Data Science Toolbox Question Sheet

04.2 Outliers and Missing Data

Daniel Lawson

Block 4

Short Questions

1. In what circumstances might measuring anomalies using p-values be helpful, or dangerous?
2. A colleague claims that regression is particularly useful for detecting anomalies because you can access leverage of each data point. Discuss.
3. A colleague used a density-based approach based on k-Nearest Neighbours to detect several clusters in their data, as well as a set of "outliers" in low density regions. In what sense are these outliers?
4. How can you know that you have correctly dealt with missing data?
5. Describe 3 types of missing data, and order them in terms of the difficulty to handle correctly.
6. Describe how model-based inference can be used to deal with missing data.
7. Describe what is meant by "imputation" and give two examples, explaining when each would be used.
8. What is conservative imputation and how does it differ compared to mean imputation? In an example, describe an imputation process that is conservative.
9. What is the difference between complete-case analysis and available case analysis? Under which circumstances are each appropriate?
10. You have collected logs of connections to a particular machine on your network. Amongst other traffic, they contain a number of connections from a particular IP address, all with size zero. Consider the mean size with and without removing these duplicated events. Which is the more accurate measure of connection size, and why?