Lab 2 - Data Exploration Exercise  
Sept. 07, 2018

Lab Aims

The purpose of this lab is to simulate the initial few steps you, as an aspiring healthcare researcher would take to explore a new dataset. In this case, you have been given a dataset designed to investigate the effectiveness of indwelling arterial catheters (IAC) in hemodynamically stable patients with respiratory failure for mortality outcomes.  

I have also thrown in a few extra data quality issues, feel free to get my attention should you find any of them.

Data Understanding - Acquisition and Description

To begin: download the lab notebook (Sakai, Webpage), upload it to your notebook directory and use pandas to read in the provided dataset into a dataframe object.

• Q1: Print out the total number of instances and features.
  - For reference, the data is expected to have 1737 instances and 14 attributes.
  - If not, describe what has happened, and correct it.

• Q2: Explore the feature space. Do the types of features align with the data dictionary provided here?
  - If not, identify the offending instance(s) and replace the instance with np.nan
  - You will then need to retype the column (it should be a float)

Summary Statistics and Visualization

One common first step for analysis is always to summarize and visualize the data. It is also a quick way to get a rough idea of potential outliers.

• Q3: Provide the Mean, Standard Deviation Range, Min/Max for the following:
  - Length of Stay
  - First Hemoglobin

  \(^1\) I have also thrown in a few extra data quality issues, feel free to get my attention should you find any of them.

  \(^2\) Hint: Look for duplicated features and blank rows (i.e. those with a high percentage of missing data). Rows that are over >90% missing can be dropped.

  \(^3\) Can you do this in a single function call?
- First SAPS I score
- First White blood cell count

- **Q4:** Compute the arithmetic and geometric means for length of stay. How do they differ?
- **Q5:** Create a scatterplot matrix between the four values above (color points by mortality outcome)\(^4\)
- **Q6:** Group-by: Day of the week and compute the mean first SOFA value

---

**Missing Data and Noise**

Before starting any analysis it is important to filter out points that are obvious noise, or correct general errors, and to understand the amount of data available after cleaning\(^5\)

- **Q7:** Identify the level of missing data for each feature
- **Q8:** Are there obvious instances of noise? Check for impossible values, or extreme values.  \(^6\)
- **Q9:** How about inconsistent values?  \(^7\)

---

**Reduction and Transformation**

We will begin by looking at the various ways in which we can sample our dataset. While likely not needed during exploration of this small data, it is critically important during validation of models.

- **Q10:** Sample 50% of the dataframe, how many patients died?
- **Q11:** Sample exactly 174 instances, how many patients died?

Next it is often necessary to break continuous data into nominal or ordinal scales to address various research questions (high vs. low performers, patients who live vs. those who die, etc.)

- **Q12:** Discretize the Age columns into quartiles (4 equal components).
- **Q13:** Provide the cut off values for each quartile, and the mean of the data contained in each quartile\(^8\).

---

\(^4\) I recommend the use of the Seaborn package, but you can do this natively in Pandas as well with the use of a dictionary mapping outcomes to colors.

\(^5\) Note, it is not uncommon for this to result in a much smaller subset!

\(^6\) Hint, I’ve limited these to continuous data. Think both practically and clinically possible

\(^7\) I’ve limited this to nominal features. Use the pandas’ unique function to quickly check

\(^8\) The Pandas groupby function can make this very easy, just make sure you add the transformed Age values into the dataframe
Similarity

Similarity forms the basis for many aspects of analysis. From improvement in filling missing data, to more advanced forms of discretization, to patient groupings, and even with some predictive models themselves.

- **Q14:** Compute the similarity between patients and index 0 and 1 on the following columns, for this problem do not worry about the scale of data
  - First SOFA
  - First SAPS
  - Age