Charlie Gorman

CSCE580 Artificial Intelligence

Quiz 1 Responses

Q1.

a. Open data is information that is openly available to anybody. Furthermore, data that anyone can access, modify, share, and use for analytics, models, or statistics. An example of open data that I produce and that others can use is about my own health data. I could track and plot daily calorie intake, hours slept, and data from a fitness watch (Apple Watch, Garmin, Whoop, etc.) and publish the anonymized results, possibly including height, weight, etc. If it's shared in a clear manner, others could use it for health analytics or statistics, as well as training a model for whatever their goal is.
b. – "Missing at random". Data could be missing at random because a subset of the data or participants (i.e. from a survey) do not reply to a question possibly due to privacy concerns (address, income, phone number, etc.). "Missing not at random". Data could be missing not at random because of a group's inability to participate (i.e. inability to collect the information that is being asked of them, broken tool / sensors).

   –By "omission", you can effectively ignore missing values and remove them. The assumption here is that the missing values are random and will not mess up the dataset. The issue with this assumption is if the missing values are not random, this can introduce bias or skewing to the data set that isn't accurate.

   –By "imputation", you can replace, or impute, the missing values with known values, such as the mean, median, or a prediction based on the dataset and model being used. The assumption here is that the estimations or imputations are sound approximations of the missing data. The risk with this assumption is if the imputations are not good approximations of the data. This can introduce disruptions in the natural patterns of the data, possibly creating correlations that were not there to begin.