DS 4420: Machine Learning and Data Mining 2 **(Due: Feb. 28 2025)**

Homework Assignment #4

Professor: Eric Gerber *Name:* ____________________

Instructions: Please include the following information on the first page of your completed homework write-up:

1. Your name
2. DS 4420
3. Homework #4

You will submit **up to three files** to Gradescope for this homework:

- A .pdf file with handwritten/latex typeset answers to the math problems
- A .ipynb file with all python code written
- A .r/.rmd file with all R code written (or, a .html/.pdf file with the knitted .rmd code/output)

Answers that are not supported by reasoning/work will not receive full credit. **Homework is due by 11:59 pm, via Gradescope, on the date above.** Late submissions will **not** be accepted, but you may receive extra credit for early submission (see syllabus for details).

You will also be graded on organization/neatness of the submitted files.

Discrete Distribution (40 points)

You will need two data files for the below problems: the **ds4420_spotify.csv** file and the **drg_points.csv** file. The former contains features grabbed using Spotify's API on the songs you listed as your favorite on the getting to know you form at the beginning of the semester, while the latter contains Dr. Gerber's performance on the varsity basketball team during his senior year of high school.

(1: 10 points) Assume the Spotify data set is representative of the type of playlist all DS 4420 students would create.

(a) Call two random variables X (whether a song is explicit or not) and Y (how many songs on a playlist of size n are explicit). Identify the reasonable distributions for X and Y with a short justification for why those probability distributions seem appropriate for these variables.

(b) In **both Python and R** compute the Maximum Likelihood Estimate MLE(s) of the parameter(s) under the distributions from (a) and then estimate the probability that on a playlist of $n = 50$ songs from DS 4420 students, at least 10 of those songs would be explicit.

THE HOMEWORK CONTINUES ON THE NEXT PAGE

(2: 15 points) Assume the Spotify data set is representative of the type of playlist all DS 4420 students would create.

(a) Call two random variables X (the month that a song in the playlist was released) and Y (how many songs on a playlist of size n were released in each month). Identify the reasonable distributions for X and Y with a short justification for why those probability distributions seem appropriate for these variables.

(b) In **both Python and R** compute the MLE(s) of the parameter(s) under the distributions from (a) and then estimate the probability that on a playlist of $n = 12$ songs from DS 4420 students, each month is represented once.

(c) Discuss **in at least two full sentences** what practical benefit(s) there might be to analyzing the release month of songs.

(3: 15 points) During his senior year of high school, Dr. Gerber's varsity basketball team went 3-18. Perhaps unsurprisingly (given the record), Dr. Gerber was a key "contributor" for the team. He kept track of his points in each game and stored them in the **drg_points.csv** file for later user.

(a) Call the random variable X: the number of points Dr. Gerber scored in a single game. Identify a reasonable distribution for X with a short justification for why the probability distribution seems appropriate for the variable.

(b) In **both Python and R** compute the MLE(s) of the parameter(s) under the distribution from (a) and then estimate:

- The probability Dr. Gerber scores more than 2 points in a game
- The probability Dr. Gerber scores less than 2 points in a game

(c) In **both Python and R** create a histogram of the observed values of X and then overlay a density plot of the distribution you chose in (a). Consider also if the MLE(s) calculated in (b) match the relationship of the parameters of the distribution in theory. Make a determination on if you believe the distribution you chose is a good fit for these data.

Continuous Distributions (60 points)

(4) Assume the Spotify data set is representative of the type of playlist all DS 4420 students would create.

(a) Call the random variable X: the duration of a song (in seconds). Discuss **in at least two full sentences** whether you believe an exponential distribution makes sense in modeling this feature, and why/why not.

(b) In **both Python and R**, calculate the MLE for θ (the mean) of an exponential distribution fit to X and use it to estimate the following probabilities:

- The probability a random favorite song of a DS 4420 student lasts longer than 4 minutes
- The probability a random favorite song of a DS 4420 student lasts between 2 and 5 minutes
- The probability a random favorite song of a DS 4420 student lasts less than 1 minute

Note: the **pexp()** function in R and the **expon.cdf()** function in python are equivalent, but are parameterized differently: **pexp()** takes the *rate* $(1/\theta)$ as an argument, while **expon.cdf()** takes the *scale* (θ).

THE HOMEWORK CONTINUES ON THE NEXT PAGE

(5) Assume the Spotify data set is representative of the type of playlist all DS 4420 students would create.

(a) Call the random variable X: the duration of a song (in seconds). Discuss **in at least two full sentences** whether you believe a normal distribution makes sense in modeling this feature, and why/why not.

(b) In **both Python and R**, calculate the MLEs for μ (the mean) and σ (the standard deviation) of a normal distribution fit to X and use them to estimate the following probabilities:

- The probability a random favorite song of a DS 4420 student lasts longer than 4 minutes
- The probability a random favorite song of a DS 4420 student lasts between 2 and 5 minutes
- The probability a random favorite song of a DS 4420 student lasts less than 1 minute

(6) In **both Python and R** create a histogram of the observed values of X and then overlay density plots for **both** the exponential and normal distributions. Discuss:

- Which distribution seems to fit best based on the plot?
- Does the distribution that seems to best fit the data match your intuition from parts 4 (a) and 5 (a)?
- Compare the probabilities you calculated in parts 4 (b) and 5 (b). Which are different/similar between the two distribution fits?
- Do you believe either distribution can be used as an effective model for the duration of songs?

(7) Assume that X_1: artist popularity and X_2: track popularity follow a multivariate gaussian distribution with mean vector μ and covariance matrix Σ.

(a) In **both Python and R** estimate the MLEs for μ and Σ from the data. What does the structure of $\hat{\Sigma}$ suggest about the relationship between artist and track popularity? What does this mean for the **type** of multivariate gaussian we are assuming the data follow?

(b) Using the MLEs of the distribution, in **both Python and R** create a 3D density plot of the theoretical multivariate gaussian distribution. Discuss where the peak of the distribution is and what sort of values appear to be very unlikely under this distribution.

(c) In **both Python and R**, create 2D contour/heatmap density plots to compare the observed data distribution with the theoretical distribution under the MLEs. Do you believe that a multivariate gaussian is a reasonable distribution for modeling artist and track popularity jointly? Explain your reasoning.