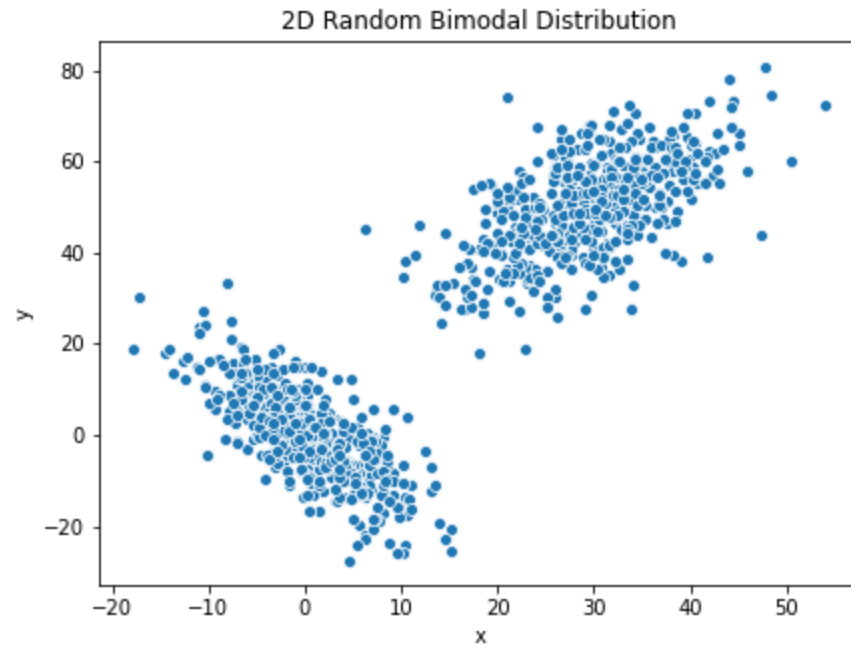# Constrained Clustering

Krishna Teja

Problem Statement

Stated as : Select a portion of the data as separate cluster and continue clustering

Random Data Generation

1. Generate 2D Multivariate normal data

2. Establish GroundTruth based on the desired conditions (x>0 & y<0)

3. Use Ground truth for constraint generation or evaluate Performance Metrics

Metrics

1. NMI

2. Rand Score

Existing Literature

- 71 papers proposing algorithms

Category	Method		
k-Means	COP-COBWEB (Wagstaff and Cardie, 2000)	Spectral Graph Theory	Adjacency Matrix Modification (Kamvar et al, 2003)
	COP-KMeans (Wagstaff et al, 2001)		Out-Of-Sample Adjacency Matrix Modification (Alzate and Suykens, 2009)
	Seed-KMeans (Basu et al, 2002)		CSP (Wang and Davidson, 2010a; Wang et al, 2014)
	Constrained-KMeans (Basu et al, 2002)		Constraint Satisfaction Lower Bound (Wang et al, 2010)
	ICOP-KMeans (Tan et al, 2010)		Inconsistent Constraints (Rangapuram and Hein, 2012)
	Sequenced Assignment COP-KMeans (Rutayisire et al, 2011)		Logical Constraint Combinations (Zhi et al, 2013)
	MLC-KMeans (Huang et al, 2008)		Distance Modification (Anand and Reddy, 2011)
	SCREEN (Tang et al, 2007)		Constraint Propagation Binary Class (Lu and Carreira-Perpiñán, 2008)
	GA Dispersion & Impurity (Demiriz et al, 1999)		Constraint Propagation Multi-Class (Lu and Ip, 2010; Chen and Feng, 2012; Ding et al, 2013)
	CVQE (Davidson and Ravi, 2005)		Kernel Matrix Learning (Zhang and Ando, 2006; Hoi et al, 2007; Li and Ding, 2008; Li and Liu, 2009)
	LCVQE (Pelleg and Baras, 2007)		Guaranteed Quality Clustering (Cucuringu et al, 2016)
	PCK-Means (Basu et al, 2004b)	Ensemble Clustering	SCEV (Iqbal et al, 2012)
	Lagrangian Relaxation (Ganji et al, 2016)		Consensus Function (Al-Razgan and Domeniconi, 2009; Xiao et al, 2016; Dimitriadou et al, 2002)
	Tabu Search (Hiep et al, 2016)		
	Fuzzy CMeans (Grira et al, 2006)	Collaborative Clustering	SAMARAH (Forestier et al, 2010a)
	Non-Negative Matrix Factorisation (Li et al, 2007)		Penta-Training (Domeniconi and Al-Razgan, 2008)
	Mathematical Program (Ng, 2000)	Declarative Approaches	SAT (Davidson et al, 2010)
	Minimal Capacity Constraints (Bradley et al, 2000)		CP (Dao et al, 2013, 2016, 2017; Guns et al, 2016)
	Balanced Clustering (Banerjee and Ghosh, 2006)		ILP Column Generation (Merle et al, 1999; Aloise et al, 2012; Babaki et al, 2014)
	Minimal Size (Demiriz et al, 2008)		Restricted Cluster Candidates (Mueller and Kramer, 2010; Ouali et al, 2016)
	Minimal Size & Balanced Clustering (Ge et al, 2007)		
Metric Learning	Euclidean (Klein et al, 2002)	Miscellaneous	Constrained EM (Shental et al, 2013)
	Mahanalobis (Bar-Hillel et al, 2003, 2005; Xing et al, 2002)		Evolutionary Algorithm (Handl and Knowles, 2006)
	Kullback-Leibler Divergence (Cohn et al, 2003)		Random Forest (Zhu et al, 2016)
	String-Edit Distance (Bilenko and Mooney, 2003)		
	LRML (Hoi et al, 2008, 2010)		
	Partially Observed Constraints (Yi et al, 2012)		
k-Means & Metric Learning	MPCK-Means (Bilenko et al, 2004)		
	HMRF-KMeans (Basu et al, 2004b)		
	Semi-Supervised Kernel *k*-Means (Kulis et al, 2005, 2009)		
	CLWC (Cheng et al, 2008)		

Interesting Algorithms

1. Cop-Kmeans (Constrained KMeans)

2. PCKMeans (Pairwise Constrained KMeans) & Others

3. Preidentify and Kmeans

References

1. Cop-Kmeans (Constrained KMeans) [1]

2. PCKMeans (Pairwise Constrained KMeans) & Others [2]

3. Preidentify and Kmeans

Implementing Algorithms

1. Cop-Kmeans (Constrained KMeans)

2. Preidentify and Kmeans

Cop-KMeans – Implementing Algorithms(contd.)

Table 1. Constrained K-means Algorithm

COP-KMEANS(data set D, must-link constraints $Con_= \subseteq D \times D$, cannot-link constraints $Con_{\neq} \subseteq D \times D$)

1. Let $C_1 \ldots C_k$ be the initial cluster centers.

2. For each point d_i in D, assign it to the closest cluster C_j **such that** VIOLATE-CONSTRAINTS$(d_i, C_j, Con_=, Con_{\neq})$ **is false. If no such cluster exists, fail (return {}).**

3. For each cluster C_i, update its center by averaging all of the points d_j that have been assigned to it.

4. Iterate between (2) and (3) until convergence.

5. Return $\{C_1 \ldots C_k\}$.

VIOLATE-CONSTRAINTS(data point d, cluster C, must-link constraints $Con_= \subseteq D \times D$, cannot-link constraints $Con_{\neq} \subseteq D \times D$)

1. For each $(d, d_=) \in Con_=$: If $d_= \notin C$, return true.

2. For each $(d, d_{\neq}) \in Con_{\neq}$: If $d_{\neq} \in C$, return true.

3. Otherwise, return false.

Cop-KMeans - Implementing Algorithms(contd.)

Steps Involved :

1. Generate all Must-Link and Cannot-Link constraints from Ground Truth of the data

2. ML - datapoints in one cluster

3. CL - datapoints from different cluster

4. When the whole GT is known, every combination can be achieved

5. In reality, we might know less than 5% of the GT

6. So we take these available info from GT(1%) and make them into constraints

7. Run the algorithm as mentioned where constraints are not violated

Custom-KMeans - Implementing Algorithms(contd.)

Steps Involved :

    1. Pre-label all the data points with the desired cluster tag

    2. Find the centroid of that specific cluster

    3. Find the remaning centroids on all of the remaining data

    4. Converge when the centroids do not move

Results

Cop-Kmeans

Prelabel-Kmeans

Limitations

Cop-Kmeans

1. Cop-KMeans is time consuming

2. Sometimes, does not converge after processing for a long time

Prelabel-Kmeans

1. Preldenitying the datapoints is not part of unsupervised/semi-supervised

2. There is no clustering happening in the selected portion