CBS COPENHAGEN BUSINESS SCHOOL
HANDELSHØJSKOLEN

# Data Mining, Machine Learning, and Deep Learning

## Compulsory assignment 1

## Question 3 : Written Assignment

**MSc in Business Administration and Data Science**

| Student ID | Student name |
|---|---|
| 143003 | Diana Laura Janikowski |
| 158283 | Georgios Kotrotsios |
| 158798 | Andrea Pérez López |
| 158398 | Celine Schuhmann |

Submission date:

**02/03/2023**

**Characters (incl. spaces): 4,621**

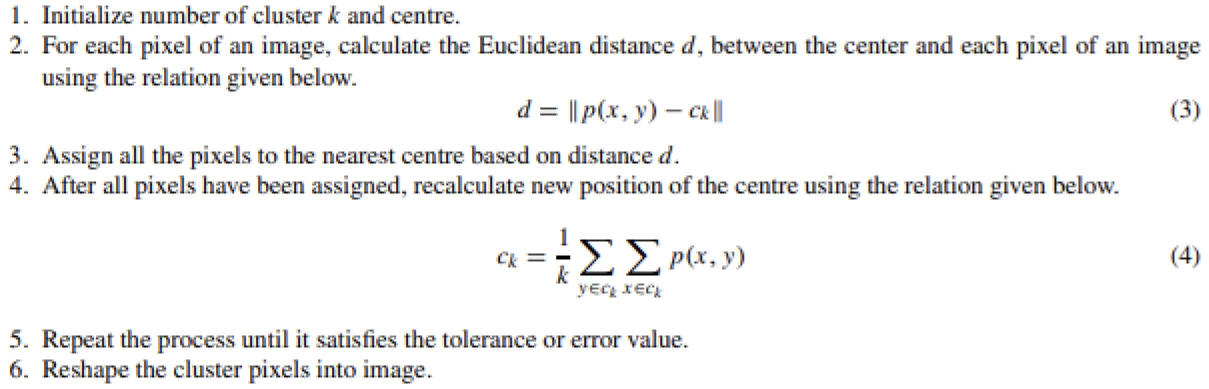
**Pages: 3**

## Clustering

Machine learning systems can be distinguished based on the supervision used in order to train them which can take the following forms: supervised, unsupervised, self-supervised, semi-supervised as well as reinforcement learning (Géron, 2023). The unsupervised learning method is utilized when the machine learning system is programmed to learn by itself. In such a case, clustering, which is the focus of this essay, is one of the tasks that can be used. According to Géron (2023), clustering is "[...] the task of identifying similar instances and assigning them to clusters, or groups of similar instances". Grouping blog visitors based on their similarity is an instance of clustering. Other areas where clustering is used include statistics, bioinformatics, image processing, etc. (Ahmed et al., 2020). In general, the definition of a cluster depends on the context. As mentioned by Géron (2023), it can be "[...] instances centered around a particular point, called a centroid [or] regions of densely packed instances [...]".

Two examples of popular clustering algorithms include k-means and DBSCAN (Géron, 2023). The first one, k-means which is also sometimes called the Lloyd–Forgy algorithm, was introduced in 1957. Disciplines such as image segmentation and handwriting recognition utilize this algorithm accompanied by deep learning (Ahmed et al., 2020). The first mentioned discipline, image segmentation, considers pixels of an image (Dhanachandra et al., 2015). Here, an image is dissected into partitions based on the pixels' similarity. Moreover, these partitions are characterized by a high difference in contrast between each other.

The following figure is an example of a k-means clustering algorithm used for image segmentation:



1. Initialize number of cluster $k$ and centre.
2. For each pixel of an image, calculate the Euclidean distance $d$, between the center and each pixel of an image using the relation given below.

$$d = \| p(x, y) - c_k \| \tag{3}$$

3. Assign all the pixels to the nearest centre based on distance $d$.
4. After all pixels have been assigned, recalculate new position of the centre using the relation given below.

$$c_k = \frac{1}{k} \sum_{y \in c_k} \sum_{x \in c_k} p(x, y) \tag{4}$$

5. Repeat the process until it satisfies the tolerance or error value.
6. Reshape the cluster pixels into image.

Figure 1. k-means clustering (Dhanachandra et al., 2015)

According to Ahmed et al., (2020), k-means clustering is characterized by two unavoidable problems related to centroid assignments as well as the assignment of the cluster number, and moreover, the algorithm's ability to work with mixed data types and features. When training a k-means cluster on a dataset the k, which is the number of clusters that the algorithm is instructed to find, must be provided (Géron, 2023). Because of that, different results can be expected (Dhanachandra et al., 2015). Therefore, the initialization of a proper number is an important task and most likely problematic as well. Moreover, the result of clustering is dependent on the initial centroid assignment, which must be also carefully selected in order for the algorithm to arrive at the desired segmentation. Additionally, it can also be argued whether k-means clustering truly belongs to the unsupervised machine learning field since the algorithm needs to be instructed in terms of centroid assignment and how many clusters it is supposed to organize the results which heavily affects the clustering results.

DBSCAN (density-based spatial clustering of applications with noise) is another popular clustering algorithm mentioned earlier. This algorithm does not require specifying the number of clusters, instead, it is necessary to initialize other variables, such as Epsilon values and the minimum number of points (Kurumalla and Rao, 2016). When using the DBSCAN algorithm, a cluster is described using wording such as core, border, or noise points:

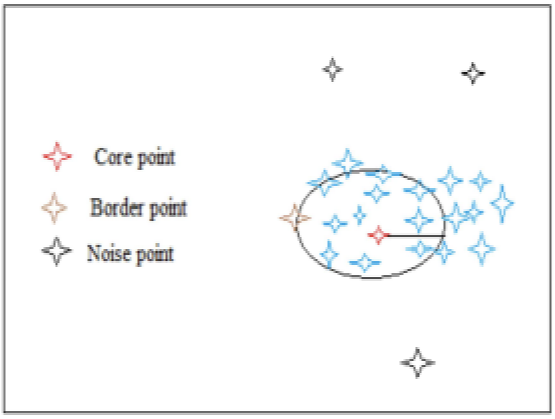


Figure 2. DBSCAN clustering (Kurumalla and Rao, 2016)

As described by Kurumalla and Rao (2016), DBSCAN clustering can be summed up in 5 points as follows: "1. Choose a random point. 2. Explore whole points that are density reachable from p rendering to Epsilon and Minimum points. 3. If p is a core point, formerly a cluster is generated.

4. If p is a border point, no point is accessible by density and DBSCAN moves to the succeeding point of the database. 5. Repeat the procedure until whole points are visited”.

The DBSCAN algorithm can be used for image segmentation as well, and the process to arrive at a segmented image as proposed by Kurumalla and Rao (2016) starts with the conversion of the RGB image to a gray one, followed by noise removal from the gray image. Afterward, the epsilon and minimum number of points values are selected and the traditional DBSCAN algorithm is used. Compared to k-means clustering, which struggles with clusters that vary in sizes and densities, as well as with clusters whose shapes are nonspherical, DBSCAN behaves well in regard to the identification of clusters regardless of their shapes (Géron, 2023). However, similar to k-means, DBSCAN clustering will struggle when the difference between clusters’ density is significant.

## Bibliography

Ahmed, M., Seraj, R. and Islam, S.M. (2020) "The K-Means Algorithm: A comprehensive survey and performance evaluation," *Electronics*, 9(8), p. 1295. Available at: https://doi.org/10.3390/electronics9081295.

Dhanachandra, N., Manglem, K. and Chanu, Y.J. (2015) "Image segmentation using K - means clustering algorithm and subtractive clustering algorithm," *Procedia Computer Science*, 54, pp. 764–771. Available at: https://doi.org/10.1016/j.procs.2015.06.090.

Kurumalla, S. and Rao, P.S. (2016) "-nearest neighbor based dbscan clustering algorithm for image segmentation," *Jatit*, 92(2), pp. 395–402.