Student Assignment Brief

This document is intended for Coventry University Group students for their own use in completing their assessed work for this module. It must not be passed to third parties or posted on any website. If you require this document in an alternative format, please contact your Module Leader.

Contents:

The work you submit for this assignment must be your own independent work, or in the case of a group assignment your own groups' work. More information is available in the 'Assignment Task' section of this assignment brief.

Assignment Information

Module Name: Principles of Data Science

Module Code: 7144CEM

Assignment Title: Individual Coursework

Assignment Due: Thursday, 07/11/2024 at 6 pm UK time

Assignment Credit: 10 credits

Word Count (or equivalent): 2000 words equivalent (not including reference list or output)

Assignment Type: Written

Percentage Grade (Applied Core Assessment). You will be provided with an overall grade between 0% and 100%. You have one opportunity to pass the assignment at or above 40%.

Assignment Task

This *Individual Coursework* involves investigating the different properties and operations involving vectors, matrices, and probability, including Markov chains and Eigenfaces. You are encouraged to explore the topic, use your initiative, and show some originality, within the time available. Make sure you read each task through carefully and answer all parts of each task. Aim to demonstrate your understanding of the topics and the relevant module learning outcomes. *Please take every opportunity to check that you are working with matrices and vectors as 2D numpy arrays of the appropriate shape.*

Task 1. Markov Chains

A manufacturing company has 200 identical machines that are used to create some particular product. The possible state of each machine is Idle (awaiting work), Working (on a task), Broken, or in Repair. Each machine is monitored every hour to determine its status.

Suppose after each hour, 5% of the Idle machines remain Idle, 93% of Idle machines are now Working, and 2% of Idle machines are now Broken. For Working machines, 10% become Idle, 86% stay Working, and 4% become Broken. For Broken machines, 80% remain Broken and 20% become in Repair. For machines in Repair, 50% become Idle, 10% become Working, and 40% remain in Repair.

(1) Construct a 4×4 transition matrix A (as a numpy array) to model the state of the machines as a Markov chain.

[5 marks]

(2) Suppose that all 200 machines are initially Working. Use numpy to calculate how many machines we would expect to see in each state at the end of each hour for the next 24 hours. Use Python to plot your results on a line graph showing one line for each state, together with a legend. Clearly comment on what you observe. Use numpy to find the eigenvectors of the transition matrix A from part (1), and explain how one of these eigenvectors is related to the number of machines we would expect to see in each state in the "long-run steady state".

[20 marks]

(3) Suppose the machines are in the "long-run steady state" from part (2) when the manufacturing company is sold to new owners. The new owners introduce a policy that any Idle machine will be switched off and prepared to be sold.
   (a) Modify the transition matrix A from part (1) to include this new policy and estimate the number of hours until there are fewer than 20 Working machines.
   (b) The expected number of hours needed for a machine starting from state i to reach state k for the first time can be calculated using matrix operations and is denoted μ_{ik}. For a particular destination state k, let N be the transition matrix from part (3)(a) but with row k and column k both deleted, and let I be the 3×3 identity matrix. The sum of each column of the matrix $(I - N)^{-1}$ gives the values μ_{ik}. Use this method to find the expected time needed for a machine to become Idle (and therefore switched off) for the first time, for each of the other possible states that a machine could start from. Clearly comment on what you observe.

[15 marks]

(4) Explain how this Idle-Working-Broken-Repair Markov chain model could be adapted to model the spread of an infectious disease in a population (assuming that the birth and death rates are roughly equal over time). Critically identify one weakness of using your (or any) Markov chain to model the spread of an infectious disease.

[10 marks]

Task 2. Eigenfaces

The Olivetti Faces dataset consists of 400 images of human faces. Five examples are given below.

Each image is 64 pixels wide by 64 pixels high. However, the Python code below assembles the 400 images as columns of the matrix `Xall`. Each image is greyscale (not colour) and all pixel values are between 0 and 1. The matrix `Xsub` is the first 200 columns of the matrix `Xall`.

The only Python libraries you may use in this task are numpy and matplotlib, except that you may use sklearn only to load the dataset (as in the Python code below).

```
import sklearn.datasets
faces = sklearn.datasets.fetch_olivetti_faces()
Xall = faces.data.T
print(Xall.shape)
print(Xall.min(), Xall.max())
Xsub = Xall[:,0:200]
print(Xsub.shape)
```

(1) Use numpy to find the mean of each row of `Xsub` and reshape as a 4096×1 vector (call this `xbar`). Adapt the Python code below to visualise this "mean face" and carefully compare it to the image stored in the column of `Xsub` corresponding to the last two digits of your Student ID. Clearly comment on what you observe.

```
import matplotlib.pyplot as plt
plt.imshow(**something**, cmap=plt.cm.gray, vmin=0, vmax=1)
```

[10 marks]

(2) Use numpy to calculate the covariance matrix of `Xsub`, giving a 4096×4096 matrix (call this matrix C). Then use numpy to find the eigenvalues (V) and eigenvectors (columns of the matrix P) of C using `np.linalg.eigh()`. Use the Python code below to reverse the entries in V and the columns of P so that the eigenvalues are in descending order. Use numpy to confirm that P is an *orthogonal* matrix (do not attempt to find P^{-1}). The eigenvectors of C (i.e. the columns of P) are called "eigenfaces" (Turk and Pentland, 1991). Visualise those eigenfaces corresponding to the 8 largest eigenvalues (be careful with `vmin` and `vmax`). Clearly comment on what you discover in terms of facial features and differences between the images.

```
V = V[::-1]
P = P[:,::-1]
```

[20 marks]

(3) In the following Python code, we apply P^T (as a matrix transformation) to the faces in `Xsub`, where `xbar` is the "mean face" (a 4096×1 vector) from part (1).

```
Ysub = P.T@(Xsub-xbar)
```

(a) Show (by example) that if `y` is any one column of `Ysub` then `P@y+xbar` perfectly recreates the corresponding column of `Xsub`. *This shows that each "face" in* `Xsub` *is a "linear combination" of eigenfaces plus the "mean face".*

(b) Choose a photo of your own face, crop it in a suitable way, resize to 64×64 pixels, and save as a ".pgm" file image format.

You may wish to use the free GIMP software (https://www.gimp.org/) to help with this process. The following GIMP menu selections may be useful:

- *Tools | Transform Tools | Crop [adjust the size on the left to be square]*
- *Image | Mode | Grayscale*
- *Image | Scale Image | set to 64x64 pixels*
- *Export as | name file with ".pgm" extension*

Then use `plt.imread()` to load the pgm image as a 2D numpy array and call it `image_me`. Then let

```
y_me = P.T@(image_me-xbar)
```

Although P is a 4096×4096 matrix, generally only those eigenfaces corresponding to the largest eigenvalues of C are "useful". If there are k such eigenvalues (with $k \geq 2$), you can partially reconstruct the image of your own face using only the first k eigenfaces, i.e., the eigenfaces corresponding to the largest k eigenvalues. Investigate how well these partial reconstructions approximate your original face image for different values of k by using `np.linalg.norm()` to provide a measure of "goodness of fit". Build an appropriate line graph as part of your investigation and show the final "best fit" partial reconstruction image. Critically discuss what you observe.

[20 marks]

Reference
Turk, M.A. and Pentland, A.P. Face recognition using Eigenfaces. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, June 1991, pp 586-591. https://ieeexplore.ieee.org/document/139758

Submission Instructions:

Ensure that your coursework is all your own work and you clearly cite and reference any sources you have used using APA style referencing. Please include both in-text citations and a list of references for each task (where relevant). **No collaboration with other students is permitted.**

Please submit one report (e.g. as a single Microsoft Word document or a single PDF document) covering all of the tasks above, clearly organised by subtask. Start each task on a new page. Make sure you include your Python code and relevant output/plots directly in the report. You must not submit a zip file. You must not submit a Jupyter notebook (but you can print a Jupyter notebook to a PDF file and submit the PDF file). Do not use screenshots to include Python code or text output into your report. If your report is a Microsoft Word file, then please use a syntax highlighter (such as http://hilite.me/) and copy-and-paste the coloured Python code into your report. Also simply copy-and-paste text output into your report.

Submission is online via Aula using the submission box provided. *Do not leave uploading too late.*

Marking and Feedback

How will my assignment be marked?

Your assignment will be marked by the module team.

How will I receive my grades and feedback?

Provisional marks will be released once internally moderated.

Feedback will be provided by the module team alongside grades release. Students will be able to access their feedback via Aula/Turnitin.

Your provisional marks and feedback should be available within 2 weeks (11 working days).

What will I be marked against?

Details of the marking criteria for this task can be found at the bottom of this assignment brief.

Assessed Module Learning Outcomes

The Learning Outcomes for this module align to the marking criteria which is provided above. Ensure you understand the marking criteria to ensure successful achievement of the assessment task. The following module learning outcomes are assessed in this task:

ML01. Demonstrate systematic knowledge and critical understanding in topics in linear algebra, probability and statistical models, relevant to data science.

Assignment Support and Academic Integrity

If you have any questions about this assignment please see the Student Guidance on Coursework for more information.

Spelling, Punctuation, and Grammar:

You are expected to use effective, accurate, and appropriate language within this assessment task.

Academic Integrity:

The work you submit must be your own, or in the case of groupwork, that of your group. All sources of information need to be acknowledged and attributed; therefore, you must provide references for all sources of information and acknowledge any tools used in the production of your work, including Artificial Intelligence (AI). We use detection software and make routine checks for evidence of academic misconduct.

Definitions of academic misconduct, including plagiarism, self-plagiarism, and collusion can be found on the Student Portal. All cases of suspected academic misconduct are referred for investigation, the outcomes of which can have profound consequences to your studies. For more

information on academic integrity please visit the Academic and Research Integrity section of the Student Portal.

Support for Students with Disabilities or Additional Needs:

If you have a disability, long-term health condition, specific learning difference, mental health diagnosis or symptoms and have discussed your support needs with health and wellbeing you may be able to access support that will help with your studies.

If you feel you may benefit from additional support, but have not disclosed a disability to the University, or have disclosed but are yet to discuss your support needs it is important to let us know so we can provide the right support for your circumstances. Visit the Student Portal to find out more.

Unable to Submit on Time?

The University wants you to do your best. However, we know that sometimes events happen which mean that you cannot submit your assessment by the deadline or sit a scheduled exam. If you think this might be the case, guidance on understanding what counts as an extenuating circumstance, and how to apply is available on the Student Portal.

Administration of Assessment

Module Leader Name: Mark Johnston

Module Leader Email: ad4039@coventry.ac.uk

Assignment Category: Written

Attempt Type: Standard

Component Code: CW

Generic Marking Rubric

PG

Mark band	Outcome	Guidelines
90-100% Distinction	Meets learning outcomes	Distinction - Exceptional work with very high degree of rigour, creativity and critical/analytic skills. Mastery of knowledge and subject-specific theories with originality and autonomy. Demonstrates exceptional ability to analyse and apply concepts within the complexities and uncertainties of the subject/discipline. Innovative research with exceptional ability in the utilisation of research methodologies. Demonstrates, creativity, originality and outstanding problem-solving skills. Work completed with very high degree of accuracy, proficiency and autonomy. Exceptional communication and expression demonstrated throughout. Student evidences the full range of technical and/or artistic skills. Work pushes the boundaries of the discipline and may be strongly considered for external publication/dissemination/presentation.
80-89% Distinction		Distinction - Outstanding work with high degree of rigour, creativity and critical/analytic skills. Near mastery of knowledge and subject-specific theories with originality and autonomy. Demonstrates outstanding ability to analyse and apply concepts within the complexities and uncertainties of the subject/discipline. Innovative research with outstanding ability in the utilisation of research methodologies. Work consistently demonstrates creativity, originality and outstanding problem-solving skills. Work completed with high degree of accuracy, proficiency and autonomy. Outstanding communication and expression demonstrated throughout. Student demonstrates a very wide range of technical and/or artistic skills. With some amendments, the work may be considered for external publication/dissemination/presentation
70-79% Distinction		Distinction - Excellent work undertaken with rigour, creativity and critical/analytic skills. Excellent degree of knowledge and subject-specific theories with originality and autonomy demonstrated. The work exhibits excellent ability to analyse and apply concepts within the complexities and uncertainties of the subject/discipline. Innovative research with excellent ability in the utilisation of research methodologies. Work demonstrates creativity, originality and excellent problem-solving skills. Work completed with very consistent levels of accuracy, proficiency and autonomy. Excellent communication and expression demonstrated throughout. Student demonstrates a very wide range of technical and/or artistic skills.

60-69% Merit		Merit - Very good work often undertaken with rigour, creativity and critical/analytic skills. Very good degree of knowledge and subject-specific theories with some originality and autonomy demonstrated. The work often exhibits the ability to fully analyse and apply concepts within the complexities and uncertainties of the subject/discipline. Very good research evidence and shows very good ability in the utilisation of research methodologies. Work demonstrates creativity, originality and problem-solving skills. Work completed with very consistent levels of accuracy, proficiency and autonomy. Very good communication and expression demonstrated throughout. Student demonstrates a wide range of technical and/or artistic skills.
50-59% Pass		Pass - Good work undertaken with some creativity and critical/analytic skills. Demonstrates knowledge and subject-specific theories with some originality and autonomy demonstrated. The work exhibits the ability to analyse and apply concepts within the complexities and uncertainties of the subject/discipline. Good research and shows some ability in the utilisation of research methodologies. Work demonstrates problem-solving skills and is completed with some level of accuracy, proficiency and autonomy. Satisfactory communication and expression demonstrated throughout. Student demonstrates some of the technical and/or artistic skills.
40-49% Pass		Pass - Assessment demonstrates some advanced knowledge and understanding of the subject informed by current practice, scholarship and research. Work may be incomplete with some irrelevant material present. Sometimes demonstrates the ability to analyse and apply concepts within the complexities and uncertainties of the subject/discipline. Acceptable research with evidence of basic ability in the utilisation of research methodologies. Demonstrates some originality, creativity and problem-solving skills but often with inconsistencies. Expression and presentation sufficient for accuracy and proficiency. Sufficient communication and expression with professional skill set. Student demonstrates some technical and/or artistic skills.
30-39% Fail	Fails to achieve learning outcomes	Fail - Very limited understanding of relevant theories, concepts and issues with deficiencies in rigour and analysis. Some relevant material may be present but be informed from very limited sources. Fundamental errors and some misunderstanding likely to be present. Demonstrates limited ability to analyse and apply concepts within the complexities and uncertainties of the subject/discipline. Limited research scope and ability in the utilisation of research methodologies. Limited originality, creativity, and struggles with problem-solving skills. Expression and presentation insufficient for accuracy and proficiency. Insufficient communication and expression and with deficiencies in professional skill set. Student demonstrates deficiencies in the range of technical and/or artistic skills.

20-29% Fail -		Fail - Clear failure demonstrating little understanding of relevant theories, concepts, issues and only a vague knowledge of the area. Little relevant material may be present and informed from very limited sources. Serious and fundamental errors and virtually no evidence of relevant research. Fundamental errors and misunderstandings likely to be present. Little or no research with no evidence of utilisation of research methodologies. No originality, creativity, and struggles with problem-solving skills. Expression and presentation insufficient for accuracy and proficiency. Insufficient communication and expression and with serious deficiencies in professional skill set. Student has clear deficiencies in range of technical and/or artistic skills.
0-19% Fail		Fail - Clear failure demonstrating no understanding of relevant theories, concepts, issues and no understanding of area. Little or no relevant material may be present and informed from minimal sources. No evidence of ability in the utilisation of research methodologies. No evidence of originality, creativity, and problem-solving skills. Expression and presentation deficient for accuracy and proficiency. Insufficient communication and expression and with deficiencies in professional skill set. Student has clear deficiencies in range of technical and/or artistic skills.