Machine Learning - Spring 2019 **March 17, 2019**

Mini Project #1

Instructor: Prof. Dr. Herbert Jaeger *Authors:* Ralph Florent, Diogo Cosin

Task Description: The project is a programming exercise that requires to code the K-means clustering Machine Learning algorithm. Once implemented, the coded algorithm should be able to regroup into K(1, 2, 3, 200) clusters of digit images from the "Optical Character Recognition (OCR)" dataset provided by Prof. Dr. H. Jaeger. Additionally, a few examples of visualizations of the obtained results should be provided.

Summary: Attached along this report, should it be found a Jupyter Notebook which contains the Python script that implements the K-means clustering algorithm from the Dr. Herbert Jaeger's Machine Learning Lecture Notes. The algorithm is described in the following steps:

- **Given**: a training data set (xi)i=1,...,N Rn, and a number K of clusters that one maximally wishes to obtain
- **Initialization**: randomly assign the training points to K sets Sj (j = 1, . . . , K).
- **Repeat**: For each set Sj, compute the mean j = Sum(x)/—Sj— for x Sj. This mean vector j is the center of gravity of the vector cluster Sj. Create new sets S'j by putting each data point xi into that set S'j where Modulus(xi j) is minimal. If some S'j remains empty, dismiss it and reduce K to K' by subtractring the number of dismissed empty sets (this happens rarely). Put Sj = S'j (for the nonempty sets) and K = K'.
- **Termination**: Stop when in one iteration the sets remain unchanged.

Analysis