BRAC University
Department of Computer Science and Engineering

CSE427 - Machine Learning

Summer, 2018

Problem Set - 01

Playing with Data

In the past, the power of a country measured by its wealth or weapons it had. But today the whole game has changed. Now, it is measured by how much data you have and how many data warriors you have.

As you are doing machine learning, consider yourself as a ”Data Warrior”, maybe an amateur one (but don’t worry you will be promoted to the next rank by the end of this semester).

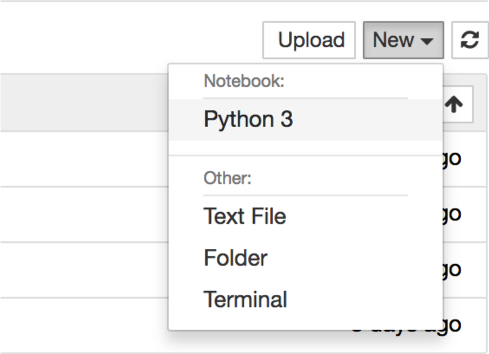
Recently you have been transferred to the data intelligence agency for learning more hands-on data science. One of the field agents of your organization gave his life to steal the census data of the California state, from the US Census Bureau. Which contains the housing price of the whole California state. This very High-Value Data or HVD has been given to your department to deal with. Your department chief M wants you to analyse the data. Now, this is your mission. Should you choose to accept it Or not?

Phase - 01: Setting a secure environment:

First, you need to set up a secure environment to analyse the data. One of the most secure system that your National Security Agency (whatever the acronym looks like) has built for machine learning and data analysis is Anaconda. You can download it from here. It is mandatory for you as a data warrior to have installed the Anaconda in your system. You should download and install the Python 3.6 version or above.

Phase - 02: Testing the environment:

Now that you have installed Anaconda. It’s time for testing the environment. First, open Anaconda. There you will find the code editor named **”jupyter notebook”**. Open it and create a new notebook for this problem set.



Now, write the traditional ”hello world” program.

```
print("Hello World!")
```

If it prints `Hello World!`, that means you are good to go.
Now, try to import some other libraries that are needed for machine learning and data science and let's see if everything is installed as it should be. Use the codes below and match with the outputs.

```
# Check the versions of libraries

# Python version
import sys
print('Python: {}'.format(sys.version))
# scipy
import scipy
print('scipy: {}'.format(scipy.__version__))
# numpy
import numpy
print('numpy: {}'.format(numpy.__version__))
# matplotlib
import matplotlib
print('matplotlib: {}'.format(matplotlib.__version__))
# pandas
import pandas
print('pandas: {}'.format(pandas.__version__))
# scikit-learn
import sklearn
print('sklearn: {}'.format(sklearn.__version__))
```

The output should be something like this.

```
Python: 3.6.3 |Anaconda custom (64-bit)| (default, Oct  6 2017, 12:04:
                                      38)
[GCC 4.2.1 Compatible Clang 4.0.1 (tags/RELEASE_401/final)]
scipy: 0.19.1
numpy: 1.13.3
matplotlib: 2.1.0
pandas: 0.20.3
sklearn: 0.19.1
```

Don't worry if the versions don't match.

Phase - 03: Getting the DATA

Now that your environment is ready, you can download the data from a secure server. Your data is in CSV (comma separated value) format. Which means you don't have to worry about your data formatting or cleaning. Your data is saved in a secure locker here. Download the dataset and keep it as it is.

Phase - 04: Knowing your Data

Now that you have your HVD, its time to know and what's inside the data. First, you need to feed the data to your system so that you can play with your data. Write the given code in your notebook to feed the data to your system (you might want to change the directory to where you have put your data, as I have kept the data file and notebook in the same directory, I didn't need to use different directory path).

```
import os
import pandas as pd

def load_data():
  csv_path = os.path.join("HVD01.csv")
  return pd.read_csv(csv_path)
```

This method will read your data's location and return it to the system.

Now, that we have fed the data to the system, let's look into it. Let's see the first 5 instances of the data. Just write the code below that should show you the first five instances of the data.

```
show_hvd = load_data()
show_hvd.head()
```

If you want you can also see the the overall information of your data with a simple line of code.

```
show_hvd.info()
```

Now that you have seen inside the data, you might have noticed that the *ocean_proximity* attribute is not a *numerical* type, it's *method*. To know more about it, use

```
show_hvd["ocean_proximity"].value_counts()
```

which returns information about that attribute.

The best way to know about your data is through ”*Histogram*”. A histogram will tell you all about your data in ”*bell shaped*” curve. Let's make histograms of the HVD01.

```
import matplotlib.pyplot as plt

show_hvd.hist(bins=50, figsize=(20,15))
plt.show()
```

This will show you the *histograms* of all your data. The figures are very much informative. It can tell you a lot about your data. If you want you can change the size of your histograms as you want, you just need to change the parameters in the code. Make yourself comfortable with the size of figure you want to make.

Questions:

1. Write down the first five instances of the data in a table.
2. How many individual entries are there in the HVD01?
3. By looking at the histograms, do you think any of the data has been capped? If yes, write the names of the attribute or attributes that you think are capped.
4. Are the histograms *tail heavy*? If yes, give your reasonings.
5. If you are asked to index the dataset, which attribute or attributes will you choose as the *unique identifier* for indexing the data? Give your reasonings.