Introduction to Data Science Homework Set #1

1. Will the cumulative number of people who have gone to college converge or diverge? Explain

The cumulative number of people who have gone will diverge. With each passing year, the cumulative number will increase, since >=0 new people will attend each year. In theory the limit is infinite, assuming there will always be colleges and a portion of the population attending college over time.

1. The release of chlorofluorocarbons (CFCs) was banned by treaty and CFCs decay in the atmosphere over time. Would we expect the quantity of CFCs to converge after an arbitrarily long period of time without new emissions? Explain

The quantity of CFCs will converge to 0, a finite limit, over a very long period of time. Assuming that people and corporations comply with the ban, there will be no new CFCs, and all existing CFCs will slowly decay. With enough time there will be no more CFCs in the atmosphere.

1. Which series converges the fastest: , , or ? Explain

The 3rd series converges fastest, and converges to 0.

Series 1 simplifies to . The first term dominates because it’s slower (has p=1, which is lower than p=3 in the second term). Therefore, series 1 has a linear convergence rate, p=1. It diverges.

Series 2 simplifies to . The first term dominates because it’s slower (has p=2, while the second term has p=5. Therefore, series 2 has a superlinear convergence rate, p=2. It converges to 0.

Series 3 has a superlinear convergence rate, p=3. Its p is the highest, meaning it converges faster. Series 3 converges to 0, but faster than series 2.

1. Consider two algorithms: (A) has quadratic convergence and requires operations and (B) has linear convergence and requires operations. What would be the characteristics of a data science project when algorithm (B) would be preferable?

Algorithm B requires operations to execute, making it less complex, though it converges at a slower rate. A project where B is preferrable is one where the dataset has fewer dimensions, has fewer function calls, and requires fewer “reads” and “writes” to memory.

1. Describe whether and how the probability of rain on a given day could be described as a series with the Markov property.

A series has the Markov property if its future state is independent of its past state, and only depends on its current state. The probability of rain on any given day can described as a series with the Markov property in the short run, in “typical” climate conditions. Absent extreme weather events recently, the chance of rain today is dependent on today’s weather conditions, rather than the rain from last week.

But the chance of rain today might not have the Markov property in the long-run or in certain contexts. In reality, weather conditions are complex and influenced by many factors, including prior weather. For example, a hurricane last week might increase the probability of rain today, and a long drought with no rain for six months might decrease the probability of rain today.

1. Describe how representation bias could affect a facial recognition algorithm?

Representation bias happens when the training data isn’t representative of the population the algorithm will be applied to. This can occur if we have selection bias during the data collection process, the population changes, or we apply an algorithm to a population with different types of faces. For example, if a facial recognition algorithm was developed in a region with mostly young White people and trained on faces from its own population, it can have high accuracy when recognizing White people. However, it will be less accurate for people of other ages and races/ethnicities, because it had fewer data points from those other types of faces to learn from. This algorithm wouldn’t perform well in other regions where there are more older and non-White people.

1. Describe how algorithmic biases (as described in the reading: <https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/>) can lead to poor outcomes and one possible approach to mitigating algorithmic bias.

Algorithms can be biased if they are trained on data that is incomplete and unrepresentative, even if programmers don’t intend to discriminate. The data that algorithms are trained on reflect historical human biases in our society such as racism, sexism, and lack of representation of minorities in certain social contexts. As a result, algorithms can perpetuate these biases and have discriminatory effects. For example, a facial recognition algorithm used by law enforcement can perpetuate discrimination against Black males if it’s trained on mugshots, since Black males are overrepresented in the criminal justice system due to systemic racism. With this bias, the algorithm might wrongfully incriminate innocent Black males and be less effective in detecting crime from other groups.

One possible approach to mitigate algorithmic bias is to require a bias impact statement for each algorithm. This is a form of self-regulation where programmers acknowledge and disclose their algorithm’s assumptions and potential biases. While this won’t prevent bias and discrimination, it will increase transparency and guide programmers address potential biases throughout the development process.