Clustering

K-Means

Degeneracy

Example

Addressing

Computation Results

Conclusion

References

Degeneracy on K-means clustering

Abdulrahman Alguwaizani
Assistant Professor

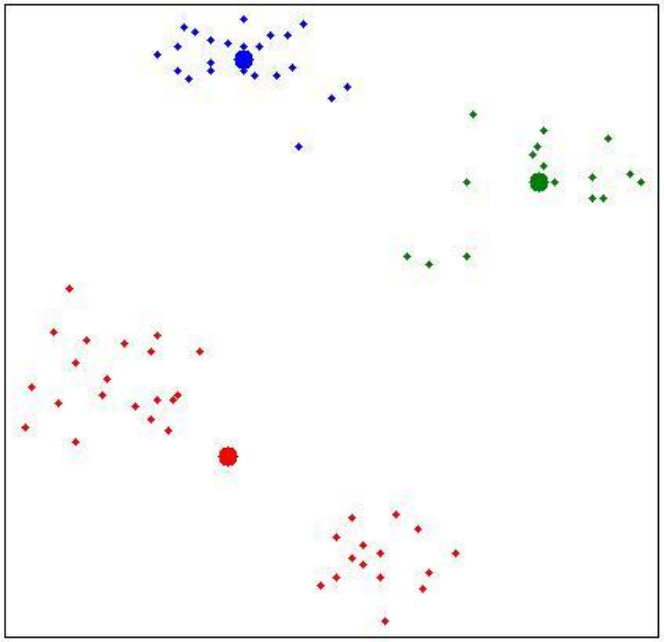
Math Department
College of Science
King Faisal University
Kingdom of Saudi Arabia

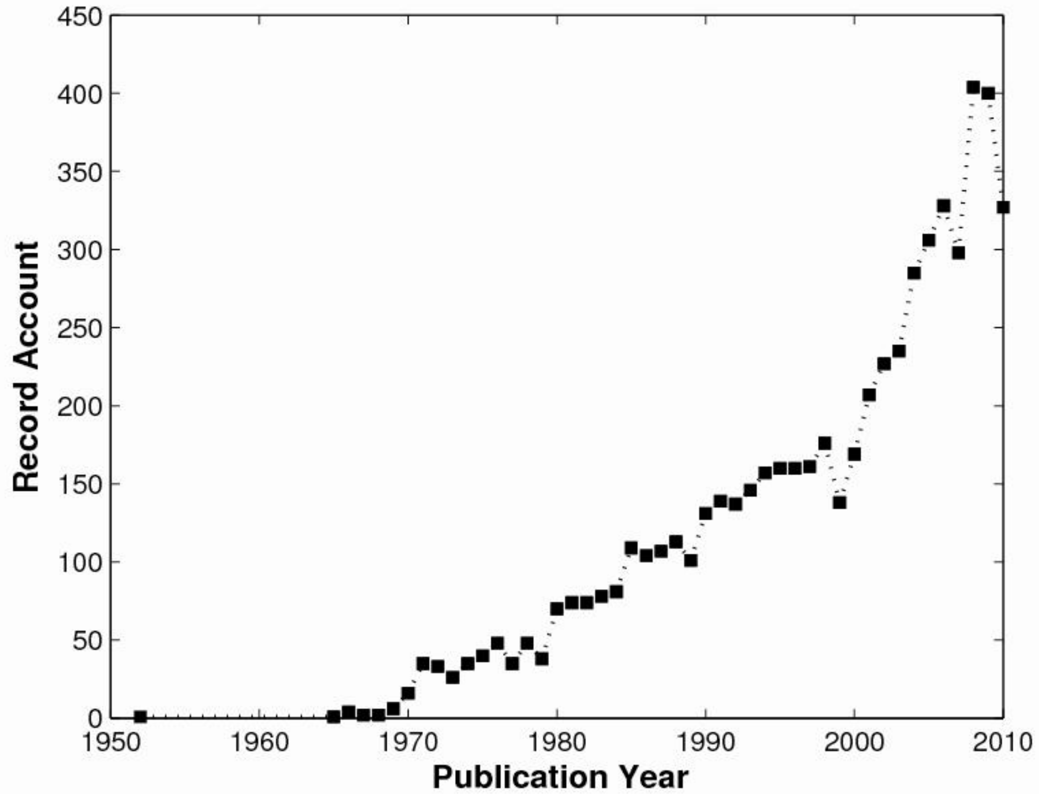
Office # 0096635899717
Mobile # 00966505920703
E-mail: aalguwaizani@kfu.edu.sa

Clustering
K-Means
Degeneracy
Example
Addressing
Computation Results
Conclusion
References

DEFINITION:

Clustering is a scientific method which addresses the following very general problem: given the data on a set of entities, find clusters, or groups of these entities, which are both homogeneous and well-separated. **Homogeneity** means that the entities in the same cluster should resemble one another. **Separation** means that entities in different clusters should differ from one another.





The growth of publications on clustering.

CLUSTERING METHODS:

1. **HIERARCHICAL CLUSTERING:**

- **Agglomerative Hierarchical Clustering.**
- **Divisive Hierarchical Clustering.**

Let δ(C1,C2) be the distance function between two clusters C1 and C2. It can be computed as:

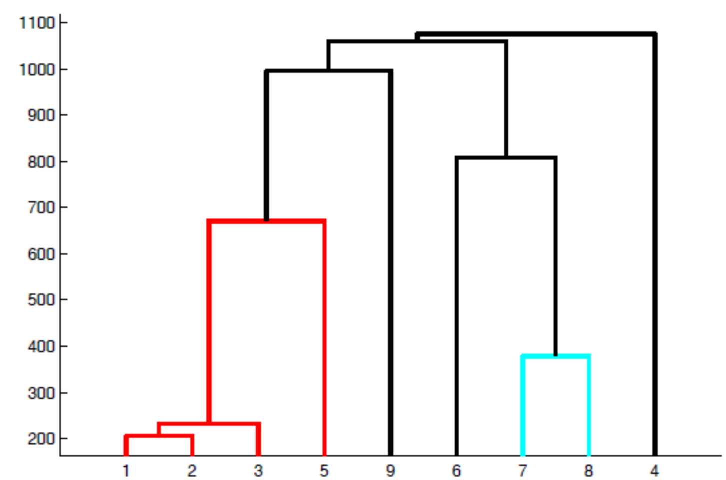
- $\delta(C_1, C_2) = \min \{ d(i,j) : i \in C_1 \quad , \quad j \in C_2 \}$. For single linkage.
- $\delta(C_1, C_2) = \max \{ d(i,j) : i \in C_1 \quad , \quad j \in C_2 \}$. For complete linkage.
- $\delta(C_1, C_2) = \frac{1}{|C_1| \cdot |C_2|} \sum_{i \in C_1} \sum_{j \in C_2} d(i,j)$. For average linkage.

- Clustering
- K-Means
- Degeneracy
- Example
- Addressing
- Computation Results
- Conclusion
- References

Example:
Consider the Table which shows the distances in miles between some United States cities . The method of clustering is single linkage. So, in the first stage BOS and NY are merged into a new cluster because 206 is the minimum distance. After applying the agglomerative algorithm, the rest of the solution can easily be concluded from the dendrogram in Figure

	1 BOS	2 NY	3 DC	4 MIA	5 CHI	6 SEA	7 SF	8 LA	9 DEN
BOS	0	206	429	1504	963	2976	3095	2979	1949
NY	206	0	233	1308	802	2815	2934	2786	1771
DC	429	233	0	1075	671	2684	2799	2631	1616
MIA	1504	1308	1075	0	1329	3273	3053	2687	2037
CHI	963	802	671	1329	0	2013	2142	2054	996
SEA	2976	2815	2684	3273	2013	0	808	1131	1307
SF	3095	2934	2799	3053	2142	808	0	379	1235
LA	2979	2786	2631	2687	2054	1131	379	0	1059
DEN	1949	1771	1616	2037	996	1307	1235	1059	0

Table 1.1: *Distances in miles between U.S. cities*



Clustering
K-Means
Degeneracy
Example
Addressing
Computation Results
Conclusion
References

2. PARTITIONING:

let $X = \{x_1, \ldots, x_N\}$ be a set of objects or entities to be clustered ($x_i \in \mathbb{R}^q$), and let C be a subset of X. Then $P_K = \{C_1, C_2, \ldots, C_K\}$ is a partition of X into K clusters if it satisfies:

(i) $C_k \neq \emptyset; \quad k = 1, 2, \ldots, K.$

(ii) $C_i \cap C_j = \emptyset; \quad i, j = 1, 2, \ldots, K; \quad i \neq j.$

(iii) $\bigcup_{k=1}^{K} C_k = X.$

Clustering

K-Means

Degeneracy

Example

Addressing

Computation Results

Conclusion

References

Minimun Sum of Squares Clustering (MSSC):

consider a set $X = \{x_1, \ldots, x_i, \ldots, x_N\}, x_i = \{x_{1i}, \ldots, x_{qi}\}$ of N entities in Euclidean space $\mathbb{R}^q$. The MSSC problem is to find a partition of X into K disjoint subsets C_j such that the sum of squared distances from each entity x_i to the centroid c_j of its cluster C_j is the minimum.

Specifically, let P_K denote the set of all partitions of X into K sets. Let partition P be defined as $P = \{C1, C2, \ldots, CK\}$.

Then MSSC can be expressed as:

$$f_{MSSC}(P) = \min_{P \in P_K} \sum_{i=1}^{N} \min_{j=1,\ldots,K} \|x_i - c_j\|^2,$$

where the centroid of cluster *j* is given as:

$$c_j = \frac{1}{|C_j|} \sum_{i \in C_j} x_i.$$

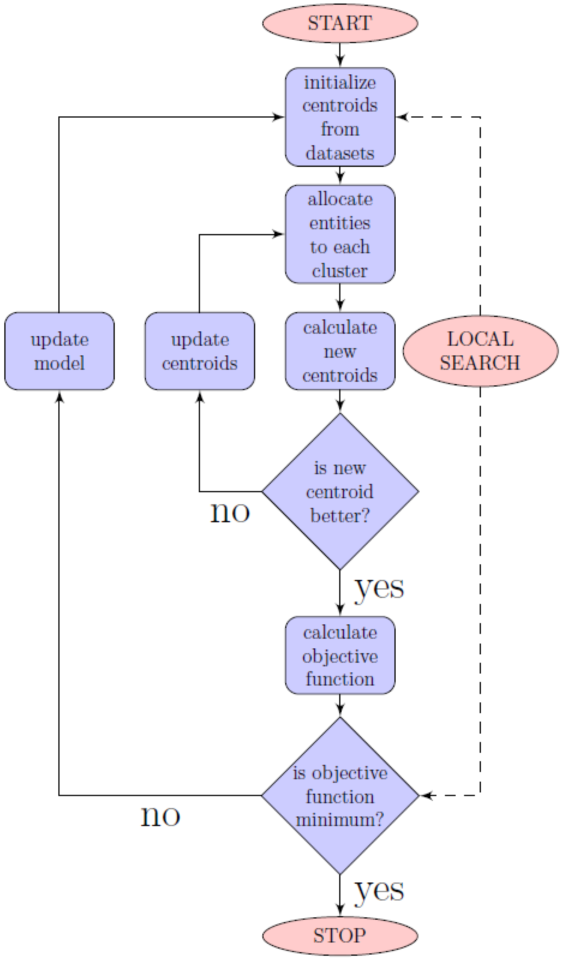


Fig. 1: The K-means clustering algorithm.

Clustering
K-Means
Degeneracy
Example
Addressing
Computation Results
Conclusion
References

K-Means Algorithm:

Algorithm 1. *K-Means algorithm* (KM) *for the MSSC problem*

Function KM $(X, K, Maxit, N, C, z)$

Choose initial centroids c_k $(k = 1, \ldots, K)$
$l \leftarrow 0$
Repeat
    $l \leftarrow l + 1$
    For $i := 1, \ldots, N$ **Do**
        $m(x_i) \leftarrow \text{argmin}_{j \in \{1,\ldots,K\}} (\|x_i - c_j\|_2)^2$
    $z = f_{MSSC}$ as in (1)
    For $j := 1, \ldots, K$ **Do**
        Calculate centroid c_j
Until m does not change or l =Maxit

Maxit: (the maximum iteration allowed)

Degeneracy of K-means clustering:
It has been observed that the final solution of MSSC problem obtained by KM heuristic depends substantially on the initial choice of centroids. Since most of these algorithms generate random initializations for centroids, the degeneracy could occur with those of bad initials or choices.

Degeneracy:
We say that solution of the clustering problem is degenerate, if either: (i) there is one or more cluster centers have no entities allocated to them or (ii) two or more cluster centers are identical.

Degree of Degeneracy:
We say that degenerate solution has degree of degeneracy equal to d if the number of empty clusters in the solution is equal to d.

Clustering
K-Means
Degeneracy
Example
Addressing
Computation Results
Conclusion
References

Initialization:
initial cluster centers are located at customer locations 75, 63 and 65 when 3 clusters are desired.

If K = 4, the same initial solution is suggested in addition to location 61 for the fourth cluster.

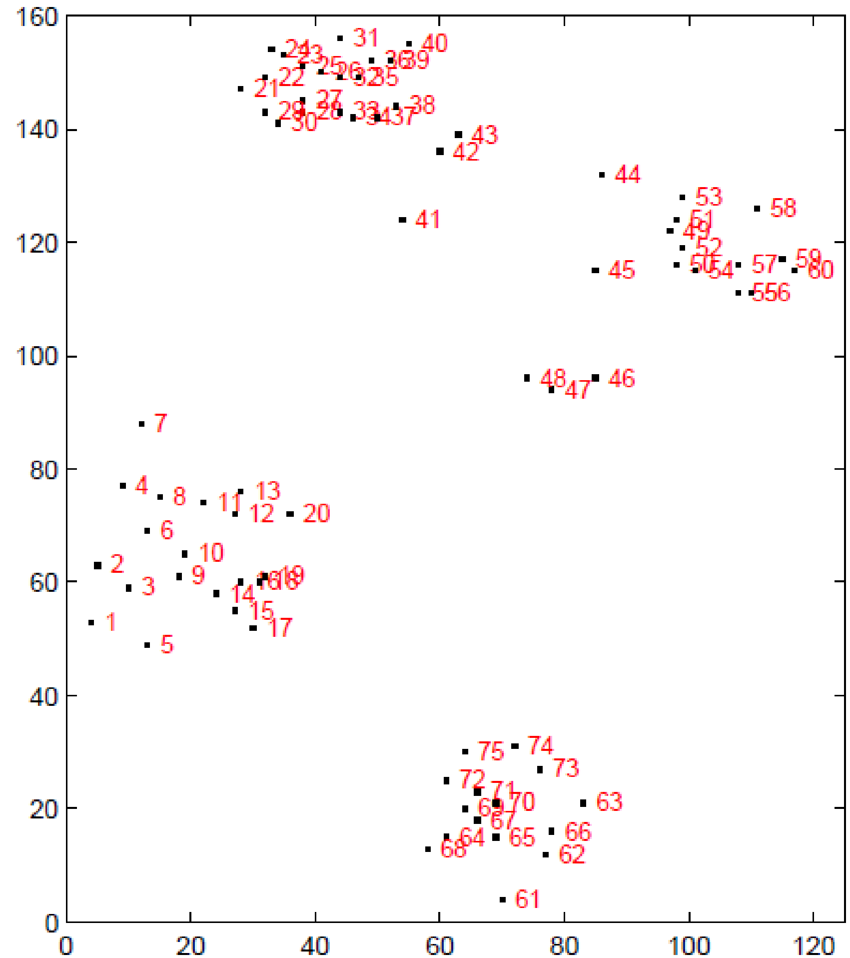


Fig. 2: Ruspini dataset representation.

[Clustering](#)

[K-Means](#)

[Degeneracy](#)

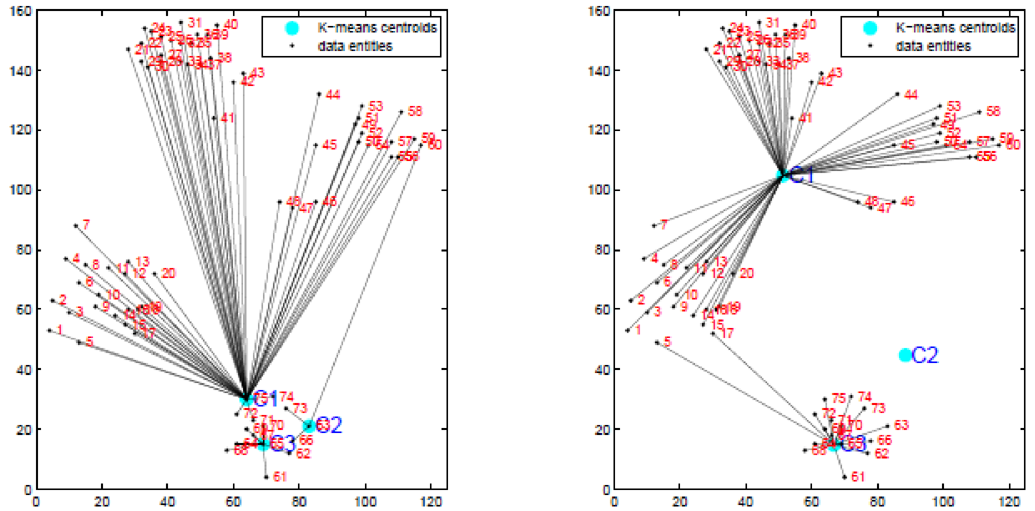
[Example](#)

[Addressing](#)

[Computation Results](#)

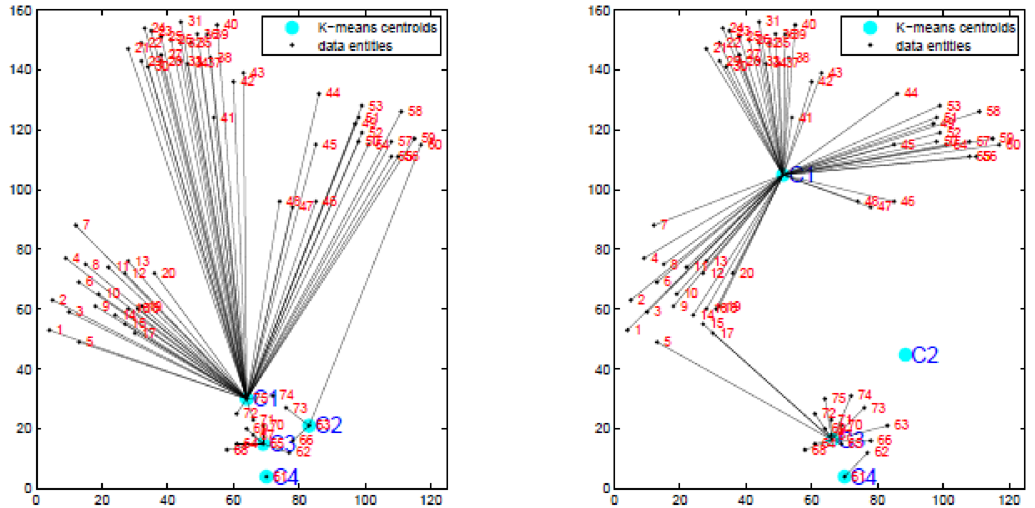
[Conclusion](#)

[References](#)



(a) Initial solution and the objective function = 536386 and $K = 3$.

(b) First Iteration and the objective function = 142857.136 and $K = 3$.



(c) Initial solution and the objective function = 536264 and $K = 4$.

(d) First Iteration and the objective function = 142199.057 and $K = 4$.

Clustering
K-Means
Degeneracy
Example
Addressing
Computation Results
Conclusion
References

Function KMDEG $(C, K, C, Maxit, N, C, z)$

1. $C^{(new)} = \{c_1, c_2, \ldots, c_K\}$ — *K centroids are chosen from X.*
2. $i \leftarrow 0$ — *i-iteration counter*
3. *repeat*
4. $i \leftarrow i + 1$
5. $C \leftarrow C^{(new)}$
6. $z = f_{MSSC}(C)$
7. *Indicate indices b_ℓ of degenerate solutions* $(\ell = 1, \ldots, g)$
8. *if* $g > 0$ *then*
9.   *for* $\ell := 1, \ldots, g$ *do*
10. $t = b_\ell$
11. $h = 1 + n * RND$ — *choose an entity h at random*
12.   *for* $\beta := 1, \ldots, q$ *do*
13. $c_{t\beta} = x_{h\beta}$
14. *until* z *does not change or* $i = Maxit$

K	mth	obj	% dev	maxit	maxdeg	time
4	KM	50589.47	1.49	8	1	0.04
4	KM+	49833.99		5		0.03
6	KM	11008.8	12.15	7	1	0.09
6	KM+	9670.8		8		0.11
8	KM	10643.21	25.67	5	1	0.11
8	KM+	7911.43		5		0.05
10	KM	9486.38	4.73	5	1	0.13
10	KM+	9037.95		5		0.09

(a) Dataset: Ruspini-75

K	mth	obj	% dev	maxit	maxdeg	time
20	KM	182.4	0.06	12	1	0.18
20	KM+	172.2		10		0.11
40	KM	144.4	0.20	12	2	0.24
40	KM+	114.8		9		0.13
60	KM	131	0.13	10	3	0.18
60	KM+	114		8		0.15
80	KM	125.3	0.21	8	9	0.34
80	KM+	99.2		7		0.11
100	KM	116	0.28	10	9	0.38
100	KM+	83.1		7		0.18

(b) Dataset: Glass-214

K	mth	obj	% dev	maxit	maxdeg	time
20	KM	8905.1	0.00	21	1	1.02
20	KM+	8888.7		23		1.08
40	KM	7230.3	0.01	15	2	1.23
40	KM+	7143.3		15		1.22
60	KM	5868.1	0.00	12	4	1.53
60	KM+	5870.2		10		1.24
80	KM	5459.7	0.04	11	13	1.89
80	KM+	5220.1		9		1.67
100	KM	4917.6	0.05	9	17	2.45
100	KM+	4648.4		12		2.11

(c) Dataset: Breast-Cancer-699

K	mth	obj	% dev	maxit	maxdeg	time
20	KM	6399635.3	0.00	19	1	0.35
20	KM+	6393096.5		23		0.39
40	KM	4596055.5	0.01	31	1	1.15
40	KM+	4572142.9		22		0.76
60	KM	4238761.4	0.01	47	1	1.77
60	KM+	4186300.5		26		1.05
80	KM	3237280.9	0.13	51	9	2.00
80	KM+	2818305.4		23		1.14
100	KM	2937550.8	0.25	27	9	2.29
100	KM+	2201357.8		21		1.77

(d) Dataset: Image Segmentation-2310

Conclusion:

The Minimum Sum Of Squares Clustering (MSSC) problem is considered and the algorithm (KM) is designed to solve it. It has been observed that the K-Means (KM) clustering heuristic for solving (MSSC) poses the property of degeneracy, i.e., the property that some clusters could remain empty (without entities) during the execution or at the code.

I explain the degenerate solutions and provide an efficient and easy procedure which removes degeneracy immediately when it appears in iterations.

Future work:

- Diagnosing the degeneracy.
- Applying VNS to improve the solution.

References:

[1] Aloise, D., A. Deshpande, P. Hansen and P. Popat, *NP-hardness of Euclidean sum-of-squares clustering*, Machine Learning 75 (2009), pp. 245–248.

[2] Aloise, D. and P. Hansen, *Clustering*, in: Handbook of Discrete and Combinatorial Mathematics, CRC Press, 2010 .

[3] Blake, C. and C. Merz, *UCI repository of machine learning databases*, http://archive.ics.uci.edu/ml/datasets.html (1998), [Online; accessed Jan 12, 2012].

[4] Brimberg, J. and N. Mladenovi´c, *Degeneracy in the multi-source Weber problem*, Mathematical Programming 85 (1999), pp. 213–220.

[5] Forgy, E., *Cluster analysis of multivariate data: efficiency versus interpretability of classifications*, Biometrics 21 (1965), pp. 768–769.

[6] Hansen, P. and N. Mladenovi´c, *J-means: a new local search heuristic for minimum sum of squares clustering*, Pattern Recognition 34 (2001), pp. 405–.413

[7] Kaufman, L. and P. Rousseeuw, *"Finding groups in data. An introduction to cluster analysis",* A JOHN WILEY & SONS INC. PUBLICATION, 1990.

References:

[8] MacQueen, J., *Some methods for classification and analysis of multivariate observations*, Berkeley, CA: University of California Press 1 (1967), pp. 281–297.

[9] Mirkin, B., *“Clustering for data mining: a data recovery approach,”* Taylor & Francis group, FL, 2005.

[10] Mladenovi´c, N. and J. Brimberg, *A degeneracy property in continuous location-allocation problems*, Les Cahiers du GERAD G-96-37 (1996).

[11] Ruspini, E., *Numerical methods for fuzzy clustering*, Information Sciences 2(1970), pp. 319–350.

[12] Xu, R. and D. Wunsch, *“Clustering,”* IEEE Press, 2009.

THANK YOU FOR LISTENING