Machine Learning, Spring 2018: Exercise Sheet 4

This is a programming exercise. It will be graded and the grade counts toward the course grade. ***Join into groups of two*** *and submit a single solution per group, indicating the group members' names on the report sheet. You can use Python or Matlab.*

Please send your type-set solutions by email to our two TA's Tianlin Liu (t.liu@jacobs-university.de) and Tayyab Mateen (t.mateen@jacobs-university.de)

Deadline for submission is Wednesday March 14, 23:59 hrs (email sending timestamp). Submissions arriving later (even a second after midnight) will be corrected but not counted for the course grade.

Task description. *Throughout this course we will be basing programming exercises on the digits dataset described in Section 4 of the lecture notes. Today's problem is the first in this series. You will later be able to re-use much code from this problem, especially code that generates graphical output*.

Here

http://minds.jacobs-university.de/sites/default/files/uploads/teaching/share/DigitsBasicRoutines.zip

you can download the digits dataset together with some elementary Matlab routines for visualization (if you use Python you'll have to translate them to Python) and some super-elementary scripts for training classifiers.

Your task: pick one of the digits (e.g. the "ones"), which gives you a dataset of 200 image vectors. Carry out a K-means clustering on your chosen sample, setting $K = 1$ (!), 2, 3, and 200 in four runs of this algorithm. Generate visualizations of the images that are coded in the respective codebook vectors that you get (for the $K = 200$ case, only visualize a few). Discuss what you see. Your discussion should include (but not be restricted to) answers to the questions (1) what is the mathematical nature of the codebook image for the case $K = 1$? (2) what is the mathematical nature of the codebook images for the case $K = 200$?

There are innumerable Matlab and Python implementations of K-means clustering on the web. Stay away from them and program your K-means clustering algorithm from scratch. It's not a big deal – only a few lines of code; doing that by yourself means you learn something useful for life... because K-means clustering is indeed very useful.

Deliverables: a typeset discussion (say, half a page of text, but can be more) with nice graphics, and in a separate text file the code that you produced. The codefile must also include your names and it must be minimally documented inline such that the TA's can quickly grasp what you are computing where in your code. The TAs will do no code-checking or code reviewing, but they may want to inspect your code to resolve possible vaguenesses in your report. The grade will be primarily based on the report

document, but poorly documented code that the TAs cannot easily understand will lead to a grade reduction.