Machine Learning Course
Second intermediate assessment – June 4, 2019

Students should do all the exercises to get the maximum score.
If you solve all the three exercises correctly, you get 33 points.
Please, justify carefully each answer.

Name: ……. Surname: ……. Student ID: ……

Exercise on the k-means clustering algorithm

Given the n=6 two-dimensional data points **x**, and their labels **y**: $\mathbf{x} = \begin{bmatrix} -2 & 0 \\ 0 & -2 \\ -2 & -1 \\ 1 & 1 \\ 1 & 0 \\ 2 & 1 \end{bmatrix}$, $\mathbf{y} = \begin{bmatrix} -1 \\ -1 \\ -1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$,

and the initial k=2 centroids $\mathbf{v} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$:

- Cluster the data points x using the **k-means clustering algorithm**, reporting the clustering labels, the updated centroids and the **objective function** at each iteration of the algorithm.
- You should use this **objective function:** $\sum_{i=1}^{n} \left|\left|\boldsymbol{x}_i - \boldsymbol{v}_k\right|\right|_1$
- For simplicity, use the L1 (Manhattan) distance instead of the L2 (Euclidean) distance, both for computing the objective function $\sum_{i=1}^{n} \left|\left|\boldsymbol{x}_i - \boldsymbol{v}_k\right|\right|_1$ (being $\boldsymbol{v}_k$ the closest centroid to $\boldsymbol{x}_i$) and for computing the distances between the data points x and the centroids v. If a point has the same distance with respect to a number of centroids, assign it to the centroid with the lowest class index in this set (e.g., if the point has the same distance w.r.t. centroid 0 and 1, assign it to centroid 0).
- Make a two-dimensional plot displaying the data points (with a clear indication to explain to which cluster each point belongs to, after the last iteration) and the final centroids.

SOLUTION

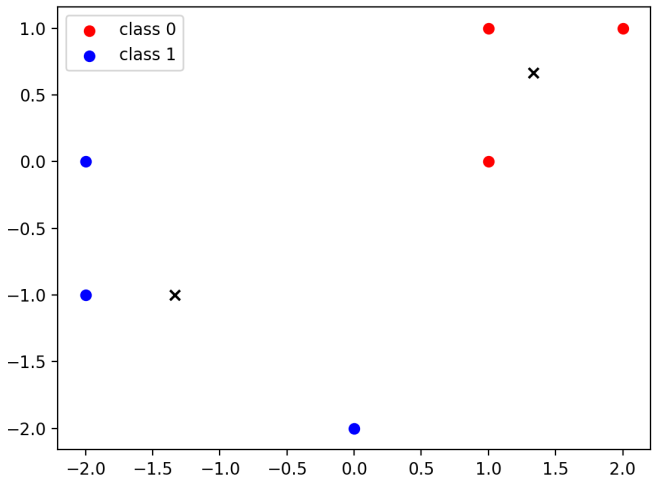
$$\mathbf{x} = \begin{bmatrix} -2 & 0 \\ 0 & -2 \\ -2 & -1 \\ 1 & 1 \\ 1 & 0 \\ 2 & 1 \end{bmatrix} \qquad \mathbf{v} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$$

$d(x_1, v_2) = |-2 - 1| + |0 - 1| = 4$

$d(x_1, v_2) = |-2 - 1| + |0 - (-1)| = 4$

iter	$\sum_{i=1}^{n} \|\|\boldsymbol{x}_i - \boldsymbol{v}_k\|\|_1$	*distance matrix*	*cluster assignments*	*current* $\boldsymbol{v}$
0	11.0	[[**4**. 4.] [4. **2**.] [5. **3**.] [**0**. 2.] [**1**. 1.] [**1**. 3.]]	0 1 1 0 0 0	[[1. 1.] [1. -1.]]
1	9.5	[[3. **2.5**] [3. **1.5**] [4. **1.5**] [**1.** 4.5] [**1.** 3.5] [**2.** 5.5]]	0 0 0 1 1 1	[[0.5 0.5] [-1. -1.5]]
2	7.33	[[4. **1.67**] [4. **2.33**] [5. **0.67**] [**0.67** 4.33] [**1.** 3.33] [**1.** 5.33]]	0 0 0 1 1 1	[[1.33 0.67] [-1.33 -1.]]

After iteration 2, the cluster assignments do not change anymore. Therefore, the algorithm stops.
The final clustering (along with the centroids) is shown below.


class 0
class 1
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0