CS 383 - Machine Learning

Assignment 4 - Clustering

1 Theory

1. Given two clusters:

$$C_1 = \{(1, 2), (0, -1)\}, C_2 = \{(0, 0), (1, 1)\}$$

what is:

(a) The weighted average intra-cluster distance if you are using euclidean distance?

$$G_i = \frac{\sum_{x,y \in C_i} d(x, y)}{2|C_i|}$$

$$G_1 = \frac{\sum_{x,y \in C_1} d(x, y)}{2|C_1|} = \frac{\sqrt{(1-0)^2 + (2--1)^2}}{2*2} = \frac{\sqrt{10}}{4}$$

$$G_2 = \frac{\sum_{x,y \in C_2} d(x, y)}{2|C_2|} = \frac{\sqrt{(1-0)^2 + (1-0)^2}}{2*2} = \frac{\sqrt{2}}{4}$$

$$W_j = \sum_{i=1} \frac{|C_i|}{N} G_i$$

$$W_j = \frac{|C_1|}{2} G_1 + \frac{|C_2|}{2} G_2 = \frac{2}{2}\frac{\sqrt{10}}{4} + \frac{2}{2}\frac{\sqrt{2}}{4} = \frac{\sqrt{10}+\sqrt{2}}{4}$$

(b) The single link similarity between the clusters if we're using cosine similarity as our similarity function?

$$sim(C_i, C_j) = min_{x \in C_i, y \in C_j}(sim(x, y))$$

$$sim(A, B) => cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|}$$

$$sim((1, 2), (0, 0)) = \frac{(1, 2) \cdot (0, 0)}{\|(1, 2)\|\|(0, 0)\|} \rightarrow \text{undefined}$$

$$sim((1, 2), (1, 1)) = \frac{(1, 2) \cdot (1, 1)}{\|(1, 2)\|\|(1, 1)\|} = \frac{3\sqrt{10}}{10}$$

$$sim((0, -1), (0, 0)) = \frac{(0, -1) \cdot (0, 0)}{\|(0, -1)\|\|(0, 0)\|} \rightarrow \text{undefined}$$

$$sim((0,-1),(1,1)) = \frac{(0,-1)\cdot(1,1)}{\|(0,-1)\|\|(1,1)\|} = \frac{-\sqrt{2}}{2}$$

$$\text{Single link similarity} = \frac{-\sqrt{2}}{2}$$

(c) The complete link similarity between the clusters if we're using cosine similarity as our similarity function?

$$sim(C_i, C_j) = max_{x\in C_i, y\in C_j}(sim(x,y))$$

$$sim(A,B) => cos(\theta) = \frac{A\cdot B}{\|A\|\|B\|}$$

$$\text{Complete link similarity} = \frac{3\sqrt{10}}{10}$$

(d) The average link similarity between the clusters if we're using cosine similarity as our similarity function?

$$sim(C_i, C_j) = \frac{1}{|C_i||C_j|}\sum_{x\in C_i}\sum_{y\in C_j}(sim(x,y))$$

$$sim(A,B) => cos(\theta) = \frac{A\cdot B}{\|A\|\|B\|}$$

$$\text{Average link similarity} = \frac{1}{4}\cdot(\frac{3\sqrt{10}}{10} + \frac{-\sqrt{2}}{2}) = 0.060394$$

2. Given an average intracluster distance for clustering level j, W_j , what is the fourth derivative at j, namely $W_j^{(4)}$?

$$W_j = \sum_{i=1}\frac{|C_i|}{N}G_i$$

$$W_j' = \frac{1}{2}\cdot(W_{j+1} - W_{j-1})$$

$$W_j'' = \frac{1}{2}\cdot(W_{j+1}' - W_{j-1}') = \frac{1}{2}\cdot(\frac{1}{2}\cdot(W_{j+2} - W_j) - \frac{1}{2}(W_j - W_{j-2})) = \frac{1}{4}\cdot(W_{j+2} - 2W_j + W_{j-2})$$

$$W_j''' = \frac{1}{4}\cdot(W_{j+2}' - 2W_j' + W_{j-2}') = \frac{1}{4}\cdot(\frac{1}{2}\cdot((W_{j+3} - W_{j+1}) - 2(W_{j+1} - W_{j-1}) + (W_{j-1} - W_{j-3}))$$

$$= \frac{1}{8}\cdot(W_{j+3} - 3W_{j+1} + 3W_{j-1} - W_{j-3})$$

$$W_j^{(4)} = \frac{1}{8}\cdot(W_{j+3}' - 3W_{j+1}' + 3W_{j-1}' - W_{j-3}') =$$

$$\frac{1}{8}\cdot\frac{1}{2}\cdot((W_{j+4} - W_{j+2}) - 3(W_{j+2} - W_j) + 3(W_j - W_{j-2}) - (W_{j-2} - W_{j-4}))$$

$$\frac{1}{16} \cdot (W_{j+4} - 4W_{j+2} + 6W_j - 4W_{j-2} + W_{j-4})$$

3. Given the output of your clustering algorithm as $C_1 = \{1, 2, 3, 4\}, C_2 = \{5, 6, 7, 8\}$, and a hand labeled clustering of $C_1 = \{3, 4\}, C_2 = \{1, 2, 5, 6, 7, 8\}$, what is the weighed average purity of the clusters created by the clustering algorithm?

   $$Purity(C_i) = \frac{1}{|C_i|} max_j N_{ij}$$

   $$Purity(C_1) = \frac{1}{|C_1|} max_j N_{1j} = \frac{1}{4} \cdot max(2, 2) = \frac{1}{2}$$

   $$Purity(C_2) = \frac{1}{|C_2|} max_j N_{2j} = \frac{1}{4} \cdot max(2, 4) = 1$$

   $avg\ purity = \frac{1}{N} \sum_{i=1}^{k} |C_i| Purity(C_i)$ $avg\ purity = \frac{1}{8} \cdot (4 \cdot \frac{1}{2} + 4 \cdot 1) = 0.75$

2 Clustering

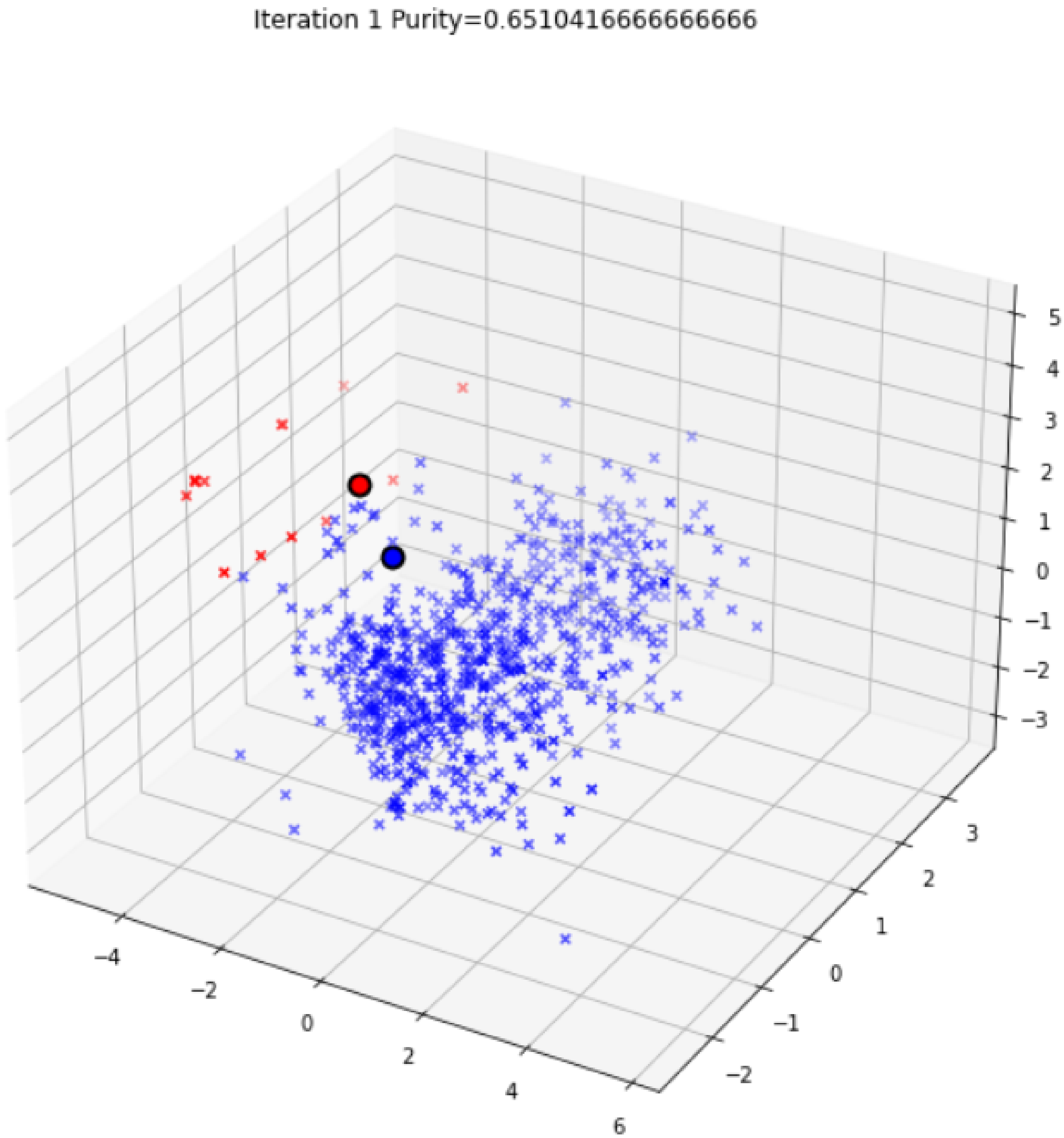

Figure 1: Initial Clustering

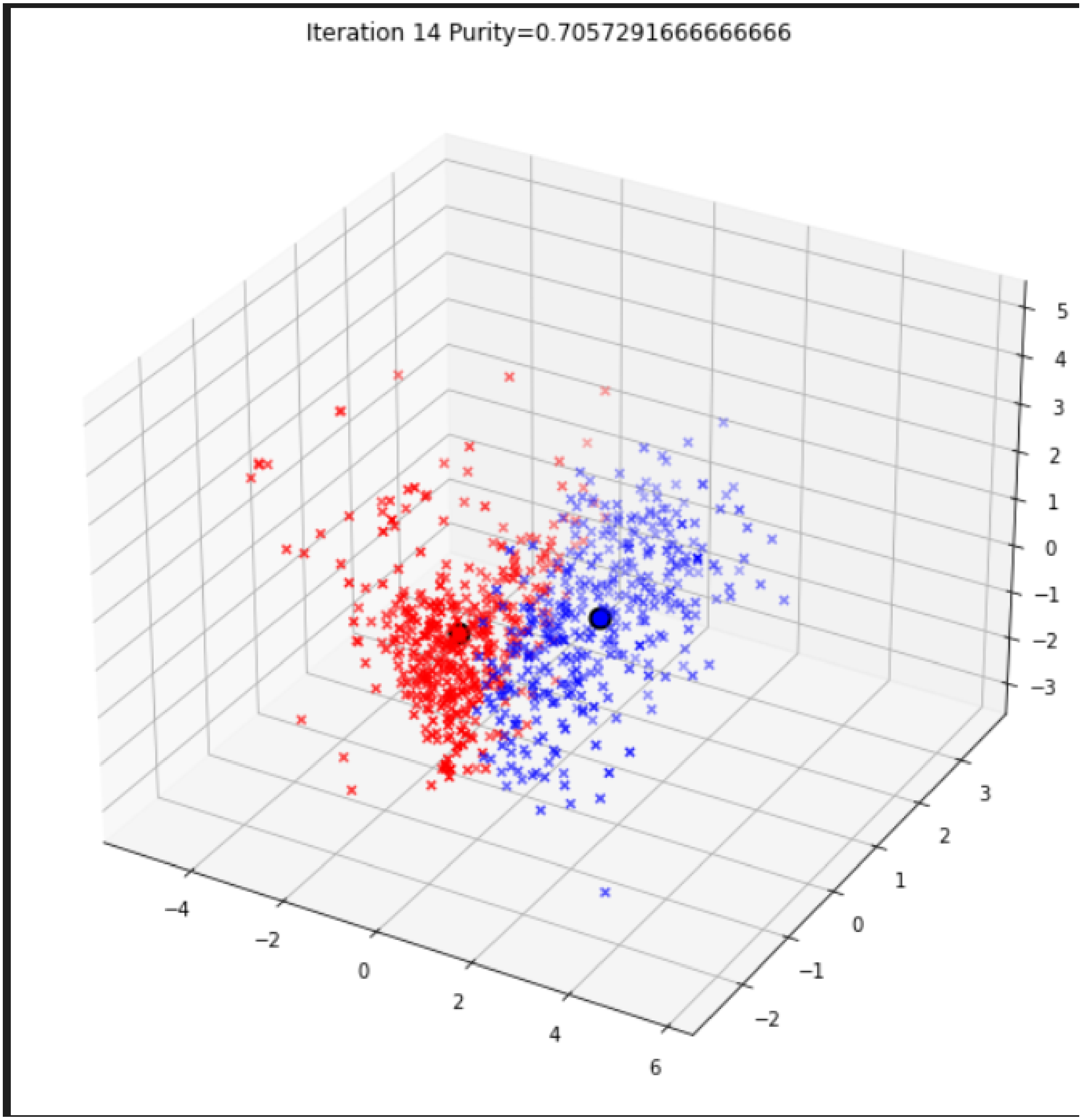

Figure 2: Final Clustering