DATA PREPROCESSING

This week we will take a deeper look at the cancer data file. Specifically, we will perform *data preprocessing* for the purposes of quality control and understanding.

**Materials – readings and videos**:

* Dr. Welch’s Introductory Lecture
  + <https://youtu.be/Z2bdf-vPfDY>
* Data preprocessing:
  + Sections 2.1-2.6 – Data Preprocessing - Intro to data mining, *Discovering Knowledge in Data*. D.R. Larose and C.D. Larose. Wiley. 2014.

<https://alice.library.ohio.edu/record=b5187242?> (click on the link “Connect to resource OhioLink”)

* + <https://en.wikipedia.org/wiki/Data_pre-processing>
* Visualization and Statistics:

1. <https://en.wikipedia.org/wiki/Scatter_plot>

C**oncepts to learn from the materials:**

1. Why preprocess data?
2. What is *data cleaning*?
3. What is an *outlier*?
4. Name and describe *graphical methods* for identifying outliers.

**Quiz**:

After learning the concepts listed above, complete the blackboard quiz no later than Monday August 30, 9:39 AM. The quiz will cover your understanding of the *concepts to learn from the materials* (see above). The quiz may include multiple choice, true-false, fill-in-the-blank, and/or matching questions.

Data Mining Activity: (*to be started after you complete the quiz*)

**Due date**: no later than Thursday Sept. 2, 11:59 pm

(submit by email to welch@ohio.edu)

Throughout the course, we will analyze a data file that lists the genetic mutations of individuals who participated in a research study (the data file is provided on Blackboard in the week 1 folder). We will employ data mining techniques to discover possible genetic causes of cancer by analyzing the information in the data file. This week we will perform *data preprocessing*.

Outlier detection is an important task to perform during the *data preprocessing phase*. Use scatterplots to determine if there are (1) any outlier samples and (2) if there are any outlier genetic mutations. Submit your scatterplots and discuss whether you observe any samples or any genetic mutations that are outliers (if so, explain why they are outliers). Specifically, generate and analyze the following scatterplots:

(1) to identify *outlier samples*:

1. y-axis: number of mutations per sample
2. x-axis: samples

(2) if there are any *outlier genetic mutations*:

1. y-axis: number of samples per mutation
2. x-axis: mutations

To gain a deeper understanding of the patterns contained in the data file, you will explore how the individual genetic mutations relate to the two categories of samples. You will perform this exploration by making a feature table as illustrated in Table 1.

Table 1. Partial feature table showing the number of times each genetic mutation occurs in each group.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Genetic Mutation (Feature)** | **T** | **C** | **NC** | **%C** | **%NC** | **C-NC** | **C/NC** | **%C-%NC** | **%C/%NC** |
| FARP1\_GRCh37\_13:99092237-99092237\_Frame-Shift-Del\_DEL\_G-G-- |  |  |  |  |  |  |  |  |  |
| QKI\_GRCh37\_6:163987695-163987695\_Intron\_DEL\_T-T-- |  |  |  |  |  |  |  |  |  |
| NHLRC2\_GRCh37\_10:115662308-115662308\_Frame-Shift-Del\_DEL\_A-A-- |  |  |  |  |  |  |  |  |  |

Where the column labels are defined as follows:

* **T** = total number of samples that contain the mutation
* **C** = number of cancer samples that contain the mutation
* **NC** = number of non-cancer samples that contain the mutation
* %**C** = percentage of cancer samples that contain the mutation
* %**NC** = percentage of non-cancer samples that contain the mutation
* %**C** - %**NC** = the difference of cancer to non-cancer samples that contain the mutation
* %**C** / %**NC** = the ratio of cancer to non-cancer samples that contain the mutation

Provide a summary feature table, showing the top 10 features (mutations) selected by each of the following measures:

* Column 1: top 10 features, when ranked by **T**
* Column 2: top 10 features, when ranked by **C**
* Column 3: top 10 features, when ranked by **NC**
* Column 4: top 10 features, when ranked by %**C**
* Column 5: top 10 features, when ranked by %**NC**
* Column 6: top 10 features, when ranked by %**C** - %**NC**
* Column 7: top 10 features, when ranked by %**C** / %**NC**

Submit an email to [welch@ohio.edu](mailto:welch@ohio.edu) that contains a brief report that discusses your findings. Additionally, attach either (a) the .csv file (it should contain the original data, along with the calculated values), or (b) the computer program that you developed for this activity and the output of your program (either a screenshot or a file).

**NOTE**: I may respond to your email submissions with questions about your methods, results, and/or interpretation. Please respond promptly to my questions.