UiT

THE ARCTIC UNIVERSITY OF NORWAY

EXAMINATION PAPER

Portfolio Assignment 3

Home exam in:	**FYS-2021 - Machine Learning**
Hand-out:	**Friday November 13, 2020, 10:00**
Hand-in:	**Friday November 20, 2020, 13:00**

The exam contains 4 pages including this cover page

Contact person:	**Karl Øyvind Mikalsen**
Email:	**karl.o.mikalsen@uit.no**

UiT
THE ARCTIC
UNIVERSITY
OF NORWAY

Before You Start

Module examination

This is the third portfolio assignment for FYS-2021 Machine Learning. Portfolio assessment of project assignments counts about 40 % of the final grade in the course. All modules in the portfolio are assessed as a whole and one combined grade is given. Note that access to the final examination requires submission and approval of project assignments. Please familiarize yourself with the Regulations for examinations at UiT: https://en.uit.no/exams/art?p_document_id=523936

The report and code should be your own individual work. Remember to cite all sources. More information on source use, plagiarism, and cheating can be found here: https://en.uit.no/sensor/art?p_document_id=684332

Portfolio instructions

Read carefully: Failure to follow the instructions below may have a negative impact on the grade of your submission, or even cause your submission be deemed invalid.

Learning to write a scientific report is an important skill that many of the courses at the Faculty of Science and Technology, including this one, aim to improve. Therefore any *question* that you answer should be contained within the report of this portfolio assignment. Answers outside of the written report, (e.g. in the comment of the code, or within a Jupyter Notebook), will not be considered as a part of your answer of the problem. You can structure your report by having a separate (sub)section with the answer for each question. Remember to cite any sources you use - be advised that your submission will be checked for plagiarism.

Make sure your report shows that you understand what you are doing. More specifically, it is important to elaborate your answers such that essential theory, equations, and intuition is included in your answers. However, your answers should still remain concise and stay focused on the core problem, e.g. there is no need to derive or prove an equation unless the problem asks you to.

Problems that ask for numeric values or plots should include these in the answer of the report.

The code should be commented in such a way that any person with programming knowledge should be able to understand how the program works. Like your report, the code must be your own individual work.

You are permitted to use standard built-in functions and/or packages (e.g. *numpy* and *matplotlib* in

Python) for reading the data and basic calculations. However: make sure that the packages you use do not over simplify your implementation! Of course, all implementations asked for in the problems should be your own work.

Hand-in format

Your report must be a single file as in portable document format (`.pdf`). The file name *has to* follow the format `portfolio3_candidate`*`XX`*`.pdf` (replace *`XX`* with your candidate number obtained from StudWeb) for anonymity. Do not put your name in the report or code. Failure to do this may compromise your anonymity. Be advised that the name of the files are visible to the reviewers.

The code you write for this assignment should be included *both* in the appendix of the report *and* submitted as separate files in WISEflow.

Follow the hand-in instruction in WISEflow and **make sure to submit before the WISEflow room closes.**

Resources

All datasets required to answer the exercises can be found in the Canvas room for the course.

Problem 1

The file `city-inner-sweden.csv` provides the inner product matrix (the **B** matrix in the book). It is generated based on the geodesic distances between 34 Swedish cities, and the corresponding names are in the file `city-names-sweden.csv`. Based on these inner products, we will in the following exercises estimate the coordinates of these cities relative to each other.

(1a) Describe the main difference between feature extraction and feature selection. Describe the multidimensional scaling (MDS) algorithm and comment on its areas of use.

(1b) Implement the multidimensional scaling algorithm, and perform MDS on the provided data. Briefly discuss what would be a sensible number of dimensions of your output.

(1c) Plot the result of your MDS scaling. Comment on your result by visually comparing it to a map (e.g. Google maps[1]). Try to explain similarities and differences.

Problem 2

The file `frey-faces.csv` contains 1965 images of faces with size 20×28 (20 horizontal pixels and 28 vertical pixels). *Be advised that this file does not contain any labels!*

(2a) Describe the k-means algorithm. Discuss how the algorithm can be initialized. Are there any problems that might occur in connection with the initialization of the algorithm?

(2b) Implement the k-means algorithm. Run your k-means algorithm with $k = 2$, $k = 4$ and $k = 10$ clusters on the `frey-faces.csv` dataset. Plot the centroids and some of the images closest to each cluster.

(2c) Use the results from your k-means algorithm for 2, 4 and 10 clusters you obtained in (b) (do *not* re-run your algorithm for this part of the problem since the results might change). Plot the border cases: the faces that are on the borders between clusters. Use these images, and the images closest to each cluster centre plotted in (b), to compare the clusters and what they represent.

[1] https://goo.gl/maps/hCoNRL1WQU82