Wrapping Up and Next Steps

Ivan Corneillet

Data Scientist

Learning Objectives

After this lesson, you should be able to:

- Summarize the different steps of the data science pipeline
- Summarize the models we learned in this class and select the appropriate model for your problem at hand
- As next steps, which models you should learn on your own

End-of-Course/Projects Countdown

Final Project, Part 5	July 7; due next session!

Announcements and Exit Tickets

Q & A

Review

Review

Introduction to Databases

Review

- Databases are often at the core of any data analysis. Most analysis starts with retrieving data from a database
- SQL (Structured Query Language) is a key language that any data scientist should understand
  - `SELECT`: Used in every query to define the resulting columns
  - `WHERE`: Filters rows based on a given condition
  - `GROUP BY`: Groups rows for aggregation
  - `JOIN`: Combines two tables based upon a given condition
- *pandas* can be used to access data from databases as well

Today

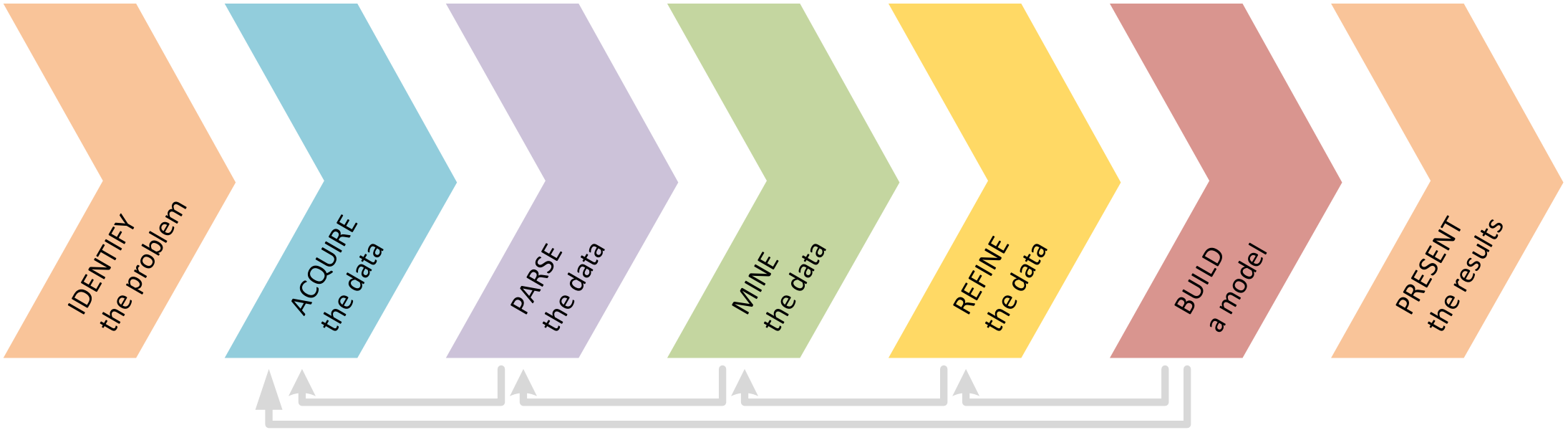
Here's what's happening today:

- Alumni Panel
- Review
- Data Science Workflow
- What models did we learn in this class? Which one should I choose? Pros and Cons?
- What else did we learn in this class?
- What's next?
- Office hours in class for final projects

Wrapping Up and Next Steps

The Data Science Workflow

The Data Science Workflow (one last time…)



What models did we learn in this class? Which one should I choose? Pros and Cons?

What models did we learn in this class?

Regression

- Linear Regressions (simple and multiple)
- k-NN (K-Nearest Neighbors)
- Regression Decision Trees and Random Forests
- AR, MA, ARMA, and ARIMA (for time series)

Classification

- Logistic Regression
- k-NN (K-Nearest Neighbors)
- Classification Decision Trees and Random Forests

What model should I use?

- Ask yourself the following questions:
  - Do I have an output or not?
    - If yes, you need to use a supervised learning technique, otherwise, you will use an unsupervised technique
      - In this class we focused on supervised learning techniques. (The exception was the Principle Component Analysis we briefly mentioned)
  - Assuming you have a supervised learning problem, is your output a quantitative variable or qualitative
    - If it is quantitative you will use one of the regression methods, otherwise you will use a classification algorithm

What model should I use? (cont.)

- Is your goal interpretation or prediction?
  - Interpretative regression models are:
    - Linear Regression
    - Simple Regression Decision Trees
  - Predictive regression models are:
    - KNN (with a lower-dimension feature space)
    - Regression Random Forest
    - AR, MA, ARMA, and ARIMA. Low order of models are relatively interpretable but higher order are not. This is mainly why they are usually only used for prediction

What model should I use? (cont.)

- Interpretative classification models are:
  - Logistic Regression
  - Simple Classification Decision Trees
- Predictive classification models are:
  - Classification Random Forest

Pros and Cons of Linear Regression

- Pros
  - Intuitive
  - Very interpretable
  - Easy to compute predictions
  - No need to standardize your data
- Cons
  - Assumes linear association among variables
  - Assumes normally distributed residuals
  - Outliers can easily affect coefficients

Pros and Cons of k-NN

- Pros
  - Intuitive
  - Very easy to compute
  - Easily capture non-linearity
- Cons
  - Not interpretable
  - Cannot be used if you have sparse data and feature space with dimension of 4 or more
  - Need to standardize your data

Pros and Cons of Logistic Regression

- Pros
  - Fit is fast; faster than Random Forests
  - Output is a (posterior) probability which is easy to interpret
- Cons
  - Limited to binary classification (need to use ensemble for more classes)

Pros and Cons of Simple Decision Trees

- Pros
  - Intuitive
  - Very interpretable
  - Easy to compute predictions
  - No need to standardize your data
- Cons
  - Low predictable power

Pros and Cons of (Fully-Grown) Decision Trees and Random Forests

- Pros
  - Good predictive power
  - Easy to compute predictions
  - No need to standardize your data
- Cons
  - Not interpretable

Pros and Cons of AR, MA, ARMA, ARIMA

- Pros
  - AR – good for smoothing patterns
  - MA – good for tackling shocks
  - ARMA – good for smoothing patterns and tackling shocks
  - ARIMA – good for smoothing patterns and tackling shocks; also takes care of linear trends in the model
- Cons
  - Not that interpretable (except when p and q are small)

What else did we learn in this class?

What else did we learn in this class?

- We discussed how important it was to tidying up data
  - Tidying data is one of the most fruitful skill you can learn as a data scientist. It will save you hours of time and make your data much easier to visualize, manipulate, and model

What else did we learn in this class? (cont.)

- Over the course, we improved our Python fluency
  - *pandas* dataframes and other Python data structures (e.g., dictionaries)
  - Loops and conditionals
  - Write basic functions to simplify our life and avoid code duplication (e.g., transforming variables on the training set then on the testing set)
  - Primer on object oriented programming

What else did we learn in this class? (cont.)

- We are no longer afraid of statistics! The following should feel very familiar now:
  - Two-Tail Hypothesis Testing
  - Normal, Student's t-, and F-distributions
  - z-scores, t-values, and p-values

What else did we learn in this class? (cont.)

- Validation
  - Divide your dataset into a train and a test sets. Train with training Data and Test it with test data. If you have a large dataset, using validation is always preferred
- Cross-Validation
  - Divides data into chunks. Then train your model on all groups but one, and then test it on the one chunk left out. Repeat on all groups. This way, you are not wasting any data. Especially useful when you have a small dataset
- Can we combine Validation and Cross-Validation?
  - Not only can you, but you should… Develop your model and tune it up using cross-validation (on your training data). Then train a final model using all your training data and then test it on your test set (completely unseen data)

What else did we learn in this class? (cont.)

- Git/GitHub
  - GitHub has become such a staple amongst the open-source development community that many developers have begun considering it a replacement for a conventional resume and some employers require applications to provide a link to and have an active contributing GitHub account in order to qualify for a job

What’s next?

What's next?

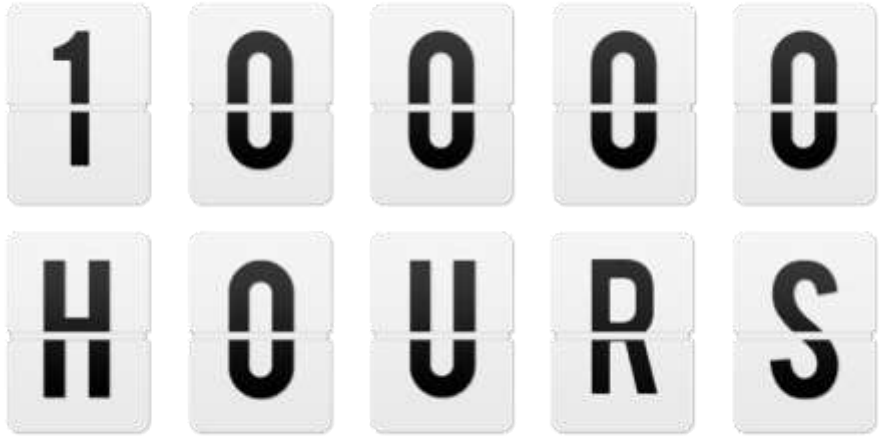
- A lot!
- This course was just an introduction to data science
- We focused on learning just a handful on models but learning them well. There are of course many more…

In short term, consider spending time learning or doing a deep dive on the following machine algorithms:

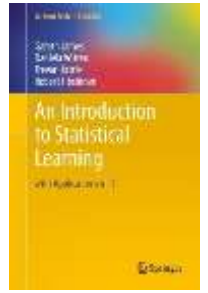
- Principal Component Analysis (a.k.a., PCA) for dimensionality reduction as a pre-processing step for other machine learning algorithms
- Supervised
  - Regularization (for linear regression and logistic regression)
  - Boosting (e.g., on Regression/Classification Decision Trees)
- Unsupervised
  - k-Means Clustering

Get your Hands Dirty, a.k.a., Practice, Practice, Practice…

- Kaggle (http://www.kaggle.com) competitions are a great way to practice everything we've learned in this class. And it's fun too!
  - Azi, Jeremiah, and Ivan are spending way too much time in this site…
  - You can compete by yourself but your can also team up with your fellow GA classmates!
- If Kaggle is not your thing, you should consider joining a study group if you haven't done so already



Longer term, consider the following resources



- An Introduction to Statistical Learning: with Applications in R (by James et al.). The e-book is available free-of-charge here

- A MOOC (Massive Open Online Courses) called Statistical Learning covering the book above is usually offered by Stanford also free-of-charge once a year during the winter. (now self-paced!) Check it out here

- For a more advanced treatment of these topics, check out The Elements of Statistical Learning: Data Mining, Inference, and Prediction (by Hastie et al.). And yes, the e-book is also free… (here)

Q & A

Next Classes

Final Project Presentations

Part 1 on 7/8, Part 2 on 7/12

Exit Ticket

Don’t forget to fill out your exit ticket here