CSCE 478/878 Recitation 1 Handout
January 8, 2019

Recitation Details

- **Time**: Tuesdays 3:30-4:45
- **Location**: Louise Pound Hall 105
- **Graduate TA**:
  - Jacob Williams
    - **Email**: jakew42@gmail.com
    - **Office Hours**: Mondays 2-3PM, Wednesdays 10:30-11:30 AM in Avery 12
- **Undergraduate TAs**:
  - Ayush Manish Agrawal
    - **Email**: ayush.agrawal7661@gmail.com
    - **Office Hours**: TBA
  - Atharva Tendle
    - **Email**: atendle13.3.98@gmail.com
    - **Office Hours**: TBA
- **Instructor**:
  - Dr. Mohammad Rashedul Hasan
    - **Email**: hasan@unl.edu
    - **Office Hours**: TBA

Introduction

Welcome to your first CSCE 478/878 Machine Learning recitation!

Here's how the recitations will work:

- There will be a short introduction or lecture at the beginning of the recitation
- You will then work through a programing task as described in the handout and/or ipython notebook for the session.
  - These will explore the practical application of machine learning, generally through the use of scikit-learn
- You will submit your ipython notebook via handin at the end of class with the naming convention: `<lastname>_<firstname>_<recitation#>.ipynb`
  - E.g. `williams_jake_1.ipynb`
  - For today's recitation, you have until Wednesday January 9 at 11:59PM to submit on handin so we can work out any technical issues
- Recitation, as a whole, is worth 10% of your final grade.
  - Each individual recitation will be worth a little less than 1% of your final grade
- Feel free to discuss with those sitting around you and ask your TAs for assistance!

Today's Task

1) Log into Canvas, navigate to Modules, and select Recitation 1
  a) Open the link to Data Scientist's Handbook 1 and read through the sections up to `Dataset`
  b) If you have not done so already, follow the instructions in the link at the end of 'What You Will Learn in this Notebook Series' to install Python and the required libraries. If you are on Windows,

it may be easiest to use the Anaconda package manager to install Python. This will install all of the necessary packages (and many you don't need…) and give you a preconfigured Anaconda Terminal with all of the paths set. Using that to start your Jupyter notebook will make your life easier. Mac and Linux users should be able to follow the instructions in the link, but are welcome to use Anaconda.

2) Download the Online News Popularity dataset from https://archive.ics.uci.edu/ml/datasets/Online+News+Popularity
3) Start up a fresh Jupyter notebook and recreate the steps performed in the Data Scientist's Handbook 1 for the Online News Popularity dataset. Follow the notes below as you go.
   a) The images in your notebook should be clear, so you will need to change parameters in some places.
   b) You will also need to choose columns that exist in the current dataset for the analyses. Descriptions of the data columns exist on the download page. Choose columns that you find interesting.
   c) For part 7b, printing all columns for this dataset will not provide intelligible outputs. Choose a shortened list of columns (3-5) for which printing counts makes sense, e.g. ' num_videos'. Most columns in this dataset use integer values, so you should remove the if statement that checks the column's datatype.
   d) For part 8b, we can use this same method to restrict our dataset to only data in which we are interested. Use logical indexing to restrict the dataset to data with at least 1 video, i.e. compare the value of ' num_videos' to 0.
   e) For part 17, one-hot encoding is frequently only desirable in columns that only have a small set of possible values. There are several binary columns in our dataset that are perfect for this, but Pandas reads them in as integer data types by default. Moreover, the get_dummies function only works for columns with categorical values. We'll need to fix this before running the code in this section.
      i) First, we'll want to select a few of the binary columns to convert to string-valued. You'll do this with the 'apply' function that you saw earlier when calculating percentiles. Here is an example: df[' is_weekend'] = df[' is_weekend'].apply(lambda x: str(x))
      ii) Find two other binary columns to apply this process to using the histograms you plotted earlier.
      iii) After running `category_df = df.select_dtypes('object')`, we'll run into one more problem: each url is unique, meaning it is a terrible choice for one-hot encoding. We'll need to drop that column from our category_df before proceeding using the code: category_df = category_df.drop('url', axis =1)
   f) For part 18, we'll also want to drop the "url" category before performing the "get_dummies" function for the same reason. If you don't, your computer will try to run ~14,000 choose 2 correlations and you don't want that.
   g) Part 20 will take a minute to run, so don't worry.
4) Once you are finished, hand in your ipython notebook with the naming scheme mentioned above

---

Grading

There are 25 sections to this assignment. Each will be worth 4% of the grade.