Data Analytics for Personalized Genomics and Precision Medicine

L5: Data Exploration and Data Cleaning

Lecturer : Yu LI (李煜) from CSE
Liyu95.com | liyu@cse.cuhk.edu.hk

Scripting Author: Kevin Jesus Martinez
1155255965 | 1155255965@link.cuhk.edu.hk

Main Points

1. Genome Assembly
2. Data Cleaning
3. Data Exploration

Genome Assembly

Since machines are only able to obtain relatively short sequence readings, a process for making those readings into the whole genome sequence is needed, this process is called genome assembly.

Despite nowadays technology has made us able to sequence longer pieces of a genome with novel methodologies like nanopore sequencing, an efficient genome assembly is still necessary. This process resembles a jigsaw puzzle, where overlap regions play an important role indicating where each pieces goes.

However, many problems arise on assembly process, such as:

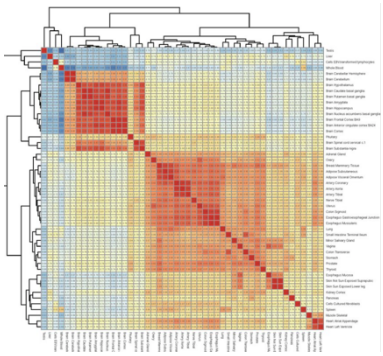
- Mutations
- Repeated sequences
- Determining the copy number of repeated genes
- Computing and time limitation

Mapping Example

We have a sequence of DNA, and several nucleotides combinations that form an aminoacid (triplets).

Each triplet would try to find a place on the sequence where it has a highest similarity. Then, the triplets with the highest similarity to the gene expression of the sequence (orange section) will be those in who we are interested.

This process would eventually take us to form a gene expression matrix, a potential project to elaborate.



Gene expression matrix

Data Cleaning

Before any other thing, a cleaning of the data that will be used is essential for obtaining correct results. Dirty data can be caused by missing values, noise and outliers, duplicate data, unnormalized data and mixed data. The cleaning process should address the mentioned obstacles in the following order:

Noise

Noise is the modification of an original value, like the distortion of the voice on a voice call or the signal lost during a telecommunication process. Fixing this problem usually involves the mean, the tendency or the expected value. Knowing if the noise follows a normal or gaussian distribution could be useful to find a more precise denoising method.

Outliers

An outlier data is a value with characteristics that are considerably different than most of the other data, like a rain on the dessert or a very short person. Although outliers can disturb the exploration and understanding of the data, they can also be informative, in some cases revealing tendencies or problems that might not be visible by the rest of the data. Hence, treating outliers can eventually be struggling, since care must be taken when deciding what to do with them.

A useful method consists of trying to find the cause of the outlier and then, if it is a typo or an irrelevant data, treat it as a missing value. But if it is not, then including it in the analysis could be a good decision.

Missing Values

Missing values usually occur when information is not collected, or the attributes may not be applicable to some cases. To handle them, these strategies can be followed:

- Eliminate the complete object.
- Estimate the missing value by considering the probabilities.
- Simply ignore it (not often recommended)

Duplicates

Usually provoked by merging tables with common registers. Duplicates are easy to fix, being important to eliminate them as they can change the distribution of the data.

Unnormalized Data

It is common to work with variables with different units, leading to a big disparity of magnitudes range, this can cause a variable to dominate over others when comparing them, making it necessary to convert all variable into the same unit.

Two normalization techniques are presented, although they are not the only ones:

1. Min-Max Normalization

$$v' = \frac{v - v^{min}}{v^{max} - v^{min}}$$

2. Z-Score Normalization
(Gaussian distribution assumed)

$$v' = \frac{v - \boldsymbol{Mean}(v)}{\mathbf{Std}(v)}$$

Normalization is usually done along the columns because these are random variables which follow the same distribution. For gene expression matrix, however, normalization can be across the columns or the rows, since both have the same distribution.

Alternatively, when facing categorial data, a process called one-hot-encoding can be executed, where binary variables (columns) are created for each category.

Data Exploration

Exploring data permit us to understand the underlying story of it and find the best way to proceed with whatever finality we have. There are two main ways to understand data, with summary values and with images.

Summary Statistics

These are numbers that summarize properties of data. Summarized properties include frequency, mean and standard deviation, median, range and the interquartile range.

1. Mean: Also known as average, it's the basic statistic to understand data.

$$\bar{x} = \frac{\sum_{i=0}^{n} x_i}{n}$$

2. Frequency: Used mainly in categorical data, indicates how many times a value is repeated.

3. Standard Deviation: It's the square root of the variance, summarizing how much difference exist between each value in the dataset.

$$\sigma = \sqrt{\frac{\sum_{i=0}^{n}(x_i - \bar{x})^2}{n-1}}$$

4. Median: Also known as trimmed mean, it's a fine way to summarize data full of outliers. It can be computed ordering the data from lower to higher and taking the one on the middle (if there are two on the middle, is the mean of these two).

5. Range: The range of possible values the data has.

$$Range = \max - min$$

6. Percentiles: The p-th percentile is a value of x such that $p\%$ of the observed values of x are less than x_p.

Visualization

Conversion of data into a visual or tubular format. Especially useful to note the general behave of data, along with patterns and possible errors and outliers.

Some of the most used visualization tools are histograms, boxplots, maps, heatmaps, matrices, scatter plots, bubble plots, bar graphs and pie graphs. The use of each of them would depend of the kind of data we are handling.