Exercise 6

Clustering Algorithms

Distance measurements

Euclidean distance	d(A , B) = $\sqrt{\sum_{i=1}^{n} (A_i - B_i)^2}$
Manhattan distance	d(A , B) = $\sum_{i=1}^{n} \lvert A_i - B_i \rvert$
Hamming distance	d(A , B) = # { i : $A_i \neq B_i$ }

k-means Algorithm:

```
Inputs: D: Dataset; k: the number of clusters to form;
Outputs: D: Labeled dataset;
Begin
    Randomly choose k instances as centroids.
    Repeat
        Calculate the distance between each instance D[i] and the k centroids;
        Assign each instance D[i] to the cluster closest to its center;
        Calculate the new center of each cluster and modify the centroid;
    Until D[i] is the same as D[i-1];
Return D;
End.
```

Questions :

Let's ignore the "*Exercise*" and “ep (ms)” and “ID” attributes.

1- Re-use the distance function seen in TP 5.

2- Write a function to calculate the centroid of a set of instances.

3- Write a function to find the cluster to which a given instance is the closest.

4- Implement the k-means algorithm.

5- Test the algorithm with k=2, k=5 and k=6. Suggest an interpretation of the obtained clustering.

Have fun !