ARINJAY-JAIN (A20447307)

final exam - Data Mining 422.

Problem 1.1

- DBSCAN → Define density in a region of feature space based on a center-based methodology

- Density of a point in a region is defined by the number of points with a radius of itself
  → hyperparameter of ~~dr~~ radius.

  Definition → Radius : Eps.

  Regions based on Eps.

  Dense Region I = core point
  " Edge = Boarder points
  Low / No Density = noise point (Background)

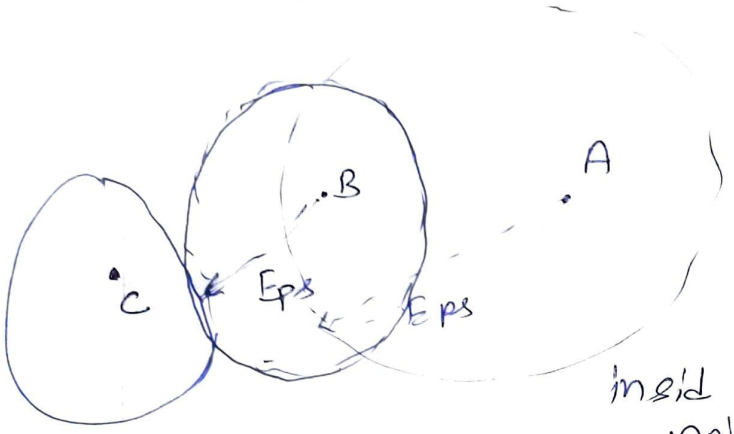
Core points with in Eps radius the naimber of points including itself is > MinPts.
→ use define.

Border points : within a neighborhood Eps radius of a core point, but does not have enough points around it to be a core point.

Noise points : Neither enough points with Eps to be a core points and ~~not~~ not ~~be~~ near within Eps to a core points.

Eg:- Eps = 10

min pts = 7

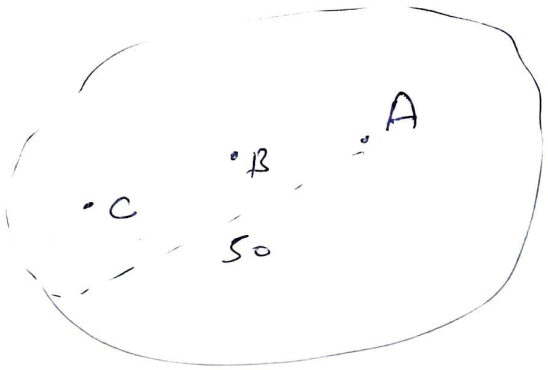


A = core point
B = border point
C = Noise point.

Noise point should not exist inside of the neighborhood a core point. But if we ~~take~~ given a large value of Eps (radius) then it could be possible we capture all the points in one cluster.

eg⇒

Eps = 50



1.2) total point ⇒ n
features = d

Multivariate Normal

→ mahalanobis distance

$$d(x, \bar{x}) = (x - \bar{x})\, \Sigma^{-1} (x - \bar{x})^T$$

Covariance matrix

	x_1	x_2
x_1	σ_1^2	$\sigma_1 \sigma_2$
x_2	$\sigma_1 \sigma_2$	σ_2^2

maximum likelihood estimate :

$$\mu = \frac{1}{n} \sum_{1}^{n} x_i$$

$$\Sigma = \frac{1}{n} \sum_{i=1}^{n} (x^{(i)} - \hat{\mu})^T (x^{(i)} - \hat{\mu}) \quad \text{(average of } d \times d \text{ matrices.)}$$

$$N(x, \mu, \Sigma) = \frac{1}{(2\pi)^{d/2}} |\Sigma|^{-1/2} \exp\left\{ -\frac{1}{2} (x - \mu)\, \Sigma^{-1} (x - \mu)^T \right\}$$

log liklihood

$$= \log\left(\frac{1}{(2\pi)^{d/2}}\right) + \log |\Sigma|^{1/2} + \frac{-1}{2} (x - \mu)\, \Sigma^{-1} (x - \mu)^T$$

if we remove ~~the~~ an anomaly then log-likehood may ~~increase~~ decrease ~~or not~~ because we removed some part of distance magnitude. like distence contribution by anomaly will not count in $(x-\mu)^T \Sigma^{-1} (x-\mu)^T$.

If the features are uncorrelated then variance (Correction distence) will ~~incre~~ increase so ~~we~~ will get high magnitude.

1.3) Root mean Square error (RMSE)

Assume that M is an n by m utility matrix with some entries blank while U and V are matrices of dim n-by-d and d-by-m for some d. Also let P = UV, the product matrix P.

1) Subtract each value of the calculated matrix from the original

2) Square each value in the new matrix

3) add each row together.

4) sum all value the n by 1 matrix

5) divide the by the total number of provided rating from our original utility matrix

6) take the Square root of this result and the value is the RMSE.

1.4 $N = 100,000 = 10^6$

1% of $N = 10^6 \times 10^{-2} = 10^4 = n_i$

length of single doc = 1000

$f_{ij} = 17$

$Tf_{ij} = \frac{17}{1000}$ $IDF = \log \frac{10^6}{10^4} = \log 10^2$

$= 2$

$Tf - IDf = \frac{17}{1000} \times 2$

$= \frac{34}{1000} = 0.034$

1.6

$N = 10000000 = 10^7$ $k = 5$ 27 letter

Check matrix dimensionality = 27×10000000

Characteristic. = 27×10^7

Signature matrix. ~~5×10^7~~

~~vector~~ we have 50 permutation. and 5 shigles.

then dim = 50×5

$r = 5$

$b = 10$

$P(D_1 \& D_2 \text{ identical in a particular bin}) = (0.8)^r$

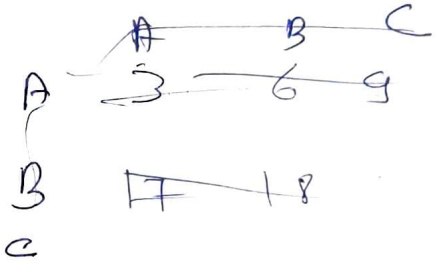
$= (0.8)^5 = 0.328.$

1.7

$A = (3, 6, 9)$

$B = (17, 18, 19)$

$C = (-2, -4, -8)$



Centroid of $A = \frac{3+6+9}{3} = \frac{18}{3} = 6$

Centroid of $B = 18$

Centroid of $C = \frac{14}{3} = 4.66$

Single link $= \min((A, B), (B, C), (A, C))$

Single link $= (A, C) = 1.34$

Complete $= \max((A, B), (B, C), (A, C))$

$= (B, C) = 13.34$

1.8) the objective of minimizing squared error cause K-mean to break the large cluster Hence K-means is not good at deriving clusters of variable sizes at least when they are not well separated. ~~Hence~~

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} d(C_i, x)^2$$

also means we are trying to maximize cluster cohesion.

$$C_k = \frac{1}{m_k} \sum_{x \in C_k} x_k$$

Optimal centroid value for over optimization solution is the mean of each cluster!

post Processing →

~~A~~im is to improve over clustering (reduce SSE) after the algorithm has complete.

- Split a cluster.
- Introduce a new cluster centroid.

SSE ↓ if we ~~d~~reduce the K (no. of cluster) but it we may increase the distance b/w Centroid C_i and X (data point) so it also a cause of ↑ SSE again.

1. 5)

Dimensionality of Markov matrix = $n \times n$. (n: web pages.)

each col represents the out links from a web page. Hence the sum is equal to 1

Spider trap:)

is a set of nodes with no dead ends and out links are with in the trap.

take $v_0 = \left(1/n, \frac{1}{n} \cdots \frac{1}{n}\right)^T$

$M^{\infty} v_0 = (0, 0, 1, 0 \cdots 0)^T$

Hence everyone ends up at K nodes.
for K node page rank score is 1.
for (n-K) node the page Rank score will be Zero (0).

to avoide this problem we com method called taxation.

2.1)

Silhouette score / coff.
for a set of sample data points is used to measure how dense and well seperated the clusters are.

mean intra-cluster distance (a) : the avg distance from x_i to other points in C

mean inter(nearest)-cluster distance (b) : the avg distance from x_i to other points in D.

Silhouette Score

$$S = \frac{b_i - a_i}{\max(b_i, a_i)}$$

range = $[-1, 1]$

lowest value (-1) tells us the samples are assigned in the wrong clusters,

high value (1) : Clusa is dense and well-separated we want $a_i \to 0$ to get S_i to 1.

Zero (0) $\Rightarrow$ overlapping cluster with sample very close to the decision boundary of the neighboring cluster.

2.2)

$\mu = 100$

$\sigma^2 = 10000$. $\quad SD = \sqrt{10^4} = 100$

$Pr(\mu - 4\sigma \leq X \leq \mu + 3\sigma) \leq 0.01\%$

≤ 0.0001

Z score $= Z = \frac{X - \mu}{\sigma}$

↑ attributes value.

C.I. $= \bar{x} \pm 4 \frac{\sigma}{\sqrt{n}}$

Confided interval of 4-Stemdaud deviation.

give the value above the mean and belero the mean.

2.3)

$U_1 =$ 4 2 3 2 4 $\quad$ Avg $U_1 = \frac{15}{5} = 3$

$U_2 =$ 5 3 4 3 5

Avg $U_2 = \frac{20}{5} = 4$

Normalize centre ⇒

$U_1 =$ 1 −1 0 −1 1

$U_2 =$ 1 −1 0 −1 1

okay both user profile is similary.

Cosine similiy = Corr = $\frac{(1\times1) + (-1\times-1) + (0) + (-1\times-1) + (1\times1)}{\sqrt{1^2+(-1)^2+(0)^2+(-1)^2+(1)^2} \times \sqrt{1^2+(-1)^2+(0)^2+(-1)^2+(1)^2}}$

$= \frac{1+1+1+1}{2\times\sqrt{4}}$

$= \frac{4+1+1+1}{4} = \frac{1}{}$

$1 > 0.75 \Rightarrow$ highly correlated.

↑

threshold.

2.4) In Content - Based - filtering.

Recommend item i to user u based upon other items $i_1, i_2 \dots i_k$ rated by user u.

methodology ⇒

- Build Item profile based on Item properties
→ we see what a user likes based on Item profiles of brought/rated item → this gives us a user profile

→ we find Item profile that match user profile for recommendation.

Use profile utility matrix.

eg ⇒ ~~family~~ animation adventure ~~drama~~ ~~docum.~~

U_1

	M_1	M_2	M_{20}	M_{100}	
U_1	3	5	2	7	← any rating

Item profile

	family	anima	adu	drama	doc. ↰ genre
i_1	0	1	0	0	0
i_2	1	0	0	0	0
⋮	⋮	⋮	⋮	⋮	⋮
i_{100}					

↗ movie

Un Weighted

	m_1	m_2	m_3	...
U_1	3	4	5	

Average for
User $u = 3$
for non doc genre movie.

weighed for (non-doc movie) ← movies

	M_1	M_2	M_3	...
U (user →)	3-3	4-3	5-3	...
	0 +	1 +	2 +	...

100

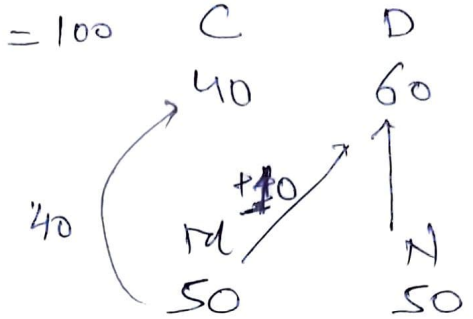
Average for
User $u = 5$
for doc genre movie. ← movies

	M_1	M_2	M_3	...
U (user →)	3-5	4-5	5-5	...
	= -2	-1	0	...

$= -3/3 = -1$

$= -3 + 0 \ldots$
100

3.1)



$$E = -\sum P(w_c) \log_2 P(w_c)$$

$$= -\sum \frac{|w_c|}{n_w} \log \frac{|w_c|}{n_w}$$

$|w_c|$ = is the count of points classified as c in cluster w.

n_w = is the count of points in cluster w.

Entropy D $\Rightarrow$ $n_w = 60$

$|w_{c_1}| = 50$ (class N)

$|w_{c_2}| = 10$ (class M)

\# Entropy D = $-\frac{50}{60} \log_2 \frac{50}{60} - \frac{10}{60} \log_2 \frac{10}{60}$

Similarly

\# Entropy C = $-\frac{40}{40} \log_2 \frac{40}{40} - \frac{0}{40} \log_2 \frac{0}{40}$

$= 0$

Q Lucky 7

1) 2018 ~~ACM~~ AM Turing Award

2) Recipes for pizza

3) $432500

4) OpenAI

5) Pong

6) ~~Hanbi~~ Hanabi

7) Finland