Programming Assignment #1

T81-559: Applications of Deep Neural Networks, Washington University September 25, 2016

Listing 1 shows a sample submission skeleton that you can use as a starting point for this assignment.

Listing 1: Sample Submission Skeleton

```
# Programming Assignment #1,
# Solution by YOUR NAME
# T81-558: Application of Deep Learning
import os
import sklearn
from sklearn.cross_validation import KFold
import pandas as pd
import numpy as np
from scipy.stats import zscore

path = "./data/"
def question1():
    print()
    print("***Question 1***")

def question2():
    print()
    print("***Question 2***")

def question3():
    print()
    print("***Question 3***")

def question4():
    print()
    print("***Question 4***")

def question5():
    print()
    print("***Question 5***")

question1()
question2()
question3()
question4()
question5()
```

Listing 2 shows what the output from this assignment would look like. Your numbers might differ from mine slightly. Questions 2-5 also generate an output CSV file. For your submission please include your Jupyter notebook and any generated CSV files that the questions specified. Name your output CSV files something such as **submit-jheaton-prog1q2.csv**. Submit a ZIP file that contains your Jupyter notebook and 4 CSV files to Blackboard. This will be 5 files total.

Listing 2: Expected Output

```
***Question 1***
Mean weight is: 2970.424623115578

***Question 2***
MPG mean: 7.141133022714575e-17
MPG standard deviation: 1.001258653739211

***Question 3***
Most efficient is: mazda rx-7 gs

***Question 4***
Training DF: 306
Validation DF: 92

***Question 5***
Iteration #1, Training Size: 318, Validation Size: 80
Iteration #2, Training Size: 318, Validation Size: 80
Iteration #3, Training Size: 318, Validation Size: 80
Iteration #4, Training Size: 319, Validation Size: 79
Iteration #5, Training Size: 319, Validation Size: 79
```

Question 1

Using Pandas, load the **auto-mpg.csv** dataset and calculate the mean (average) value for the weight. Display this value similarly to Listing 2. No data files need to be submitted for this part.

Question 2

There are missing values in the horsepower column of the **auto-mpg.csv** dataset. Using Pandas, replace these with the median value for horsepower. Next, transform the mpg, displacement, horsepower, weight, & acceleration columns to Z-Scores. It is easy to check if a column is in Z-Score form. After you perform the Z-Score transformation, the mean of the MPG column should be near zero and the standard deviation should be near one. Display these values as shown by Listing 2. Capture the dataset, with MPG transformed to a CSV file that should look similar to Listing 3:

Listing 3: Question 2 Output Sample

```
mpg,cylinders,displacement,horsepower,weight,acceleration,year,origin,name
-0.706438700650983,8,1.090603697954175,0.6731176189353574,0.6308698741675456,-1.29549834
    chevrolet chevelle malibu
-1.0907506236404463,8,1.5035143043761154,1.5899581813455976,0.8543329711977774,-1.477037
    buick skylark 320
-0.706438700650983,8,1.196231992620253,1.1970265117412089,0.5504704530138114,-1.65857724
    plymouth satellite
...
```

Notice that only the requested columns were transformed!

Question 3

We can create a calculated field that tells us the efficiency of a car in the **auto-mpg.csv** dataset. This value will be calculated as horsepower divided by displacement. This is the amount of power a car can deliver per cubic inch of displacement. Create a new dataset that has only the car name and efficiency. Sort this dataset by efficency, with the most efficient cars first. Create a CSV that looks similar to Listing 4. Also write out the name of the most efficient car, as seen in Listing 2.

Listing 4: Question 3 Output Sample

```
name,efficiency
mazda rx-7 gs,1.4285714285714286
mazda rx2 coupe,1.3857142857142857
mazda rx-4,1.375
maxda rx3,1.2857142857142858
datsun 1200,0.9583333333333334
...
```

Question 4

When working in a team of data scientists it is very important that the team members agree on the training and validation sets. To accomplish this master files are typically generated at the beginning of a project that specify exactly which data will be in the training and validation sets. Using Pandas and the **auto-mpg.csv** dataset, generate a file that contains both the testing and validation sets. Include a column to indicate if the row is part of the training (T) or validation (V) set. Add this new column, named 'set' to the dataset just after the mpg field. Create an output CSV similar to Listing 5. Report the size of each set, as shown in Listing 2.

Listing 5: Question 4 Output Sample

```
mpg,set,cylinders,displacement,horsepower,weight,acceleration,year,origin,name
28.0,T,4,120.0,79.0,2625,18.6,82,1,ford ranger
13.0,T,8,318.0,150.0,3940,13.2,76,1,plymouth volare premier v8
...
37.7,T,4,89.0,62.0,2050,17.3,81,3,toyota tercel
26.0,T,4,97.0,46.0,1950,21.0,73,2,volkswagen super beetle
33.0,V,4,91.0,53.0,1795,17.4,76,3,honda civic
19.0,V,6,232.0,100.0,2634,13.0,71,1,amc gremlin
...
16.0,V,8,318.0,150.0,4190,13.0,76,1,dodge coronet brougham
12.0,V,8,350.0,180.0,4499,12.5,73,1,oldsmobile vista cruiser
```

Question 5

If the team of data scientists are using k-fold cross validation, it is very important that they use the same folds. To accomplish this, a master file is created that has two columns that specify what set and iteration each row belongs to. Using Pandas and the **auto-mpg.csv** dataset, create an output data file that contains all 5 folding iterations (the file will be five times the length of the original), with a column that specifies if the row belongs to the training (T) or validation (V) fold. Listing 6 shows how this file would appear:

Listing 6: Question 5 Output Sample

```
set,iteration,mpg,cylinders,displacement,horsepower,weight,acceleration,year,origin,name
T,1,22.0,4,122.0,86.0,2395,16.0,72,1,ford pinto (sw)
T,1,28.0,4,97.0,92.0,2288,17.0,72,3,datsun 510 (sw)
...
T,1,28.0,4,120.0,79.0,2625,18.6,82,1,ford ranger
T,1,31.0,4,119.0,82.0,2720,19.4,82,1,chevy s-10
V,1,18.0,8,307.0,130.0,3504,12.0,70,1,chevrolet chevelle malibu
V,1,15.0,8,350.0,165.0,3693,11.5,70,1,buick skylark 320
...
V,1,21.0,4,120.0,87.0,2979,19.5,72,2,peugeot 504 (sw)
V,1,26.0,4,96.0,69.0,2189,18.0,72,2,renault 12 (sw)
T,2,18.0,8,307.0,130.0,3504,12.0,70,1,chevrolet chevelle malibu
T,2,15.0,8,350.0,165.0,3693,11.5,70,1,buick skylark 320
...
T,2,28.0,4,120.0,79.0,2625,18.6,82,1,ford ranger
T,2,31.0,4,119.0,82.0,2720,19.4,82,1,chevy s-10
...
V,2,16.0,8,318.0,150.0,4498,14.5,75,1,plymouth grand fury
V,2,14.0,8,351.0,148.0,4657,13.5,75,1,ford ltd
T,3,18.0,8,307.0,130.0,3504,12.0,70,1,chevrolet chevelle malibu
```

```
T,3,15.0,8,350.0,165.0,3693,11.5,70,1,buick skylark 320
...
...
T,3,28.0,4,120.0,79.0,2625,18.6,82,1,ford ranger
T,3,31.0,4,119.0,82.0,2720,19.4,82,1,chevy s-10
V,3,17.0,6,231.0,110.0,3907,21.0,75,1,buick century
V,3,16.0,6,250.0,105.0,3897,18.5,75,1,chevroelt chevelle malibu
...
V,3,33.5,4,98.0,83.0,2075,15.9,77,1,dodge colt m/m
V,3,30.0,4,97.0,67.0,1985,16.4,77,3,subaru dl
T,4,18.0,8,307.0,130.0,3504,12.0,70,1,chevrolet chevelle malibu
T,4,15.0,8,350.0,165.0,3693,11.5,70,1,buick skylark 320
...
T,4,28.0,4,120.0,79.0,2625,18.6,82,1,ford ranger
T,4,31.0,4,119.0,82.0,2720,19.4,82,1,chevy s-10
V,4,30.5,4,97.0,78.0,2190,14.1,77,2,volkswagen dasher
V,4,22.0,6,146.0,97.0,2815,14.5,77,3,datsun 810
...
V,4,34.3,4,97.0,78.0,2188,15.8,80,2,audi 4000
V,4,29.8,4,134.0,90.0,2711,15.5,80,3,toyota corona liftback
T,5,18.0,8,307.0,130.0,3504,12.0,70,1,chevrolet chevelle malibu
T,5,15.0,8,350.0,165.0,3693,11.5,70,1,buick skylark 320
...
T,5,34.3,4,97.0,78.0,2188,15.8,80,2,audi 4000
T,5,29.8,4,134.0,90.0,2711,15.5,80,3,toyota corona liftback
V,5,31.3,4,120.0,75.0,2542,17.5,80,3,mazda 626
V,5,37.0,4,119.0,92.0,2434,15.0,80,3,datsun 510 hatchback
...
V,5,28.0,4,120.0,79.0,2625,18.6,82,1,ford ranger
V,5,31.0,4,119.0,82.0,2720,19.4,82,1,chevy s-10
```

Essentially, the dataset is duplicated 5 times in the output file. Each of these repetitions is specified by the ’iteration’ column. Notice the values of 1,2,3,4, and then 5 for this column. Within each iteration, some of the rows are training (T), and others are validation (V). Also report the size of the training/validation sets for each iteration, as demonstrated by Listing 2.