Outliers and Missing data (Part 1, Outliers)

Daniel Lawson University of Bristol

Lecture 04.2.1 (v1.0.1)

Signposting

- How do we identify **Bad Data**? That is, data that is misleading either due to missingness out atypicality.
  - This is one of the key ways that **Data Science Goes Wrong**.
  - Most researchers and practitioners do less than they should to understand their data.
- We use several approaches from previous lectures; this is as early in the course as it fits.
- This is part 1 of Lecture 4.2:
  - Part 1 is about outliers,
  - Part 2 is about missing data.

Intended Learning Outcomes

- ILO1 Be able to **access and process cyber security data** into a format suitable for mathematical reasoning
- ILO2 Be able to **use and apply basic machine learning** tools
- ILO3 Be able to make and report appropriate inferences from the results of applying basic tools to data

Bad Data: Missing and Misleading data

- The most time-consuming part of any real-world data analysis is **data cleaning**.
- This takes two main forms:
  - **Imputing** missing data where possible
  - **Removing** bad data where necessary
- It is **vital** that this is handled properly in order to gain appropriate insight from data.

Quality Control: Diagnosing bad data

- Most of **QC** is about figuring out whether your data are really what you thought they were.
  - Did you **measure** what you set out to measure?
  - Are there **systematic effects** that were unexpected?
- In many disciplines there are well-defined ways to spot issues.
- Cyber data tends to be more bespoke and therefore the problems are more unique.

Problems associated with Cyber data

- Cyber data is pretty poor!
- Typical problems include:
  - **Mass dropout**: whole sections of data missing, due to failure or system overload
  - **Feature dropout**: Some characteristic of the data is not captured properly for all or a subset of the data. For example, UDP packet sizes reported as 0
  - **Change in character**: if the data change due e.g. to an update, the data recording mechanism may not track this resulting in any of the problems above
  - **Unexpected data**: Much data is reported as an accumulation of something. If e.g. the termination condition is missed, a hash key duplicated, or the data unexpectedly large, reporting of the data can be wild.

Statistical tools for bad data

- There are two main tools available:

1. **Exploratory Data Analysis**
   - Does it look generally look the way it should?
   - Methods involve both plots and data summaries
   - We looked at this in Block 1
2. **Outlier Detection**
   - What specific parts of the data look unusual?
   - Methods focus on anomoly detection

Key questions to ask

1. Do my data contain important **missingness**?
   - What aspects of the truth am I not seeing?
   - How would I know?
   - What impact could missingness have on my analysis?
2. Do my data containing important **outliers**?
   - What do we mean by an outlier?
   - What impact will they have on my subsequent analysis?
   - What should I do about them?

Example: Not Missing At Random

- **Q1** of the workshop.

```
library("knitr")
conndataM=conndata
for(i in c(9,10,11,16:19))
  conndataM[,i]=as.numeric(conndataM[,i])
for(i in c(7,8)) conndataM[,i]=as.factor(conndataM[,i])
mtab=table(data.frame(
    missingduration=is.na(conndataM[,"duration"]),
    proto=conndataM[,"proto"]))
```

Anomaly Detection

- Anomaly detection uses the core methods we have seen throughout.
- For example, Density estimation (Block 4), cluster analysis (Block 3), regression (Block 2), etc.
- These models:
  - provide a baseline measure of **what is Normal**?
  - Against which **Unusual** is measured.

Measuring “Unusual” with p-values

- It is straightforward to use any model that can output a p-value as a measure of anomaly.
- Since a p-value is a Uniform random variable under the null, there is a wide literature available to assess whether the dataset as a whole is anomalous.
- **The problem**: In any cyber dataset, there is no plausible null hypothesis.
  - The data will “look weird” by any statistical measure.

Measuring "Unusual" with descriptive statistics

- **Thresholding**:
  - We saw in the "boxplot" that outliers were defined as all observations at least $3/2$ IQR above Q_3 or below Q_1.
  - This comes from reasoning about Normal distributions. However, the idea of thresholding based on intuition is probably the most common way to proceed.
  - Thresholding can be applied to p-values when they are not interpreted literally.
  - Removed values should be investigated to understand why they are unusual.
- Thresholds might be obtained by:
  - reference to other datasets,
  - theory,
  - bootstrapping,
  - ... etc!

Example: diagnosing outliers

- **Activity 2** of the workshop.

```
thist=hist(conndataM[,"duration"],breaks=101)
plot(thist$mids,thist$density,log="y",type="b",
     xlab="duration",ylab="histogram density")
```

Measuring "Unusual" with models

- Many modelling paradigms **explicitly handle outliers**. Some examples:
- Regression:
  - Measure leverage of each point (not always the same as outliers)
  - **Robust regression** methods fit better in the presence of outliers
- **Density-based** clustering (DBSCAN)
  - Points in low density regions may be outliers
  - An empirical p-value can be constructed from the set of points in lower-density regions.
- **Isolation Forests**
  - Random Forest-based technique (covered later).
  - Based on identifying "points that are easy to distinguish with a decision tree".
- Many other methods offer $Pr(data|model)$.

Duplicates and sample density

- **Sample density** obviously affects inference.
  - It is desirable that the sampling density reflects the density of the data to be predicted.
- Missing data often makes many records, that **should otherwise be different**, appear the **same**.
  - This dramatically affects density estimation.
- One solution is to work only with unique records.
  - This solves some types of bias but not others, e.g. overrepresentation of particular regions of continuous variables.
  - No longer a density, but a **plausible region**.

Batch and similar effects

- Correlation analyses of features with properties of the data that should not matter are a vital tool in Quality Control.
- Some quantities are known apriori not to affect some feature.
  - For example, if data are observed in batches, the batch number shouldn't matter.
  - In regression analyses, minor batch effects can be regressed out (included in the model).
  - Major batch effects require the data to be discarded or treated specially.
- As always, **Correlation $\neq$ Causation**.
  - So observing that e.g. different sources of data have different structures of traffic going over them isn't a smoking gun for a QC problem.
  - e.g. in Cyber data, they might measure different sorts of traffic.

Example of batch effects

- Is there a batch effect by day?
- **Activity 3** of the workshop:

```
conndataM[,"day"]=1 + (conndataM[,"ts"]>daybreak)
conndataM[,"logduration"]=log(conndataM[,"duration"])
lm1=lm(logduration~proto+service +ts +id.resp_p+day,
       data=conndataM[conndataM[,"duration"]<1200,])
summary(lm1)
```

Robust algorithms

- Most algorithms have robust alternatives, e.g.
  - Robust regression, (quantile regression),
  - Robust clustering,
  - Robust Kernel Density Estimation,
  - ... etc. Find one for your problem.
- Generally, robustness comes at a **cost**:
  - Increased computational complexity due to e.g. lack of integrability: e.g. Normal kernel replaced by Laplace,
  - Harder optimisation problem, e.g. more local minima, **non-convex solution**,
  - Or just not the model you wanted?
- **Robustness is not a general property** but defined with respect to some class of models.
  - There are many different "Robust algorithms for X" with different properties.
- "Too many" outliers will change the model anyway. How many is too many?

Removing outliers

- "An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism." Charu Aggarwal, IBM Research
- When outliers are detected, what should you do with them?
- Should we switch to a robust algorithm and take the hit?
- Or remove outliers for the purpose of model building?
- Or add an "outlier model", e.g. a larger normal distribution in Gaussian Mixture Modelling?

Reflection

- How do we know that the class of outliers detected is the “right” ones?
- Do we expect more outliers in a test dataset?
- How might we test that an algorithm is the “right kind” of robust?
- By the end of the course, you should:
  - Be able to check data for outliers,
  - Be able to perform basic outlier detection,
  - Be able to reason about what outlier removal will do.

Signposting

- Further Reading:
  - "A Survey of Outlier Detection Methodologies" by Victoria Hodge & Jim Austin, Artificial Intelligence Review 22:85–126 (2004).
  - Outlier Analysis by Charu C. Aggarwal. NB: Not freely available.
  - Chapter 10 of The Elements of Statistical Learning: Data Mining, Inference, and Prediction (Friedman, Hastie and Tibshirani) discusses the robustness to outliers for various methods.