Explainability
- Week 12
 - Neural Networks
 - Hyperparameters
 - Cross validation
- Week 13-14
 - Recommender systems
 - Ensembling
 - Topics

Where to go after this course?

- SQL – there is a lot of data in databases
 - Do you need a Database class?
 - Not for this course.
 - Professionally, it's a little trickier. You can learn what you need in a week or so and get pretty good after a month or 2.
- Do you need to scale your compute?
 - Not for this class
 - Professionally, yes. You don't use a laptop. At a minimum use a local desktop with a GPU (or GPUs)
 - What if your data is huge and will not fit in memory? Next step is something cloud based (pyspark, Dask..etc this area is changing fast)
- Time series data (store sales by day, NLP etc.)
- You have to learn to use Linux
- Start creating your own projects