PHY00047M

MSc Degree Assessments 2023/24

Course:

DATA ANALYTICS AND MACHINE LEARNING ASSIGNMENT

Allocation of Marks:

This assignment is marked out of 20.

Instructions:

Answer ALL sections.

This is an open assignment. You may refer to notes, textbooks and online information. Submit typeset solutions (Jupyter Notebook) in a single PDF file with the following information at the top of the front page:

- The module code and assignment number (*e.g.* "PHY00047M Mini Assignment 1").
- The number of pages in the PDF.

Name your PDF using the module code and the assignment number or name, e.g. `PHY00047M_1.pdf`. **Ensure that no identifying information appears anywhere within your submitted file.** You may prepare the PDF by printing your Jupyter Notebook to a PDF (ensure all cells and code is visible). Any diagrams must be included in the PDF, and pages must be numbered.

You may not discuss this assignment, through direct or indirect means, with any student or any other person until the results and feedback have been released.

If – and only if – the VLE will not accept your PDF, you may e-mail it to phys-emergency-vle-submissions@york.ac.uk, ensuring that the subject line of your e-mail and the name of the attached file consist of the module code, and the assignment number.

Data Analytics and Machine Learning Assignment - PHY00047M

Answer ALL questions.

1 Problem 1

Given the data file data1.dat perform the following tasks (while explicitly giving the formulas used where appropriate, even if you use built-in functions):

1. Plot a histogram of the data and discuss the distribution and your choice of histogram range and binning. What is the mode of the distribution from a visual inspection? [2]
2. Determine the mean [1]
3. Determine the variance [1]
4. Determine the skew [1]
5. Discuss your findings with expectations from visually inspecting the histogram. [1]

2 Problem 2

Given the two-dimensional data file data2.dat perform the following tasks:

1. Plot a histogram of the data and discuss the distribution and your choice of histogram range and binning [2]
2. Determine the means [1]
3. Determine the variances [1]
4. Determine the covariance and the Pearson coefficient [1]
5. Discuss your findings with expectations from visually inspecting the histogram. [1]

3 Problem 3

In this problem we will work a with Pandas DataFrame and check some of its functionality. Given the data in data3.dat perform the following tasks:

1. Create a Pandas DataFrame that has the following values and indexes as shown

	Temp	Humidity	Time
Event1	22	50	10:00
Event2	25	55	11:00
Event3	24	60	12:00
Event4	27	57	13:00
Event5	21	58	14:00

Table 1: Data to be inserted in DataFrame

in Table 1.

[1]

2. Using Pandas functions, sort DataFrame in decreasing order of humidity save them in a new DataFrame and print it. [1]

3. Create an array (size 5) of random integers between 1 and 100. Then add this array to the above DataFrame [1]

4 Problem 4

Given the data in data3.csv perform the following tasks:

1. Read in the data in a Pandas data frame, and describe the data structure, including number of entries, columns. [1]

2. Provide data description for the included variables and their correlations [1]

3. Plot a contour of how variable 3 depends on variable 1 and 2 (Hint: Reshape variables as arrays of [100,100].) [1]

4. Using conditional filtering, create DataFrame that has all data if the variable 3 is larger than 0.0. How many rows does the new frame have now? [1]

5. How many values of variable 2 are larger than 1.0 when variable 3 is smaller than -0.5? [1]

End of paper