Homework 1

Data Science
January 23, 2020

- Collaboration: Homeworks are individual work. Visit here for more info on academic integrity.
- Formatting: You are supposed to upload two files (`hw1_writeup.pdf` and `hw1_code.py`). For the pdf file, You can produce the file however you like (e.g. LaTeX, Microsoft Word, scanner), as long as it is readable.

Problem 1 (Probability Distribution)

Answer all parts of the questions below in your `hw1_writeup.pdf` file.

(a)[2pt] Choose your favorite (univariate) distribution and write down the probability density function. Clearly indicate what are the parameters of the distribution.

(b)[2pt] Use a Python plot library (e.g. matplotlib) to make a plot of the probability density function above. Include the plot in your `hw1_writeup.pdf` file. (you don't have to include any code in your answer for this part)

(c)[2pt] Specify one or more applications of your chosen data distribution. Briefly explain what properties/characteristics of this distribution make it a suitable distribution for this application.

Problem 2 (Data Estimators)

Create a python file `hw1_code.py` and do the following (make sure you import necessary packages like numpy etc.)

(a)[3pt] Choose a univariate distribution and sample 1000 samples in python. (You might be interested in this link for numpy random sampling.) Store the samples as a numpy array called `sample_data`. Plot a histogram of `sample_data` and include the histogram in `hw1_writeup.pdf`

(b)[6pt] Define a function `data_estimators` that inputs a numpy array and returns a python list containing the sample mean, standard deviation, skewness, kurtosis. You are free to use any python package (e.g. SciPy) that comes with the needed statistical functions. Use `data_estimators` to compute the sample mean, standard deviation, skewness, kurtosis of `sample_data`

(c)[1pt] (Answer this part of the question in `hw1_writup.pdf` only. No extra code is required to appear in `hw1_code.py`). Generate another 1000 samples from the same distribution as in (a) and compute the sample mean of the new samples. Is the new sample mean the same as the old sample mean? Why?