Problem Set 1

Data Science 602: Data Analysis and Machine Learning

Spring 2022

Please submit your code as a Jupyter notebook. Problems 3 – 5 do not require code and can be submitted either as a Word document or a Jupyter notebook. A template is available on the Google shared drive.

1. **Basic data analysis** Using the `weather` dataset, provide the following summary statistics:
   (a) Over the entire datasets, which observations had the hottest and coldest temperatures? What were those temperatures?
   (b) In 2020, what month had the hottest average temperature? The coldest?
   (c) In 2020, how many days were associated with rain (i.e., liquid precipitation exceeding zero for at least one observation)?
2. **Summary data analysis** Open and clean the `citations` dataset using the script provided. From the clean dataset, provide the following summary statistics:
   (a) Which violation types, on average, incur the highest fines?
   (b) In 2020, which violation type was the most common? What percentage of total citations issued that year did it comprise?
   (c) In 2020, which violation type was responsible for the highest fine revenue (i.e., highest sum of fine amounts)? What percentage of total fine revenue did it comprise for the year?
   (d) Show the fines charged against the 10 tags (license plates) with the highest total fines assessed in 2020. For each, show the total number of citations, and mean/standard deviation of the fine amount.
   (e) Due to accuracy concerns, Baltimore City suspended fixed speed camera enforcement for a multi-year period that is covered by the dataset. Using the data, show when the hiatus began and when citations resumed.
3. **Identifying 'interesting' patterns** Frawley, et al., [1] define an 'interesting' pattern as interesting if it is "novel, useful, and non-trivial to compute." Explore either of the above datasets to identify an interesting pattern. Describe the pattern and why it is novel, useful, and non-trivial.
4. **Data analytics models** Martínez-Plumed et al. [2] define a Data Science Trajectories (DST) model that builds on the CRISP-DM framework. For what types of projects would you use CRISP-DM? In contrast, when would an alternative model like DST be more suitable?
5. **Machine Learning, Safety, and Fairness** Review the source materials for one of the examples of unfairness in machine learning discussed in class, and answer the following questions.
   (a) Mehrabi et al. [3] define fairness in machine learning as "absence of any prejudice or favoritism toward an individual or group based on their inherent or acquired characteristics". Using this definition, how does the application of the machine learning approach result in an unfair outcome?
   (b) What practices, if any, mitigated the impact of the algorithm's unfairness?
   (c) What lessons can be learned from this example?

References

[1] William J. Frawley, Gregory Piatetsky-Shapiro, and Christopher J. Matheus. "Knowledge Discovery in Databases: An Overview". In: *AI Magazine* 13.3 (1992), p. 57. DOI: 10.1609/aimag.v13i3.1011. URL: https://ojs.aaai.org/index.php/aimagazine/article/view/1011.

[2] Fernando Martínez-Plumed et al. "CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories". English. In: *IEEE Transactions on Knowledge and Data Engineering* (Dec. 2019). ISSN: 1041-4347. DOI: 10.1109/TKDE.2019.2962680.

[3] Ninareh Mehrabi et al. "A survey on bias and fairness in machine learning". In: *ACM Computing Surveys (CSUR)* 54.6 (2021), pp. 1–35.