Solemne 3

Gabriel Gonzalez Cancino 18641538-8

1. Indique si cada item es verdadero o falso. Las respuestas correctas tienen 2 pts. No hay penalizacion´ en las incorrectas. [total: 30 pts].

a) A medida que la ventana de kernel se hace mas grande, la varianza de las predicciones del algoritmo de Kernel Regression (KR) tiende a disminuir.

Verdadero

b) El algoritmo Locally Weighted Regression (LWR) aplicado en data de testeo con M datos requiere aplicar M regresiones lineales.

Verdadero

c) El paso mas difícil del algoritmo de Case-Based Reasoning (CBR) corresponde a la adaptacion de la solucion encontrada al caso actual.

Verdadero

d) El algoritmo PEBLS no requiere un algoritmo de entrenamiento ya que el procesamiento se realiza durante el testeo, de forma analoga que el algoritmo KNN.

Falso

e) El algoritmo K-Means se puede aplicar a datos de tipo nominal y por lo tanto es recomendable su uso en este caso.

Falso

f) Dado 100 000 registros bidimensionales y asumiendo 1000 means iniciales, el algoritmo Mean Shift podría devolver un solo cluster.

Verdadero

g) Considerando el numero de datos, el clustering aglomerativo tiene una complejidad computacional mayor a la del algoritmo Mean Shift.

Verdadero

h) A medida que el parametro epsilon (´) se incrementa, en general el algoritmo DB-SCAN obtendra una menor cantidad de clusters donde el mınimo es 1. Asuma que lo aplico a un dataset de millones de datos y que se fija MinPts con el valor de 2.

Verdadero

i) La tasa de falsos positivos puede ser igual que la tasa de accuracy.

Verdadero

j) Asuma que tengo de 1000 personas donde de ellas 100 estan realmente infectadas. Ahora imagine que aplico un test donde detecta a todos (¡los 1000!) como infectados. Naturalmente,

asuma que la clase positiva corresponde a caso donde una persona esta infectada. Este test tiene al máximo valor posible de sensibilidad.

Verdadero

2. Responda la siguientes preguntas, justificandolas:

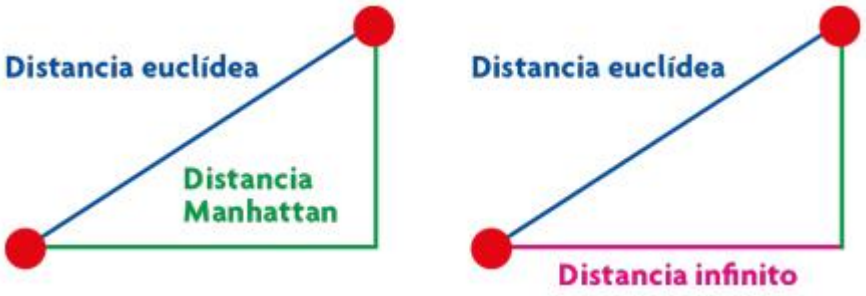
a) [12 pts.] Indique una base de datos de 6 entradas unidimensionales todas diferentes entre si, y 2 centros; tal que al aplicar K-Means este requiera al menos dos calculos de medias. Ahora agrege una nueva variable a la original, es decir ahora es un dataset bidimensional, tal que requiere solo 1 calculo ´ de medias. Recuerde que K-Means consiste en los pasos de: 1. calculo de media (excepto primera iteracion ya que es conocida) y 2. asignacion de centros. Hint: considere distancia de Manhattan para mayor facilidad en paso de asignacion de centros.

2 calculos

¡--1—2—3----11—12—13--¡

u1=	11			dist 1	dist 2	U			
u2 =	16	1	p1	10	15	1		u3=	14
		2	p2	9	14	1		u4=	0
		3	p3	8	13	1			
		11	p5	0	5	1			
		12	p6	1	4	1			
		13	p7	2	3	1			
		iteracion 2							
				dist 1 = 14	dist 2 = 0	U			
		1	p1	13	1	1		u5=	2
		2	p2	12	2	1		u6=	12
		3	p3	11	3	1			
		11	p5	3	11	2			
		12	p6	2	12	2			
		13	p7	1	13	2			
		iteracion 3							
u5 =	2			dist 1 = 2	dist 2= 12	U		u7=	2
u6 =	12	1	p1	1	11	1		u8=	12
		2	p2	0	10	1			
		3	p3	1	9	1		u5=u7	
		11	p5	9	1	2		u6=u8	
		12	p6	10	0	2			
		13	p7	11	1	2			

Ahora agregando otro punto (1,1) y utilizando manhatan utilizar la distancia de las sumas absolutas



por lo tanto es como si se agregara un 2 en cuanto a la distancia ya que seria la suma de “cateto rosado y cateto verde” desde el origen y en cuanto a la distancia con el centroide se calculara manualmente

parte 2

iteracion 1

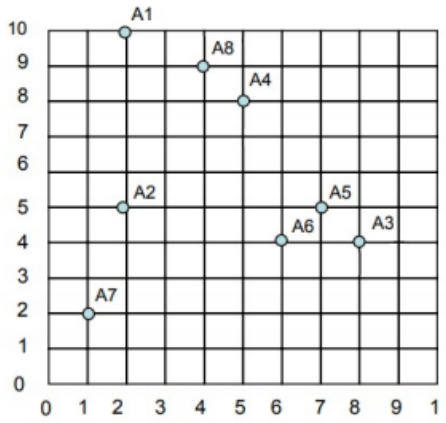
			dist 1 = (0,2)	dist 2 =(0,12	U		
2	2	p1	2	12	1	u9=	2
12	1	p2	1	11	1	u10=	12
	2	p3	0	10	1		
	3	p4	1	9	1		
	11	p5	9	1	2		
	12	p6	10	0	2		
	13	p7	11	1	2		

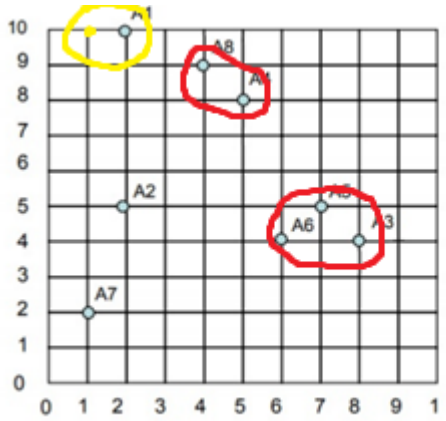
parte 2

iteracion 2

			dist 1 = (0,2)	dist 2 =(0,12	U		
2	2	p1	2	12	1	u11=	2
12	1	p2	1	11	1	u12=	12
	2	p3	0	10	1		
	3	p4	1	9	1		
	11	p5	9	1	2		
	12	p6	10	0	2		
	13	p7	11	1	2		

c) [12 pts.] Asumiendo el dataset en Figura 1, aplique DBSCAN with MinPoints=1 (ignorando al mismo punto del cual se obtienen vecindades) y epsilon=2. ¿Cuantos clusters obtiene y cuales son?. ¿Que nuevo punto agregaria a dataset para que obtenga 1 cluster mas?.





Los clusters son los encerrados en rojo de radio 2

Cluster 1: A6,A5 y A3

Cluster 2: A4 y A8

Y para formar un nuevo cluster es creando uno en (1,10) dibujado en amarillo

d) [8 pts.] Indique 4 diferencias entre algoritmos K-Means y DB-SCAN. Indique un caso donde preferiria K-Means y otro donde prefiera DB-SCAN.

DB-SCAN	Kmeans
No requiere que uno especifique el número de clústeres en los datos a priori	Requiere que uno especifique el número de clústeres en los datos a priori
DBSCAN tiene una noción de ruido y es robusto para los valores atípicos.	Kmeans no tiene noción de ruido
Es mas lento	Es mas rápido

DBSCAN no puede agrupar bien los conjuntos de datos con grandes diferencias en las densidades, ya que la combinación minPts no se puede elegir adecuadamente para todos los clústeres.	Kmeans trabaja mejor con conjuntos de datos con grandes diferencias en las densidades

-DB-scan lo preferiría cuando los datos se distribuyen esféricamente

-Kmean lo usaría cuando se presentan datos de alta dimensión

e) [8 pts.] Asuma que tengo de 1000 pollos en granja donde de ellas 100 estan realmente infectadas con una gripe. Ahora un companero del curso de Ciencia de Datos le plantea que usted debe indicar que ˜ basta que asuma que cualquier pollo en granja tiene la enfermedad. El le asegura que si aplica criterio va a obtener una sensibilidad mayor al 99 % (0.99). ¿Es correcta la afirmacion sobre sensibilidad?. ´ ¿La parece correcto el criterio?. Si no le parece correcta, ¿que metrica utilizarıa para mostrar error a su companero?. Indique su razonamiento en ˜ ultima pregunta. ´

	true	false			
true	100	900		Sensivility	1
false	0	0		acuracy	0,1

La afirmación de la sensibilidad es correcta. No es correcto el criterio ya que , el accuracy es demasiado bajo indicando solamente del 0,1 ya que según lo calculado y esta métrica es la mas asertiva a mi parecer.

$$\frac{100+0}{1000} = 0{,}1$$