Daily Lesson Plans

In this file is outlined the flow of daily lessons for the boot camp. Each day's topics are briefly summarized with bullet points.

Day 1: Collecting Data

If we are asking students to complete an end to end data science project, they need to be able to find data. This lesson should give them the tools necessary to do that. Here is a brief outline of the flow of the lesson

- Briefly overview various popular online data sites, for example display Kaggle.com, show how you can find data sets, challenges, and kernels.
- Introduce the beautifulsoup package.
- Go through a couple simple beautiful soup examples.
- Explain APIs in a nutshell.
- Give one or two examples of python API packages:
  - One where no security credentials are needed,
  - One that requires security credentials.
- Basic reading and writing of data to file.
- Potential HW:
  - Find an interesting online data set and save it to your laptop.
  - Use beautifulsoup to scrape an online spreadsheet, or save some txt to file.
  - If one of the python APIs interested you practice scraping some data with it.

Day 2: Data Types, Cleaning, and Exploration

We have data. Before we can make predictive models we usually have to clean or pre-process the data. In this lesson we overview common cleaning/exploration procedures.

- Overview of data types:
  - Continuous data,
  - Categorical and Ordinal data,
  - Text data.
- Continuous data:
  - Describing and visualizing
  - Common scaling procedures,
  - Handling missing data.

- Categorical and Ordinal data,
  - Describing and visualizing,
  - Dummy variables,
  - Handling missing data
- General tips and tricks for handling missing data.
- Text data.
  - Differences from numerical data,
  - Python strings built in methods,
  - Common cleaning techniques,
  - Resources for natural language processing.
- When would you encounter messy data?
  - Give examples of "messy" data sets.
- Potential HW:
  - Clean some messy data sets.

Day 3: Supervised Learning - Regression 1
In this lesson we start supervised learning with regression modeling.

- What is supervised learning?
  - Using "labeled" data to teach a machine to generalize trends in data to "unlabeled" data,
  - Two flavors of problem: Regression and Classification.
- What is regression?
  - In a machine learning setting, problems where the "labeled" data has a continuous measure, for example predicting opening weekend ticket sales.
  - Be careful not to confuse with a more statistical definition, for example, logistic regression is a statistical regression model, but is most often used as a classification algorithm
- The most basic model: Simple Linear Regression
  - Introduce model, explain assumptions.
  - Give example using real data set.
  - Introduce Training-Test Split
  - Interpreting results

  - Model validation for Simple Linear Regression

- Potential HW
  - Build SLR model with new data
  - Lay seeds for why you might want "multiple" training-test splits, i.e. cross-validation

Day 4: Supervised Learning: Regression 2

We build on Day 3 by extending the simple linear regression model to multiple predictors and nonlinear relationships between predictor and output.

- More than one predictor: Multiple Linear Regression
  - Introduce model, explain assumptions.
  - Give example, ideally extending example from SLR.
  - Overfitting example with synthetic data.
  - Discuss variable selection algorithms.
  - Additional model validations for MLR.
- More than just linear relationships: Polynomial Regression
- Additional topics for exploration:
  - Handling continuous vs categorical predictors.
  - Data pre-processing.
  - cross-validation
- Potential HW
  - Build the "best" regression model possible for a given data set using the techniques covered to this point

Day 5: Supervised Learning: Regression 3

In our final regression day we cover two additional regression algorithms, these address one solution to overfitting, regularization.

- Overfitting to the extreme
  - Show an example of overfitting with MLR using synthetic random data,
  - Show an example of overfitting with polynomial regression using synthetic data.
- Cost functions
  - Remind students of goal in least squares,
  - If not already defined, define cost/loss functions

- Fighting overfitting
  - Add penalty term to cost,
  - Introduce concept of hyperparameter
- Ridge Regression
- Lasso
- Potential HW:
  - Wrap up regression section by introducing local regression,
  - Introduce idea of logistic regression.

Day 6: Supervised Learning: Classification 1

We will start off classification with a general outlay of the problem. We will then progress through the general classification workflow with logistic regression.

- Explaining the Classification Problem
- Logistic Regression
  - The model,
  - General model validation for classification problems,
  - Precision vs. recall,
  - Confusion matrix,
  - ROC curve
- Nearest Neighbors
- Potential HW
  - Build a logistic regression model,
  - Use CV to tune number of neighbors,
  - Introduce Naive Bayes.

Day 7: Supervised Learning: Classification 2

In today's lesson we learn about tree based methods for classification. We start by building a decision tree model and then branch into random forest models.

- What is the essence of a tree based method?
  - We are segmenting the data to make predictions based on some sort of loss optimization criterion.
  - Called a tree because we can make a tree (in the graph theoretical sense) to represent the various decision rules.

- Classification Decision Trees:
  - The algorithm,
  - Give an example,
  - Advantages and Disadvantages.
- Methods of Improving Decision Trees:
  - Bagging,
  - Pasting,
  - Boosting.
- Random Forest Models
  - The algorithm,
  - Extend decision tree example,
  - Advantages and Disadvantages.
- Potential HW:
  - Build a decision tree model,
  - Build a Random Forest model,
  - Walk through using decision trees for regression tasks.

Day 8: Supervised Learning: Classification 3

In the final day of classification we wrap up with support vector machines. We will also discuss voting methods as a way to include all of the classification algorithms we have discussed.

- SVMs:
  - Separating Hyperplanes,
  - The maximal margin classifier,
  - Support vector classifier,
  - Support machines
- Voting Methods
  - Summarize all techniques up to the point,
  - Explain idea behind the voting method,
  - Give example.
- Potential HW:
  - Solve a classification problem using SVMs,

  - Solve a classification problem using multiple algorithms, compare and contrast them, and then combine them into a single voting algorithm.

Day 9: Unsupervised Learning: Dimensionality Reduction
To start off the unsupervised learning portion of the course we discuss the differences between supervised and unsupervised learning. We then present dimensionality reduction techniques.

- Differences between supervised and unsupervised learning.
- What is dimensionality reduction? Why do it?
  - Reducing the number of variables in our data set,
  - Memory purposes,
  - Algorithm speed,
  - Data Visualization,
  - Feature Extraction.
- Principal Components Analysis (PCA):
  - Concept,
  - The algorithm,
  - Interpretation of results
  - Example for visualization,
  - Example for feature extraction.
- A Sampling of manifold techniques:
  - Isomap,
  - t-SNE.
- Potential HW:
  - Use PCA to extract import features from a dataset,
  - Introduce kernel or sparse PCA,
  - Introduce an additional manifold technique.
  - Work through an anomaly detection example using dimensionality reduction.

Day 10: Unsupervised Learning: Clustering
We complete the unsupervised learning section with a discussion on clustering techniques.

- What is clustering?
  - Grouping of unlabeled data based on some sort of similarity measure.
  - Why might we want to do this?

- k-means clustering:
  - Give algorithm,
  - How to evaluate results,
  - Advantages and disadvantages.
- Hierarchical clustering:
  - The algorithm,
  - Dendograms,
  - Evaluating the results,
  - Advantages and disadvantages.
- DBScan:
  - The algorithm,
  - Evaluating the results,
  - Advantages and disadvantages.
- Potential HW:
  - Apply each of the clustering algorithms to different data sets,
  - Which algorithms perform "best" on what kinds of data sets,
  - Work through a group segmentation example using clustering.

Day 11: Presenting Results

In this day we discuss more advanced plotting techniques that can help participants communicate their results to others. We will also cover any presentation tips using pandas that have not already been covered.

- More advanced matplotlib.
- Diving into seaborn.
- Introduction to interactive plotting in bokeh.
- Odds and ends left to cover with pandas.

Day 12: Leave this day open for spillover from previous days

Note: it is also important for us to cover being able to run code from terminal/command line. So working that in at some point would be good.