Math 232 – Computing Assignment 4

Due Date: Dec 2, at 11:00pm.

You must upload to Crowdmark both your code (as a .pdf file) (to Code - Computing Assignment 3) and your report (to Report - Computing Assignment 3). The assignment is due at 11:00pm. If Crowdmark indicates that you submitted late, you will be given 0 on the assignment. Your computing report must be exactly 1 page. There will be a penalty given if your report is longer than one page.

- Please read the **Guidelines for Computing Assignments in Canvas** first.
- Keep in mind that Canvas discussions are open forums.
- Acknowledge any collaborations and assistance from colleagues/TAs/instructor.

Programming Preamble:

Matlab: `x=[1 1 1]'` produces a column vector. The ' indicates transpose.

Matlab: `help plot` gives you all the information you need to make your plots look good - labelling the axes, putting on a title etc.

Computing Assignment

Required submission: 1 page PDF report and Matlab or Python code (.m or .py respectively, exported as a .pdf) uploaded to Canvas.

Before you begin, consult the 'Least Squares preamble' file posted on Canvas

Machine Learning

This assignment will give you a glimpse into one of the main areas of Machine Learning (Data Science); that of regression and classification.

We study two modelling examples that use historical data on the behaviour of a specified group of people to make predictions on their future behaviour. In particular, we study student success in Math 232 based on their study habits (number of hours spent studying Math 232) and their grades in previous mathematics courses. (Note: all the data here is purely hypothetical.)

1. Our first approach to this problem will use (hypothetical) historical data on the number of hours per week students spent studying math 232 in a previous semester, and their subsequent grade in the course. What we want to obtain from this data is a criteria for "success" in math 232 (i.e., grade $> 50\%$) based solely on how much time a student spends studying the course materials per week.

Data for this assignment is contained in the excel file CA4data.xslx that is posted on Canvas. This (hypothetical) data set contains the number of hours per week students spent studying math 232 during a past semester, their grade in a previously taken math course (math 152, taken before they took math 232), and their final grade in math 232 in that semester. In Q1 below, we will use only the studying hours, and Q2 below will use both the studying hours and the grades in math 152 to build a model from which we can make predictions about student 'success' in math 232.

Read in the data file CA4data.xslx that is available for download on Canvas. Open the file in excel and remove the first line (which labels the columns). Use these commands to create (column) vectors with data taken from the columns in the file;

```
>> data = 'CA4data.xlsx';
>> D=xlsread(data);
>> h=D(:,1);
>> k=D(:,2);
>> g=D(:,3)
```

`h,k,g` are column vectors with the data #hours studied per week, grade in math 152, and the grade received in math 232, respectively.

The data set for the first question is $D1 = \{(h_i, g_i),\ i = 1, \ldots, 45\}$. We look for a linear model of the form $g = a + bh$. That is, we look for the best fitting (least squares) line through the data points $D1$ in $\mathbb{R}^2$.

1a. Let the function `gg=a+bx` (Matlab notation) be your best fit linear model through the data (here you have to first compute what those parameters `a, b` are using our theory of least squares approximation). The Matlab command

```
>> x=0:0.2:12;
>> gg=a+b*x;
plot(x,gg)
```

will plot this line.

Plot your best fit linear line along with the data set using these commands;

```
>> plot(h,g,'k.')
>> hold on
>> plot(x,gg)
>> hold on
>> axis([0 15 20 100])
>> hold on
>> xlabel('hours per week'); ylabel('grade')
```

From your linear model of hours spent studying vs final grade in math 232, what do you predict is the "success" point; that is, the (minimum) number of hours a student should spend per week studying math 232 in order to pass the course? (this will be a point h_o on the $h-$line (‘hours per week’) where $h \geq h_o$ indicates a grade of at least 50% will be obtained).

1b. Now find the best fit *quadratic* model to the data ; $g = a + b_1 h + b_2 h^2$. That is, the least squares quadratic that fits the data. Plot this quadratic along with the data points. From this quadratic find the predicted success rate. (Notice here that although we are fitting a *nonlinear* function to the data set, the least squares algorithm is *still* a linear problem!)

2. In this second approach we include another ’diagnostic’ from the group of students; that of their previously taken math 152 grade. So now we use *both* the number of hours they spent studying math 232 *and* their grade in math 152 to make a prediction on student success in math 232.

   Thus, our data set is $D2 = \{(h_i, k_i, g_i),\ i = 1, \ldots, 45\}$. You can take a look at this 3-dimension scatter plot with the commands;

   ```
   >> plot3(h,k,g,’k.’)
   >> grid on
   ```

   (Note that you can drag the mouse on the figure to rotate it and obtain different perspectives).

   Our linear model for this data set is $g = a + bh + ck$. That is, we look for the best fitting *plane* through the data points in $\mathbb{R}^3$.

   Compute (using the algorithm discussed in class) the least squares plane that fits the data points. You can plot the plane with the data points using these commands;

   ```
   >> plot3(h,k,g,’k.’)
   >> axis([0 15 20 100 20 100])
   >> grid on
   >> hold on
   >> [X,Y]=meshgrid(x,y);
   >> ggg=a+b*X+c*Y;
   >> mesh(ggg)
   ```

   where `a+b*X+c*Y` is your least squares fitting plane to the data.

   From this determine the ‘success line’ in the h, k-plane; one side of this line predicts the student will receive a grade of $\geq$ 50%. Plot this line in the h, k-plane indicating the ‘success’ side of it. (Hint: Recall that the intersection of two planes is a line (in 3 dimensions); then project this line down into the h, k-plane.)

   (Remark: In a similar way as one fitted the least squares quadratic curve through the 2 dimensional scatter plot as in Q1, one could here find the least squares *quadratic* surface $g = a_o + a_1 h + a_2 k + a_3 hk + a_4 h^2 + a_5 k^2$ through the 3 dimensional scatter plot. Again, it is a *linear* problem determining the parameters $a_1, \ldots, a_5$. We will NOT be doing that as part of this assignment.)