**LAB 1 of DSA230/CIS492 - Cleveland State University**

**2:00pm-3:15pm of 10/13/2022**

**Due time: 11:30 pm of 10/14/2022**

*[100 points in total] Please finish this lab assignment independently. No discussion with classmates. You may use notes, slides or internets for your references. If you have any questions, you can ask the TA for help. If you cannot finish all the tasks, do not worry and just submit whatever you have already written for partial credits. Lab submission late within 24 hours will receive 20% deduction for late submission. Lab submission late more than 24 hours but within 48 hours will receive 40% deduction for late submission. Lab submission late more than 48 hours is not accepted except a significant emergency. Code must be written in Python3 and submitted via Blackboard. Sorry about no email attachment for submission.*

**Objective**

Create a python3 program named ‘*gaussian.py’* to learn Gaussian distributions in data science.

**Specification:**

In data science, usually we need to find the hidden distributions of the data. A frequently-used way is to learn a *Gaussian distribution* (also called Normal distribution) from training data (https://en.wikipedia.org/wiki/Normal\_distribution). Based on the learned Gaussian distribution from training data, given a new testing data, we can evaluate the probability of the new testing data fitting with the training data. In Gaussian distribution, there are only two important parameters: *mean* and *standard deviation*. In this lab, we will read three text files and learn *mean* and *standard deviation* for each of them.

**Files**: Atlantic\_salmon.txt, Largemouth\_bass.txt, Rainbow\_trout.txt

One text file contains some numbers to represent the length of each fish instance (in inches) of that class in the pool. Please **read** the numbers, **ignore the noisy data**, and **calculate** the corresponding *mean* and *standard deviation* for different fish classes. Then, **generate/write** a new file called ‘*result.txt’* in the same path of ‘*gaussian.py’* to summarize the learned Gaussian distributions. **The noisy data to be ignored are String, Negative Numbers, Numbers Greater than 100.**

**Example:** The content of the generated file ‘*result.txt*’ should be in the following format:

Mean and standard deviation of Gaussian of Atlantic salmon are xxx and xxx.

Mean and standard deviation of Gaussian of Largemouth bass are xxx and xxx.

Mean and standard deviation of Gaussian of Rainbow trout are xxx and xxx.

**Grading:**

You only need to submit the *gaussian.py* file on Blackboard. TA will run your program and check whether your code could generate the new and correct *result.txt* file or not. You can use NumPy’s *.mean()* and *.std()* functions to calculate *mean* and *standard deviation*. Manual computation will receive 0 points.

**Your code should be in the following function-based style:**

import numpy as np

def extract\_data\_salmon(filename):

# write your code to return mean and std of salmon

def extract\_data\_bass(filename):

# write your code to return mean and std of bass

def extract\_data\_trout(filename):

# write your code to return mean and std of trout

mean1, std1 = extract\_data\_salmon('Atlantic\_salmon.txt')

mean2, std2 = extract\_data\_bass('Largemouth\_bass.txt')

mean3, std3 = extract\_data\_trout('Rainbow\_trout.txt')

print(mean1, std1)

print(mean2, std2)

print(mean3, std3)

# write your code here to generate/write a new file named 'result.txt'