Advanced Practical Course In Computer Networking

Introduction Session – April 18 2016

David Koll

What is this course about?

- „Data Science“ is a major current buzzword
- Wide applications in both academia and industry

- Many, many, many use cases
- Almost as many algorithms as use cases

What is this course about?

- In this course: apply algorithms that you have learned in other courses to real-world datasets as you will need to as a data scientist

- Will help you to know what kind of work you can expect
- Will help you to learn to independently work on such problems

- We will not teach any algorithms here!
- We will also not teach any API here!

Prerequisites [1/2]

- As stated on the course website: *you know your data science algorithms.*
- In particular, make sure you have a good understanding of **regression and classification methods.**
  - E.g.: Linear regression, lasso regression, logistic regression, decision trees, boosting methods, …
  - Terms like coefficients, precision/recall, feature selection, feature engineering have some meaning to you

Prerequisites [2/2]

- **Programming knowledge:** any of Python, Matlab, R, …
- Ideally: knowledge of data science tools/APIs
  - E.g.: Scikit learn, Pandas, Numpy, Seaborn, Graphlab, Ipython Notebook, etc. in Python
  - E.g.: Caret, ggplot2, etc. in R

- It will be entirely your choice which language/tool you use to solve the tasks
  - Recommended: Python
  - If you consider yourself very fit in one of the tools, maybe use this course to get familiar with another one

Warning

- Do **NOT** take this course if there is too much unknown on the previous slide.
- Reason: You will not have enough time to catch up with everything
- It is fine if you do not know one or two algorithms or techniques

Some hints…

- *If you are very fit in theory, but haven‘t done practical work:*

- Look at Python in combination with…

  - Pandas/Numpy or SFrame/Numpy for data handling
  - Scikit learn or Graphlab Create for training your models
  - Seaborn for data visualization

- There are lots and lots of tutorials in the web!

Tasks and Grading [1/2]

- As stated on the course website, there will be three tasks in total:

- First task: warmup analysis of a simple dataset (10%)
  - Data is cleaned
  - Data is easy to understand
  - Find insights in dataset
  - Use insights to come up with a simple predictor
  - Need to successfully complete in order to register for the course

Tasks and Grading [2/2]

- Second/Third tasks: more challenging (35% each)
  - Data may be „dirty“ and harder to understand
  - Problems to be solved can be much more challenging
  - Intended: one classification and one regression problem

- Finally, written report (20%)

Submission of Work

- First task: submit PDF and code to TA
- Second/Third tasks: presentations
- Written report contains your analysis and predictor code, and a summary (i.e., discussion of your results) of tasks 2 and 3

Role of Instructor/TA

- You should predominantly work on your own
- We can provide you with basic guidance if you run into trouble
- We are not there to give you solutions
  - Remember all teams work on the same topics – need to keep fairness up!