CARNEGIE MELLON UNIVERSITY 10-606

HOMEWORK 3

DUE: Saturday, Oct. 9, 2021

https://www.cs.cmu.edu/~10606

INSTRUCTIONS

- **Format:** Use the provided LaTeX template to write your answers in the appropriate locations within the *.tex files and then compile a pdf for submission. We try to mark these areas with STUDENT SOLUTION HERE comments. Make sure that you don't change the size or location of any of the answer boxes and that your answers are within the dedicated regions for each question/part. If you do not follow this format, we may deduct points.

  You may also type you answer or write by hand on the digital or printed pdf. Illegible handwriting will lead to lost points. However, we suggest that try to do at least some of your work directly in LaTeX.

- **How to submit:** Submit to Gradescope a pdf with your answers. Again, make sure your answer boxes are aligned with the original pdf template.

- **Policy:** See the course website for homework policies, including late policy, and academic integrity policies.

Name	
Andrew ID	
Hours to complete (nearest hour)	

1 MLE [15 pts]

Assume we have data $\mathcal{D} = \{x^{(i)}\}_{i=1}^{N}$ and we assume our $x^{(i)} \in \mathbb{R}$ are i.i.d from the a distribution with the following density function:

$$f(x) = \frac{\lambda^x}{x!} e^{-\lambda}$$

1. **[5 pts]** Write the expression for the likelihood of the data for this distribution with parameter λ.

Work

2. **[5 pts]** Write the expression for the log-likelihood of the data for this distribution with parameter λ.

Work

3. **[5 pts]** Give the equation for MLE of λ in terms of the the data.

Work

2 Data standardization [8 pts]

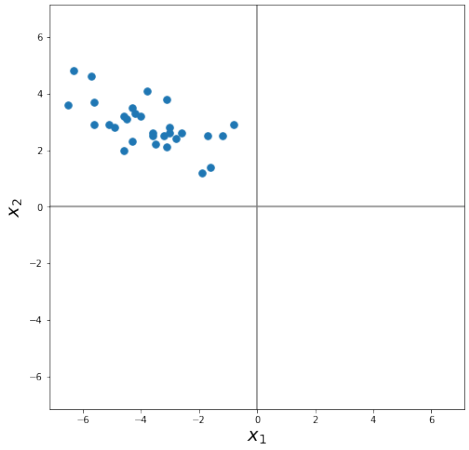
Data standardization is a common data preprocessing step that can be helpful for a number of machine techniques.

Data standardization makes the data such that the measurement for each input feature X_j has zero mean and variance equal to one. For every input feature in your data, X_j, you must subtract the mean of those feature values across all N data points and divide it by the associated standard deviation:

$$\hat{x}_j^{(i)} = \frac{x_j^{(i)} - \mu_j}{\sigma_j}$$

where $\mu_j = \frac{1}{N}\sum_{i=1}^{N} x_j^{(i)}$ and $\sigma_j = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(x_j^{(i)} - \mu_j\right)^2}$.

Consider our dataset with two input feature measurements, X_1 and X_2, and $N = 30$ data points, $\mathcal{D} = \{x_1^{(i)}, x_2^{(i)}\}_{i=1}^{N}$. The values for are data are stored in the following Jupyter notebook along with code to help plot data: https://drive.google.com/file/d/1AFEwgJk0GgsPw72SbToxPsWaaw9SyYB5



Standardize this dataset and plot the resulting data points below. You may use NumPy and matplotlib, but you may not use other Python libraries, such as scikit-learn.

Plot

3 Collaboration Policy

After you have completed all other components of this assignment, report your answers to the following collaboration questions.

1. Did you receive any help whatsoever from anyone in solving this assignment? If so, include full details including names of people who helped you and the exact nature of help you received.

2. Did you give any help whatsoever to anyone in solving this assignment? If so, include full details including names of people you helped and the exact nature of help you offered.

3. Did you find or come across code that implements any part of this assignment? If so, include full details including the source of the code and how you used it in the assignment.