SFU faculty of science
department of mathematics

MACM 316- D100 COMPUTING REPORT 3

Due Date: Monday, July 11, 2022

- Please follow the *Guidelines for Computing Assignments* found on Canvas.
- The computing report grading scheme is also described there.
- Your one-page report is to be submitted via Crowdmark by 11:59pm on Monday, July 11, 2022
- Acknowledge any and all collaborations and assistance from classmates/TAs/instructor.

Begin by downloading the data file: `Data_CR3.mat` and the two example files: `Investigate1D.m` and `Example3D.m`

Warmup in 1D

Start by running the file `Investigate1D.m`. Based on the figures do the following:

- Notice the dataset $\{x_2, y_2\}$ contains all of $\{x_1, y_1\}$ and a lot of additional values of x_i where $y_i = 0$. How does this affect the linear fit? How does this affect the nonlinear fit? Which method would you say is more robust in this situation? Why?
- Use the fitting coefficients to classify the data $\{xC\}$. Does it matter which coefficients you use?
- Examine the surface of the error function used in the nonlinear fitting. What does the staircase structure suggest about how well steepest descents would work in this case?

Classification of tumours

The data set `Data_30D` contains results from 30 different tests related to diagnosing cancer. The data set `Data_3D` is a subset in only 3 variables.

Run the file `Example3D.m`. It performs linear and nonlinear fits and plots the results.

The plane in the linear fit is found by solving the equation:

$$\beta_0 + \beta_1 x + \beta_2 y + \beta_3 z = 1/2\ .$$

Classify the data `Classify_Data3D` by evaluating

$$P = \beta_0 + \beta_1 x_i + \beta_2 y_i + \beta_3 z_i$$

for each of the rows of `Classify_Data3D`. If $P > 1/2$ then it is type A, else it is B.

The plane in the nonlinear fit is found by solving the equation

$$\frac{1}{1 + e^{\beta_0 + \beta_1 x + \beta_2 y + \beta_3 z}} = 1/2 \qquad \Rightarrow \qquad \beta_0 + \beta_1 x + \beta_2 y + \beta_3 z = 0$$

Classify the data `Classify_Data3D` by evaluating

$$f(\vec{x}) = \frac{1}{1 + e^{\beta_0 + \beta_1 x + \beta_2 y + \beta_3 z}}$$

for each of the rows of `Classify_Data3D`. If $f > 1/2$ then it is type A, else it is B.

Explain which method you believe gives better results and why. (HINT: what are the errors in the linear and nonlinear fits?) Include a plot showing the training and unknown data and your classification of it.

Now do the same for the full data set and classify `Classify_Data30D`.

There are three possible outcomes of the classification:

- It is correct.
- A cancerous tumour is incorrectly classified as benign.
- A benign tumour is incorrectly classified as cancerous.

Suppose that $y = 0$ corresponds to cancerous tumours. How would you adjust the classification if you are willing to incorrectly classify some benign tumours as cancerous if it means few to no cancerous tumours are missed? You can include a plot of the 3D data using your new method.