Machine Learning

Spring Semester 2021
Prof. Dr. Peter Zaspel
Abhieshree Dhami, Kristijan Spirkoski, Kristian
Sterjo

Assignment Sheet 10.

Submit on **Tuesday, April 20, 2021, 10:00**.

Excercise 1. (K-means clustering)

In this task, you carry out K-means clustering with $K = 3$ by paper and pencil. To this end, you are given the following data

i	x_i	$C^{(0)}(i)$
1	$(15,7)^\top$	2
2	$(0,-9)^\top$	1
3	$(-2,-5)^\top$	1
4	$(2,3)^\top$	0
5	$(3,7)^\top$	2
6	$(18,12)^\top$	0

with its initial cluster assignment. Let the clustering algorithm "run" until it finalized its assignment. Finally, draw the just computed clusters in a two-dimensional scatter plot.

(4 Points)

Excercise 2. (Principle component analysis)

You are given the data set

$$\{(1,0)^\top, (0.5,1)^\top, (1,0.5)^\top\}\,.$$

Manually compute the principle components of this data set by solving the associated eigenvalue problem.

(4 Points)

Excercise 3. (Principle component analysis for data compression)

A well-known application of principle component analysis is lossy data compression. In this application, you are given a large data set $\{\mathbf{x}\}_{i=1}^N$ with $\mathbf{x}_i \in \mathbb{R}^D$ and reduce it to a data set $\{\tilde{\mathbf{x}}\}_{i=1}^N$ with $\tilde{\mathbf{x}}_i \in \mathbb{R}^d$ where $d < D$, while storing a matrix that allows to reconstruct an approximation to the $\mathbf{x}_i$ from the vectors $\tilde{\mathbf{x}}_i$.

Develop and give a compression and a decompression algorithm which carry out the above described data compression and decompression tasks by using principle component analysis. You can either try to develop the idea by yourself or do a research in the internet. In the latter case, please quote the source from which you took the information.

(4 Points)

Programming Exercise 1. In this task, you are supposed to apply your just developed data compression algorithm to the MNIST data set. To this end, consult again Example 9.4, which provides access to the hand-written digits for the value 8 in MNIST. Compress the images of these digits to a dimensionality of $d = 784, 512, 256, 128, 64, 32$, decompress them again, and plot the decompressed digits.

Reference solutions will only be provided in Python+Matplotlib. The submission format for Python is a Jupyter notebook. The submission format for C/C++ is standard source files. Choose an appropriate format for the Gnuplot-related submission.

(4 Points)