Machine Learning with Python for Education and Personnel Economists - Exercises

Michael E. Rose, PhD
Course at Swiss Leading House, March 2020

Create one GitHub repository for yourself. For each question below, create one script that contains both code and answers (as comments at the end of the script). All scripts need to adhere to PEP8 and must be readable to someone that knows Python.

Save the script in the main folder, properly named in correspondence to the name of the exercise. Apart from the script, there should be one folder named "output" to store output such as figures and tables.

1 Exercises for Pandas and Plotting

Sketch for today: French Taunting

1. Coding workflow
   - Read "Code and Data" by M. Gentzkow and J. Shapiro, chapters 2, 3 and 4.
   - What brought Gentzkow and Shapiro to the conclusion, that version control is a necessity?

2. Occupations
   - Import the pipe-separated dataset from `https://raw.githubusercontent.com/justmarkham/DAT8/master/data/u.user` into a DataFrame. The data is on occupations and demographic information.
   - Print the last 10 entries and the first 25 entries.
   - What is the type of each column?
   - How many different occupations are there in the dataset? What is the most frequent occupation?
   - What are the age values with the least occurrence?
   - Create a histogram for occupations.
   - Set "user_id" as index and name the index "User".
   - Create a histogram for occupations again - what is the effect of setting an index first? In the figure, sort the x-values alphabetically.
   - Save the figure as `./output/occupations.pdf`

3. Euro 2012 I
   - Read the data from `./data/Euro_2012.csv` into a DataFrame with column "Teams" as index. The data is on the UEFA Championship 2012 (Euro 2012).
   - How many teams played in the Euro 2012?
   - What are your dtypes? Turn the non-numeric columns into numeric columns before proceeding.
   - Which team has the highest shooting accuracy?
   - Plot shooting accuracy versus passing accuracy.
   - Which team has the second-most shots on target?
   - Eliminate Italy temporarily from the dataset. Which team has the second-most shots on target now? (Hint: Use method-chaining so that Italy is only dropped temporarily!)
   - How many penalty goals did England score?
   - Present only the Shooting Accuracy from England, Italy and Russia.
   - Create a new DataFrame called discipline using the columns "Yellow Cards" and "Red Cards" (and the index).
   - Sort discipline primarily by red cars and secondarily by yellow cards.
   - Output the data as tab-separated textfile `./output/discipline.tsv`.

4. Tips
    - Load seaborn's tips dataset using `seaborn.load_dataset("tips")` into a DataFrame.
    - Convert the short weekday names to their long version (i.e. "Thursday" instead of "Thu") using `.replace()`.
    - Create a lineplot of "tips" on "total_bill" with markers, line styles and color by "day", and facets by "sex".
    - Label the axis so that the unit becomes apparent.
    - Add a title to the legend.
    - Save the figure as `./output/tips.pdf`

5. Alcohol
    - Load the data from `./data/drinks.csv` into a DataFrame. The data is on country's alcoholic consumption.
    - Which continent drinks most beer and wine on average?
    - Create a Boxenplot of "beer_servings" by continent.
    - Reshape the data and create a 3x1 figure for "beer_servings", "wine_servings" and "spirit_servings" by continent (i.e. three boxenplots that share their y-axis and a their color coding).
    - Save the figure as `./output/alcohol.pdf`

6. Iris
    - Read the Iris dataset from https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data directly from the Internet and without header.
    - Name the columns in the following way: "sepal_length (in cm)", "sepal_width (in cm)", "petal_length (in cm)", "petal_width (in cm)" and "class".
    - Set values of the rows 10 to 29 of the column 'petal_length (in cm)' to missing.
    - Replace missing values with 1.0.
    - Save the comma-separated file as `./output/iris.csv` without index.
    - Visualize the distribution of all of the continuous variables by "class" with a catplot of your choice.
    - Save the figure as `./output/iris.pdf`.

7. Memory
    - Load the comma-separated data from https://query.data.world/s/wsjbxdqhw6z6izgdxijv5p2lfqh7gx into a DataFrame.
    - Inspect the DataFrame using `.info()` and with `.info(memory_usage="deep")`. What is the difference between the two calls? How much space does the DataFrame require in memory?
    - Create a copy of the object with only columns of type object by using `.select_dtypes(include=['object'])`.
    - Look at the summary of this object new (using `.describe()`). Which columns have very few unique values compared to the number of observations?
    - Does it make sense to convert a column of type object to type category if more than 50% of the observations contain unique values? Why/Why not?
    - Convert all columns of type object of the original dataset to type category where you deem this appropriate.
    - What is the final size in memory?
    - Could above routine have speeded up somewhere? Hint: Look at the documentation for .read_csv().