COLUMBIA UNIVERSITY
Data Science Institute

Welcome to the Data Science Bootcamp

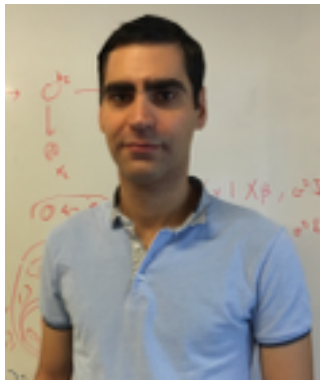
January 9-13th, 2016

COLUMBIA UNIVERSITY
Data Science Institute

Who we are

Kriste Krstovski

James McInerney



Francisco Ruiz

COLUMBIA UNIVERSITY
Data Science Institute

Bootcamp Schedule

- Monday: Introduction to Data Science
- Tuesday: Algorithms & Classification
- Wednesday: Machine Learning
- Thursday: Advanced Topics in Data Science
- Friday: Capstone Project

Monday: Introduction to Data Science

- Bootcamp outline (30 min talk)
- Data summarization and visualization (15 min talk)
- Introduction to probability (15 min talk)
  - Lab (2 h): 2016 presidential election results, NYC Open Data project
- Linear regression (30 min talk)
  - Lab (1 h): linear regression on synthetic data

Tuesday: Algorithms & Classification

- Introduction to machine learning (10 min talk)
- Logistic regression (20 min talk)
- Clustering (30 min talk)
  - Lab (2 h): logistic regression (mushroom data), clustering (NYPD vehicle collisions)
- Gradient descent (30 min talk)
  - Lab (1 h): optimizing multivariate functions

Wednesday: Machine Learning

- Probabilistic modeling (30 min talk)
- Topic models (30 min talk)
  - Lab (2 h): LDA on NYT articles, ML/MAP on breast cancer data
- Evaluation approaches (30 min talk)
  - Lab (1 h): computing evaluation measures on various model outputs

Thursday: Advanced Topics

- Neural networks (1 hour talk)
  - Lab (2 h): deep neural networks in TensorFlow (MNIST dataset)
- Language models and vector space model (30 min talk)
  - Lab (1 h): n-grams on books, tf-idf weighting

Friday: Capstone Project

- Option of one of several projects for individual study:
  - Classification methods
  - Convolutional neural networks
  - Dimensionality reduction
  - Hidden Markov models
  - Polylingual topic models
- Short presentations/discussions at the end of the day

Friday

- Course Assessment
- Certificate of Completion

Parameters of the Bootcamp

- There are 2 hours of mandatory preparation for each day:
  - Combination of online videos and assigned readings
- Ask questions, at any time: it will help everyone
- Can you open the sample IPython notebook in the project folder?
  - If not, please get help immediately (don't wait to watch the rest of this talk)

Goals

- Familiarize yourself with various approaches across different data science domains
- Be able to utilize the knowledge gained in your own research
- Motivation to further explore approaches that may be useful for your current or future research projects

Projects

- Carried out on Friday
- Extend on a topic covered during Monday-Thursday
- We propose 5 projects to choose from

Projects

1. Work individually on a project

2. Work in groups to discuss the results

3. One person from each group presents to the class

Projects

- Classification methods
  - Goal: Learn new techniques for binary and multiclass classification
  - Techniques: Random forests, KNN
  - Evaluate the performance according to different metrics

- Dimensionality reduction
  - Goal: Learn techniques to reduce the dimensionality of a high-dimensional dataset
  - Techniques: PCA, ICA, LSA, LDA
  - Evaluate the quality of the low-dimensional representation

- Time series: Hidden Markov models
  - Goal: Learn to model datasets with temporal dependencies
  - Techniques: Probabilistic modeling (hidden Markov models)
  - Implement inference and assess the results

Projects

- Polylingual topic models
  - Goal: Learn a probabilistic model for documents in different languages
  - Techniques: Probabilistic modeling (based on latent Dirichlet allocation)
  - Exploratory data analysis

- Convolutional neural networks for image classification
  - Goal: Learn about the architecture of ConvNets for multiclass image classification
  - Techniques: Implementation of ConvNets on Tensorflow
  - Evaluate the performance and compare different schemes

End of Introduction

- Enjoy the course!

- Many thanks to the other members of the organizing committee:
  - Prof. Patricia Culligan, Dr Andreas Mueller, Prof. Tian Zheng, Jonathan Stark

- Email any of us:
  - Kriste (kriste.krstovski@columbia.edu)
  - James (contact@jamesmc.com)
  - Francisco (f.ruiz@columbia.edu)

COLUMBIA UNIVERSITY
Data Science Institute