Machine Learning, Spring 2017: Exercise Sheet 2

Please send your type-set solutions by email to our two TA's Xu He ("Owen") x.he@jacobs-university.de and Felix Schmoll f.schmoll@jacobs-university.de. Join into groups of two or three and submit a single solution sheet per group, indicating the group members' names on the sheet.

Deadline for submission is Friday Feb 24, 23:59 hrs (email sending timestamp). Submissions arriving later (even a second after midnight) will be corrected but not counted for the course grade.

Problem 1 (easy, informal). Make a list of 5 classification tasks of real-world relevance and present them in a format similar to the table at the beginning of Section 4 of the lecture notes. The purpose of this task is to make you aware of the (almost) universality of the notion of "classification" – once you start thinking of examples you'll find that many relevant real-life problems can be cast as picking the right labels for patterns.

Problem 2 (programming). *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.* – At

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 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 them by yourself means you learn something useful for life... because K-means clustering is indeed useful.

Deliverables: a typeset discussion (say, half a page of text, but can be more) with nice graphics, and the code that you produced. The code must be minimally documented inline such that the TA's can quickly grasp what you are computing where in your code.