CLASSIFICATION

This week we will begin to focus on the problem of developing a genetic test that can identify individuals who may develop cancer.

**Materials**:

* Dr. Welch’s Introductory Lecture
  + <https://www.youtube.com/watch?v=NA84liIG-YI>
* Chapter 7.1 - Classification Task

*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/Binary_classification>
* <https://en.wikipedia.org/wiki/False_positives_and_false_negatives>
* <https://en.wikipedia.org/wiki/Confusion_matrix>
* <https://en.wikipedia.org/wiki/Bar_chart>
* <https://chartio.com/learn/charts/stacked-bar-chart-complete-guide/>
* <https://en.wikipedia.org/wiki/Pie_chart#Doughnut_chart>

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

* the classification task
* binary classification
* TP, FP, TN, FN
* confusion matrix
* stacked bar chart
* donut chart

**Quiz**:

After learning the concepts listed above, complete the BlackBoard quiz no later than Monday Sept 6, 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.

**Supplementray Materials**:

* <https://www.ted.com/talks/jim_collins_how_we_re_using_ai_to_discover_new_antibiotics/up-next>
* <https://en.wikipedia.org/wiki/Statistical_classification>

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

**Due date**: no later than Thursday Sept. 9, 11:59 pm

(submit by email to [welch@ohio.edu](mailto:welch@ohio.edu))

Perform the following classification tasks on the 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).

* + - 1. Generate a confusion matrix for each of the following genetic mutations:
* RNF43\_GRCh37\_17:56435161-56435161\_Frame-Shift-Del\_DEL\_C-C--
* RPL22\_GRCh37\_1:6257785-6257785\_Frame-Shift-Del\_DEL\_T-T--

HINT: Assume that you’re using a specific genetic mutation as a diagnostic test for cancer.  For example, you may select a mutation of the RNF43 gene to use as a test for cancer. If you use that test, how many times would you correctly diagnose an individual as having cancer? The number of times that you tell someone they have cancer, and they actually do have cancer, would be the number of *true positives*. What would be a *true negative* in this case?

Relatedly, how many times would you be wrong, if you use a mutation of the RNF43 gene as a test for cancer? The number of times that you tell someone they have cancer, but they actually don’t have cancer, would be the number of *false positives*. What would be a *false negative* in this case?

Given the answers to these questions, you have everything that you need to construct a confusion matrix. If your still uncertain, please review the following article from this week’s materials: <https://en.wikipedia.org/wiki/Confusion_matrix>.

* + - 1. Use stacked bar charts to represent the information contained in the confusion matrices that you generated for the RNF43 and RPL22 mutations (see above). One stacked bar chart should show TP and FP stacked on top of each other, and a second stacked bar chart should show TN and FN stacked on top of each other.
      2. Use a donut chart to represent the information contained in the confusion matrices that you generated for RNF43 and RPL22. Hint: use one ‘slice’ of the donut for each of the following: TP, FP, TN, and FN.
      3. Which of the two mutations listed above is more useful for classification of the **C** and **NC** samples? Explain your answer by using the results of questions 1, 2, and 3.
      4. If we use the quantity “**TP** – **FP**”to classify samples, which feature would be the most useful for classifying samples? (I.e., which genetic mutation yields the maximal value for the quantity “**TP** – **FP**”?) What is the maximum value of “**TP** – **FP**” across all features?
      5. If we use the quantity “%**TP** - %**FP**”to classify samples, which feature would be the most useful for classifying samples? (I.e., which genetic mutation yields the maximal value for the quantity “%**TP** - %**FP**”?) What is the maximum value of “%**TP** – %**FP**” across all features?

Submit an email to [welch@ohio.edu](mailto:welch@ohio.edu) that contains a brief report that includes your confusion matrices, stacked bar charts and donut chart, and which discusses how your findings relate to the problem of developing a genetic test that can identify individuals who may develop cancer. 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.