COM 307 Machine Learning/Data Mining Project 1: Simulated data (due by September 28 before class start)

In this first project you will be using simulated data to see how different assumptions and approaches can give varying results with the same known underlying distribution (you can conclusively *know* the underlying distribution only when using simulated data). I would like you to randomly generate the following data sets (each student will have slightly different results due to this randomness):

Imagine a "loaded die" that has a 10% chance of showing 1, 10% chance of showing 2, 10% chance of showing 3, 10% chance of showing 4, 10% chance of showing 5, and 50% chance of showing 6. Generate:

- A dataset ("dataset A") generated from rolling this die 20 times
- Another dataset ("dataset B") generated from rolling this die 200 times
- A final dataset ("dataset C") generated from rolling this die 2000 times

Each dataset should be independent (that is, don't use the 20 die rolls from dataset A as a fraction of datasets B or C). You need to only generate each dataset once (even though each dataset will be used 4 times below).

For each of these datasets, you will be trying to learn the underlying distribution (which you know because it is simulated data, but you can pretend that you do not) under 4 different scenarios:

- Scenario 1: You have no idea if this die is loaded or not, so you will not be using a prior probability.
- Scenario 2: You are completely unsure if your die is fair or not, so your prior probability is uniform (that is an equal probability of any theta).
- Scenario 3: You think that your die favors 6 with the following distribution: p(theta=0.5) = 45%, p(theta=0.6) = 35%, p(theta=1/6) = 20%.
- Scenario 4: You think that your die favors 6 with the following distribution: p(theta=0.5) = 52%, p(theta=0.6) = 28%, p(theta=1/6) = 20%.

Therefore, you will be doing a total of 12 machine learning "runs" (4 scenarios each with 3 datasets).

You are free to use any programming or scripting language of your choice. If your language doesn't support random numbers (or, more likely, you don't know/don't want to figure out how to generate random numbers in your language), feel free to use https://www.random.org/ to generate random values and import them as text for further analysis.

Now, for the actual machine learning. You will use Maximum Likelihood Estimation (MLE) to compute the probability of each state (the numbers one through six) for scenario 1 for each dataset. For scenarios 2 through 4, you will use Bayesian inference with the prior probabilities specific for each scenario ($p(\theta=0.5)=0.5$ for scenario 4's calculation for rolling a "6," for example).

Important note: For scenarios 2, 3, and 4, you will be computing the relative likelihood of the different thetas for "6" only (not for values 1 through 5). This makes it analogous to the coin flip example for class in that there are only two outcomes – "6" or "not 6." That is, you don't care if the die roll is a 1, a 2, a 3, a 4, or a 5, you just care that it isn't 6.

Grading will be based on:

1) Proper implementation of the machine learning algorithms (MLE & Bayesian inference). [55 points]
2) Providing your datasets to allow for easier grading/debugging on my end. [10 points, but it can influence other aspects of your grade if missing]
3) Your explanation of your findings. [35 points]

[Submit]

Submit as a zipped file the following files:

1) Your code (in whatever language you used)
2) Your datasets in text for (can be as a single file or split)
3) A brief write-up of what you found during this exercise. What methods/scenarios did well? What methods/scenarios did not? Did the methods/scenarios all perform well for some datasets? Were there other datasets where it made a bigger difference what method/scenario was used? Please provide enough detail to allow me to evaluate whether or not you understand how the datasets and scenarios influenced one another. You do not need to worry about formatting, punctuation, or the like. As long as I can understand what you're trying to say, that's sufficient.