University of Fribourg
Departement of Informatics

Lecturer: Dr. Andreas Fischer
Teaching Assistants: Linda Studer

Spring Semester

Pattern Recognition

Exercise 1b - March 6, 2023

First Clustering Task

Deadline: March 20, 2023 (end of day)

The aim of this exercise is to implement your first clustering task based on the MNIST dataset,[1] the same dataset you have already used for the previous classification task with KNN.

In this exercise implement the **K-Means** algorithm, apply it to the dataset (train.csv), and validate the clustering. The definition of the algorithms as well as the validation methods can be found in the slides.

Data and Features

You can download the training dataset from ILIAS containing 26,999 entries. The csv contains the label (the number written in the respective image) and the pixel values for a 28×28 grey scale image in the range from 0 to 255. The image is represented as a one-dimensional array so you need to reshape the data if you want to display the image to the screen. As features use the pixel values; there is no need to derive more sophisticated features from the data. Use the standard Euclidean distance to compute the dissimilarity of two images.

K-Means

Create your own implementation of the K-Means algorithm (see lecture notes). Use it to cluster the contents of the training set.

Apply *at least two* of the following cluster validation methods: C-Index, Goodman-Kruskal-Index, Dunn-Index, or Davis-Bouldin-Index.

Expected Output

- Source Code of your implementation
- Validation Values for $K = \{5, 7, 9, 10, 12, 15\}$
- Short report / text file with results

[1] http://yann.lecun.com/exdb/mnist/