COGS9: Introduction to Data Science
Assignment #4: Machine Learning
Due date: Friday 2024 November 29 23:59:59
Grading: 10% of overall course grade; 40 points total
Download the editable version of this document and add your responses in the locations indicated. Please respond using the blue font color used in the response text, as it makes the assignments easier to grade. This is assignment is to be completed *individually*. Once completed, save the document as a PDF and submit on Gradescope. Be sure to assign pages to each answer when you submit (see the Gradescope instructions here.)
Instructions: For this assignment, you'll be using interactive widgets throughout this piece from the Parametric Press, titled "The Myth of the Impartial Machine". Navigate to the following URL to get started:
https://parametric.press/issue-01/the-myth-of-the-impartial-machine/
Read and understand the whole piece first and then go through to answer the questions throughout. No answer should be more than 1-2 paragraphs.
Question 1. (4 pts) What is the relationship between machine learning and artificial intelligence?
AI refers to the broader idea of making computers act intelligently without being explicitly programmed and Machine learning refers to the use of statistical methods to identify patterns in data. Machine Learning is a subset of AI.
Question 2. **(10 pts)** Sampling Bias: We often aren't able to collect information about entire populations and are forced to take a sample from the population. In the "Sampling Errors Can Lead to Biased Models" Interactive widget, collect 10 samples, each of sample size 10. That is, set the "Sample size" slider to equal 10, generate a sample, record the sample mean, and do that 10 times. Record each sample mean in the table:

Sample	1	2	3	4	5	6	7	8	9	10
Sample mean	13500	9400	7900	9500	8100	9300	8600	8500	12100	10100

Interpret the results of your 10 samples, commenting on the effects of sampling bias and how it relates to the true population mean. This is an *interpretation*, which is somewhat open-ended. Your job here is to talk about how the samples you drew from that population compare to the true mean. How often do the samples diverge from it, how far, etc.
The choice sampled can easily be an overestimate or underestimate of the population mean depending on if a higher proportion of higher earners or lower earners are drawn in the sample. This leads to the sampling means to reflect very different sample means compared to the population mean.

Question 3. (4 pts) Briefly explain the sampling bias issue underlying the ImageNet data discussed in the article. If you were in charge, brainstorm at least one way you would go about addressing this bias.

The problem with the ImageNet data was that I sampled 45% of the data from the population from the United States when only 4% of the world's population is from the United states. Additionally the data from China and India only represented 3% of the data when it accounts for about 36% of the world's population. If I was in charge of this project I would subdivide the different countries into its own identification process or I would make sure to get a better representative sample of the world's population.

Question 4 (10 pts). In the "Algorithms can amplify bias found in data" section, adjust the bias and accuracy sliders for five different combinations. Record each combination you try in the table below.

	Bias	Model Accuracy	Result
ex	70/30	70%	Not incentivized to amplify bias
1	50/50	60%	Not incentivized to amplify bias
2	80/20	80%	Not incentivized to amplify bias
3	100/0	80%	incentivized to amplify bias
4	100/0	50%	Incentivized to amplify bias
5	60/40	90%	Not incentivized to amplify bias

Using the data you collected, explain the relationship between model bias and accuracy to determine when models are incentivized to amplify bias. Hint: in the paragraph preceding the interactive used for this question, it notes that when the training data are biased toward women, if the model *always selects women* it performs better.

The less biased a model is the better but datasets tend to follow realism (More women cook then men). As the bias is increased the models are incentivized to amplify bias and as the model accuracy increases the models are not incentivized to amplify bias. You want models that minimize bias because the bias can make the models more inaccurate then they normally are. As there is more bias you need higher levels of accuracy for the model to not be incentivized to amplify bia

Question 5. (6 pts) In the Runaway Feedback Loops simulation. Set the simulator to have the same number of crimes occur in precinct A and precinct B to be “4”. Run the simulation, taking the time to understand the results as the simulation runs. Explain the results of the simulation, using the outcome of the simulation to explain what has happened with regards to crime and policing.

In A and B there is the same amount of crimes that occur in both locations. Since the model begins with a slight bias towards crime in location A the police officers are first sent to location A. As the crimes are observed in location A and not location B the model will then continue to send the police officers to location as It is reinforced that crimes are committed there. This cycle repeats furthering the models predictions that crimes are being committed in Location A to the point where the model will send all officers to Location A even though there is an equal amount of crimes being committed in Location A and B. This feedback loop makes the data and results extremely lopsided.

Question 6. (6 pts) Choose one of the methods explained in the article for how people are tackling bias. Explain briefly what the approach is. Then, think critically about the pros and cons of this approach.

The De-biasing approach is making sure that the samples are representative samples. This leads to datasets that have less bias. The pros to this approach is that the data collected is more ideal and greatly helps reduce bias in modeling and predictions. A con of this approach is the cost and inconvenience of getting a completely representative bias.