CLUSTERING

QUESTION 1

We can cluster in one dimension as well as in many dimensions. In this problem, we are going to cluster numbers on the real line. The particular numbers (data points) are 1, 4, 9, 16, 25, 36, 49, 64, 81, and 100, i.e., the squares of 1 through 10. We shall use a k-means algorithm, with two clusters. You can verify easily that no matter which two points we choose as the initial centroids, some prefix of the sequence of squares will go into the cluster of the smaller and the remaining suffix goes into the other cluster. As a result, there are only nine different clusterings that can be achieved, ranging from $\{1\}\{4,9,...,100\}$ through $\{1,4,...,81\}\{100\}$. We then go through a reclustering phase, where the centroids of the two clusters are recalculated and all points are reassigned to the nearer of the two new centroids. For each of the nine possible clusterings, calculate how many points are reclassified during the reclustering phase. List out pair of initial centroids that results in *exactly one* point being reclassified.

SOLUTION

Heratim 3 M1, M2 a	2 2	No.	dent let	tel	ided c	make 1
				her	LILES LE	The state of the s
M1 = 9	M2	44 6	1			161
20 1 1	4	-		-		
initial cli	when	- COVE	~ 2 5	2002		
CI > FA	1	C7= 4		Pet L		
eran eran i	F 197	Total Control	Comme			-
almote						WA
Day 4	(u-v)	-	- 51	15	Milante	from MI
-		-	D2_	15	7	M M 2
	44	100	Francisco (C)			
Data point		63				
4	2	65	_		4	
9	0	55	CI	-	300.70	The Real Property lives
	7	98		- 19	2023	-
16.	14	29	CI			-
		10000	ca.			1000
49	111111111111111111111111111111111111111	15	Ca		** ***	CI WHILE WAR
64	55	0	C2		-	A SHADOW TO SERVICE AND ADDRESS OF THE PARTY
81	72	17	CZ			200
(00)	91	36	CZ			2011
1,000	- "	24	102			
C - 8	1.0	9 16 1	15 7			-
	_					-
C3 =4	467	11/2	1,12,15	00 7	-	
Name and Address	2					
Haution		6931	cur		100	
WI -	T.T.	FIEL	6415		11	

	Datapoint	10_	D 2	Cluster	No. 1 No. 1 No.
	1'	10	65	CI	
	4	7	62	CIN	Walter Tal Car
	9	2	57	CI	The Park I
	16	5	20	CL	
	25	14	YI	Charles	Maria Spiritary
	36	25	30	CI	10 10 15 -51
	49	48	17	C2	Tell land
	64	53	2	_C2	
	81	70	15	C2	N. Comments
_	100	89	34	C2	
	C1 = { 1,	1,9,16	, 25, 36	}	But sprate to
	C2 = { 49	,64,8	1,100	}	
	iteration	3	3 8		10 20
	M1 = 1	+4+9	+16 + 25	1+36	15 155
			6	8 1	D 18
		15.16		7777	200 61
	M2 =	49 +64	+81+10	00	B 8 0
	=	73.5	4		D 0 3

-					Fig. 0
1	Duta poi unt	M	12	cluster	
	1	1446	72.5	el ac	(Both failure)
	4	11.16	69.5	CI	THE COURT OF THE C
	9	6.16	64.5		EHAD SA S
	16	6.84	52.5		10-11
	25	9.84	48.5	CI	2002 63 65
	36	20.84	39.5	100	
	49	33.84	24.5		PS-12 23 211 11 12 12
	64	48.84	梅鞋		PERMIT PERMIT NAMED IN COLUMN TO A STATE OF THE PERMIT NAMED IN COLUMN TO A ST
	18	65.84			walker of the only of
	100	84.84	26.5	C2	V 32 00 08
				11411	
1	C1={1,	4,9,16	,25,30	7	a formal document
	C2 = \$	49,6	4,81,10	07	ADMINISTRA
		12 1	AL DE	25	0 0 0 10 10 10 10
ğ	As ove ca	ur see	thed	lata points	in cluster Cl and C2 in iteration 3
1	and ite	evotion	2 av	e same.	data points are said to be
	Converge	d and	l sta	ble.	the said is se
	At is	tial	a tool	1 (91	4), data print 36 is reclassified.
	Outo	dala	noi A	21)	1), doubt print 36 is reclarified.
	Uning	CACCO	police	26 15	reclassified in reclustering phase.
		-	1	N 55	ON THE SEC. STORE

OUESTION 2

Suppose we want to assign points to one of two cluster centroids, either (0,0) or (100,40). Depending on whether we use the L_1 or L_2 norm, a point (x,y) could be clustered with a different one of these two centroids. For this problem, you should work out the conditions under which a point will be clustered with the centroid (0,0) when the L_1 norm is used, but clustered with the centroid (100,40) when the L_2 norm is used. List out those points.

SOLUTION

Given centroids are (0,0), (100, 40).

Given a point (x, y) which could be clustered with a different one of these two centroids.

L1 norm is the Manhattan Distance and L2 norm is the Euclidean Distance. After L1 norm and L2 norm are calculated the values of x and y are 55, 5.

When L1 norm is applied on point (55, 5), the point is clustered with centroid (0, 0). When L2 norm is applied on point (55, 5), the point is clustered with centroid (100, 40).

QUESTION 3

Suppose our data set consists of the perfect squares 1, 4, 9, 16, 25, 36, 49, and 64, which are points in one dimension. Perform a hierarchical chustering on these points, as follows. Initially, each point is in a cluster by itself. At each step, merge the two clusters with the closest centroids, and continue until only two clusters remain. Which centroid of a cluster that exists at some time during this process? Positions are represented to the nearest 0.1.

SOLUTION

	Questim 3	Card	mid ance	dista	nce i	s c	alculo	ded	wing endidenn
1	Cluster/cestmid	110000000	4	9	16 2				44
	The Later	0_	All .	Sm	allest	dish	tinet	in lab	- E - CO
	u	(3)	0			Sec.	5000	1 103	bank has
	9	8	5	0		-121	1652	- Done	Su sonal
	16	15	12	7	0	40			
(10)	25	24	21	16	9	0	Met	50	Harrie
5000	36	35	32	27	20	18	0	dora	Outer
	49	48	45	40	33	29	13	0	
	69	63	60	55	48	39	28	K	0

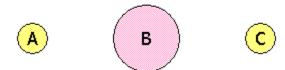
3	dustr Centraid	(P)	9	16	25	36	49	64	The same	The	144=5/1=7
	(1,4) C=2.5	0	_ <				The same of	V.L			
	9	(6.5)	0	malle	nt di	Stand	2				
	16	13.5	7	0							
	25	22.5	16	9	0			-		-	
	36	33.5		20	11	0				-	
	49	46.5	40	33	24	12	~				
	64	61.5	55	48	39	28	15	0		-	

0	cluster/con 1-1	1		_			18	= 0
	cluster/centrial	(47)	16 25	36 49	64			
	16			_				
	25	11-4	90					
	36	31-4						
	49	44.4		1 13 0		-		
	64			28 15	0			
9	(1,4,9) C=4.6	(14,9,	14) (14	15) 16	49	64		
	(14,25) c=20.5	15.9	0					
	36	31.4	15.5	0				
	49	44.4	28.5	(13)	0			
	64		43.5		15	0		
3	cluste/actorid	(LY,9)	(14,25)	(3649)) 64			
	(49) c=4.6	0	119					
	(4,25) c=20.5	15.9	0				5	
	(3449) C=42.5	37.9	22	0			EQ.	
	64	59.4	43.5	21.5	0			1900-

(74,14,25) C= 11 (74,49) C=47.5	21.5
	3 215 0
Clase for boid	
Light Confroid	(14,9,1425) (36,49,64)
144/16/28 C=11	0
(1549,64) C=49.6	38.6

QUESTION 4

Suppose that the true data consists of three clusters, as suggested by the diagram below:



There is a large cluster B centered around the origin (0,0), with 8000 points uniformly distributed in a circle of radius 2. There are two small clusters, A and C, each with 1000 points uniformly distributed in a circle of radius 1. The center of A is at (-10,0) and the center of C is at (10,0).

Suppose we choose three initial centroids x, y, and z, and cluster the points according to which of x, y, or z they are closest. The result will be three *apparent* clusters, which may or may not coincide with the *true* clusters A, B, and C. Say that one of the true clusters is *correct* if there is an apparent cluster that consists of all and only the points in that true cluster. Assuming initial centroids x, y, and z are chosen independently and at random, what is the probability that A is correct? What is the probability that C is correct? What is the probability that both are correct?

SOLUTION

Given centroids are x, y, z We can assign each of x, y, z to A, B, C in 27 possible ways.

Chance of being in A is 1000/10000 = 0.1

Chance of being in B is 8000/10000 = 0.8Chance of being in C is 1000/10000 = 0.1

There are 6 different cases to interchange x, y, z in A, B, C which will be total 27.

The probability that A is correct is 24% The probability that C is correct is 24% The probability that A & C are correct is 4.8%

QUESTION 5

Perform a hierarchical clustering of the following six points:



using the *complete-link* proximity measure (the distance between two clusters is the largest distance between any two points, one from each cluster). Find out a cluster at some stage of the agglomeration?

SOLUTION

sints	on S	(0,0) , 0	4 (10	(01)	. c	(21/2	D. D.	(33):	33),	E(5,23
	P	(29	5,6)					Tinas.	-		
Alta	nce	wet	ñπ,	d=	10	b	ni)2+	(4,-9,)2		
dest	unte	not	nin		10.		1	1000	44.00	10	780
	A	8	6 1	0.0	EE		100	-	10000		-
A	6							37			
B	(19)	0	1	-			(and a	Veril			1111
C	30	15	0	41			prai	-	0	Silvia	SE MILL
D	44	-	100						D. S.I	3400	she.
5	23	18	17	21	0		100		412		
F	29	18	16	28_	31	0			0.3430		Tex.
To	42	4	-	1			1		Mary	Lu	160
	of 1		-	10.4	111	ha.					
-	> D P	C TLD	1200-22								
								a	Carried Street	Line	- Alican
राक्त्र	9 00	mplet	te 1	Unto	ge_	di	sterie	9 0	D	6	E (p.
tonis n	ng co	mpled	te di	CC B	ge_	di	sterie	9	D	61	E (m)
todia h	ng on	mplet d(c,1	te), di	CC B	3	di	AG	9	D	6	E (In
max Max	ng co	mpled d(c, 1 (A)	te 1 1), d - 30 d(0,1	(CE)) (A,	di	sterie	C	D	6	E (S)
Maria Maria	19 cm	mpled d(c,1 18) (18)	te 1, di = 30 d(0,1	(CE.	35 (A.	di A B	AG O	C	D	6	B. (A)
Mar Mar Mar Mar	1 (30, (d(0), (d(0), (43,	mpled d(c,15) (15) (12)	te 10, di = 30 d(0,1 = 17	(C.E.	(A,	di A B	AB O 30	0 17	0	6	E (S)
max Max Max Max	19 cm	mpled d(c,15) (15) (12)	te 10, di = 30 d(0,1 = 17	(C.E.	(A,	di A B	AB O 30	0 17 17	0	-	E D
Mar Mar Mar Mar	1 (30, (d(0), (d(0), (43,	mpled d(c,15) (15) (12)	te 10, di = 30 d(0,1 = 17	(C.E.	(A,	A B C DE	AB 0 30 47 27	0 17 17	0 27	-	E
Mar Mar Mar Mar	1 (30, (d(0), (d(0), (43,	mpled d(c,15) (15) (12)	te 10, di = 30 d(0,1 = 17	(C.E.	(A,	A B C DE	AB 0 30 47 27	0 17 17	0 27	-	E CO
max max	1 (30, (d(0), (d(0), (43,	mplet d(c,16) (16) (12) (12) (18)	te 10, di = 30 d(0,1 = 17	(C.E.	(A,	A B C DE	AB 0 30 47 27	0 17 17	0 27	-	D
max max max	19 cm (130, (14), (13),	mplet d(c,1),(s) (,1s) (,1s) (,1s)	te	(C. E.	(A,	CL ABO DE F	AB 0 30 47 27	0 17 17	0 27	-	E O
max max max	19 co 100 (100 (100 (100) 100 (100) 100 (100)	mplet d(c,1),(s) (,1s) (,1s) (,1s)	te	(C.E.	(A)	CLU A BO C DE F	AB 0 30 47 27	0 17 17 16	0 27 18	31	2/2-1
max max max max	19 co 100 (100 (100 (100) 100 (100) 100 (100)	mplet d(c,1 ,15) ,22) ,18) =	te	(C.E.	(A)	CLU A BO C DE F	AB 0 30 47 27	0 17 19 16	0 27 28	31 (0,F)	A,FB)
max max max	19 co 100 (12) 1 (12) 1 (12) 1 (22) 1 (22) 1 (22) 1 (22)	mplet d(c,1 ,15) ,22) ,18) =	te	(C.E.	(A)	CLU A BO C DE F	AB 0 30 47 27	0 17 13 16 NO.	0 27 28 (CA,)	O, F	A,FB))= 29
max max max	19 co 100 (100 (100 (100) 100) 10	mplet d(c,1 ,15) ,15) ,12) ,12) ,12) ,12)	te	(C.E.	(A)	CLU A BO C DE F	AB 0 30 47 27	0 17 13 16 NO.	0 27 28 (CA,)	O, F	A,FB)

