سوال ۱،

$$Recall(i,j) = \frac{n_{ij}}{n_i} \qquad Precision(i,j) = \frac{n_{ij}}{n_j}$$

$$F_{measure}(i,j) = \frac{۲\,Recall(i,j) \times Precision(i,j)}{(Recall(i,j) + Precision(i,j))}$$

$$cluster\ 1 = \{P_۱, P_۲, P_۳, P_۴, P_۵, P_۶, P_۷, P_۸\}$$

↳ class A: $Recall(A,1) = \frac{۳}{۳} = 1 \qquad Precision(A,1) = \frac{۳}{۸}$

$$f_{measure} = \frac{۲ \times ۱ \times \frac{۳}{۸}}{۱ + \frac{۳}{۸}} = \frac{\frac{۶}{۸}}{\frac{۱۱}{۸}} = \frac{۶}{۱۱}$$

↳ class B, $Recall(B,1) = \frac{۵}{۵} = 1$, $Precision(B,1) = \frac{۵}{۸}$

$$f_{measure}(B,1) = \frac{۲ \times ۱ \times \frac{۵}{۸}}{۱ + \frac{۵}{۸}} = \frac{\frac{۵}{۴}}{\frac{۱۳}{۸}} = \frac{۵}{\cancel{۴}} \times \frac{\cancel{۸}^{۲}}{۱۳} = \frac{۱۰}{۱۳}$$

$$cluster2 = \{P_۱, P_۲, P_۴, P_۵\}$$

↳ class A: $Recall(A,۲) = \frac{۲}{۳}$ $\quad Precision(A,۲) = \frac{۲}{۴} = \frac{۱}{۲}$, $f = \frac{۲ \times \frac{۲}{۳} \times \frac{۱}{۲}}{\frac{۲}{۳} + \frac{۱}{۲}} =$

$$\frac{۲}{۳} \Big/ \frac{۴+۳}{۶} = \frac{۲}{۳} \times \frac{۶}{۷} = \frac{۴}{۷}$$

↳ class B: $Recall(B,۲) = \frac{۲}{۵}$ $\quad Precision(B,۲) = \frac{۲}{۴} = \frac{۱}{۲}$

$$f_{measure} = \frac{۲ \times \frac{۲}{۵} \times \frac{۱}{۲}}{\frac{۲}{۵} + \frac{۱}{۲}} = \frac{\frac{۲}{۵}}{\frac{۴+۵}{۱۰}} = \frac{\frac{۲}{۵}}{\frac{۹}{۱۰}} = \frac{۲}{\cancel{۵}} \times \frac{\cancel{۱۰}^{۲}}{۹} = \frac{۴}{۹}$$

cluster 3 = $\{P_۳, P_۶, P_۷, P_۸\}$

↳ class A : Recall$(A, ۳) = \frac{۱}{۳}$, Precision$(A, ۳) = \frac{۱}{۴}$ $f(A,۳) = \frac{۲ \times \frac{۱}{۳} \times \frac{۱}{۴}}{\frac{۱}{۳} + \frac{۱}{۴}}$

$$= \frac{\frac{۱}{۶}}{\frac{۷}{۱۲}} = \frac{۱}{۶} \times \frac{۱۲}{۷} = \frac{۲}{۷}$$

↳ class B : $R(B, ۳) = \frac{۳}{۵}$, $P(B,۳) = \frac{۳}{۴}$, $f(B,۳) = \frac{۲ \times \frac{۳}{۵} \times \frac{۳}{۴}}{\frac{۳}{۵} + \frac{۳}{۴}} =$

$$\frac{\frac{۹}{۱۰}}{\frac{۱۲+۱۵}{۲۰}} = \frac{\frac{۹}{۱۰}}{\frac{۲۷}{۲۰}} = \frac{۹}{۱۰} \times \frac{۲۰}{۲۷} = \frac{۲}{۳}$$

---

cluster 4 = $\{P_۱, P_۲\}$ → class A : $R(A, ۴) = \frac{۲}{۳}$, $P(A, ۴) = \frac{۲}{۲} = ۱$

$f(A, ۴) = \frac{۲ \times ۱ \times \frac{۲}{۳}}{۱ + \frac{۲}{۳}} = \frac{۴/۳}{۵/۳} = \frac{۴}{۵}$

class B : $R(B, ۴) = \frac{۰}{۵} = ۰$, $P(B, ۴) = \frac{۰}{۲} = ۰$ $f = \frac{۲ \times ۰ \times ۰}{۰ + ۰}$ تعریف نشده

cluster 5 = $\{P_۴, P_۵\}$ → class A : $R(A, ۵) = \frac{۰}{۳} = ۰$, $P(A, ۵) = \frac{۰}{۲} = ۰$

↳ $f = \frac{۰}{۰}$ تعریف نشده

class B : $R(B, ۵) = \frac{۲}{۵}$ $P(B, ۵) = \frac{۲}{۲} = ۱$ $f = \frac{۲ \times ۱ \times \frac{۲}{۵}}{\frac{۲}{۵} + ۱} = \frac{۴/۵}{۷/۵} = \frac{۴}{۷}$

---

cluster 6 = $\{P_۳, P_۶\}$ → class A : $R(A, ۶) = \frac{۱}{۳}$, $P(A, ۶) = \frac{۱}{۲}$

$f = \frac{۲ \times \frac{۱}{۳} \times \frac{۱}{۲}}{\frac{۱}{۳} + \frac{۱}{۲}} = \frac{\frac{۱}{۳}}{\frac{۵}{۶}} = \frac{۱}{۳} \times \frac{۶}{۵} = \frac{۲}{۵}$, class B : $R(B, ۶) = \frac{۱}{۵}$

$P(B, ۶) = \frac{۱}{۲}$ $f(B, ۶) = \frac{۲ \times \frac{۱}{۲} \times \frac{۱}{۵}}{\frac{۱}{۲} + \frac{۱}{۵}} = \frac{\frac{۱}{۵}}{\frac{۷}{۱۰}} = \frac{۱}{۵} \times \frac{۱۰}{۷} = \frac{۲}{۷}$

$$\text{cluster } 7 = \{P_۷ , P_۸\} \rightarrow \text{class } A,\ R(A,۷) = \frac{۰}{۳} = ۰ ,\quad P(A,۷) = \frac{۰}{۲} = ۰$$

$f = \frac{۰}{۰}$ تعریف نشده , $\text{class } B:\ R(B,۷) = \frac{۲}{۵} \quad P(B,۷) = \frac{۲}{۲} = ۱$

$$f = \frac{۲ \times ۲/۵ \times ۱}{۲/۵ + ۱} = \frac{۴/۵}{۷/۵} = \frac{۴}{۷}$$

$$F\text{measure}(A) = \max(f\text{measure}(A,i) \mid ۱ \le i \le ۷\} = \max\left(\frac{۹}{۱۱}, \frac{۴}{۷}, \frac{۲}{۷}, \frac{۴}{۵}, \ldots, ۰, ۰, \frac{۲}{۵}, \frac{۴}{۷}\right) = \frac{۴}{۵}$$

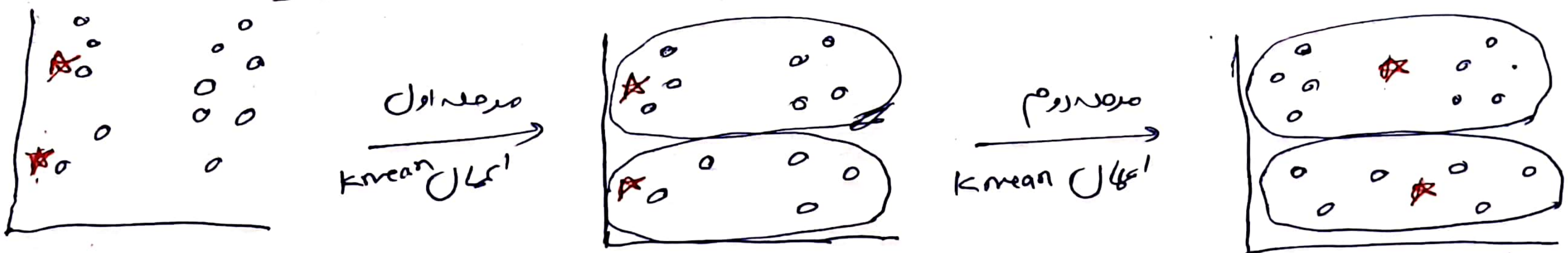
$$f\text{measure}(B) = \max(f\text{measure}(B,i) \mid ۱ \le i \le ۷\} = \max\left(\frac{۱۰}{۱۲}, \frac{۴}{۹}, \frac{۲}{۳}, ۰, \ldots, \frac{۴}{۷}, \frac{۲}{۷}, \frac{۴}{۷}\right) = \frac{۱۰}{۱۲}$$

$$f\_\text{measure} = \frac{n_A}{n} \times F(A) + \frac{n_B}{n} \times F(B) = \frac{۳}{۸} \times \frac{۴}{۵} + \frac{۵}{۸} \times \frac{۱۰}{۱۲} = \frac{۳}{۱۰} + \frac{۲۵}{۵۲}$$

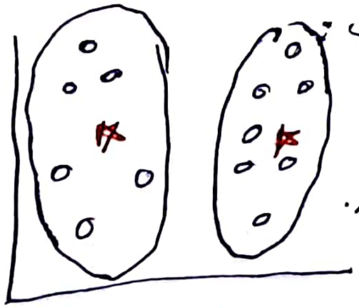
$$= \frac{۱۵۶ + ۲۵۰}{۵۲۰} = \frac{۴۰۶}{۵۲۰} = ۰/۷۸$$

سوال ۲: یکی از عوامل مهم تعیین الگوریتم ممکن است شکل داده باشد. مثلاً اگر شکل خوشه‌ها دایره‌ای نداشته باشد یا چگالی متفاوت داشته باشند الگوریتم kmeans ممکن است همگی را در یک خوشه قرار دهد ولی الگوریتم Db scan در برابر اشکال عجیب مقاومت داشته و می‌تواند به خوبی با آن‌ها رفتار کند. از طرف دیگر وجود نویز می‌تواند باعث شود مراکز خوشه‌ها تحت تأثیر قرار گرفته و به سمت دیگری متمایل شوند به گونه‌ای که تمامی نقاط از مرکز خوشه‌ای که نویز در آن قرار گرفته فاصله عمیقی داشته باشند. در این حالت است که تمامی نقاط در همان خوشه قرار می‌گیرند اما می‌دانیم که Dbscan و نسخه complete link از روش سلسله مراتبی می‌توانند این مشکل را حل کنند. برای مشکل چگال و اشکال داده‌ها نقد single link از خوشه‌بندی سلسله مراتبی نیز می‌تواند به خوبی عمل کند و اشکال غیر بیضوی را به خوبی کلاستربندی می‌کند.

سوال ۳: مثال اول: اگر چنین نمونه داده‌ای داشته باشیم و مراکز تصادفی ابتدایی را این گونه انتخاب کنیم:



که این حالت کلاستربندی مناسب نیست و حالت ایده‌آل این گونه است



که اگر بتوانیم مراکز را به درستی یکی در سمت چپ و یکی در سمت راست حالت اولیه قرار گرفته بودند کلاستربندی به درستی همین حالت انجام می‌شد.

مثال دوم

سوال ۳: می‌دانیم که الگوریتم k means برای داده‌های پرت حساس است زیرا که مرکزها را تصادفی انتخاب می‌کنیم یعنی داده‌ها را به صورت تصادفی به k خوشه تقسیم کردیم و سپس روی هر خوشه میانگین می‌گیریم تا مرکز آن خوشه پیدا شود. اگر داده‌ی پرت به صورت نقطه در یک خوشه قرار گرفته باشد میانگین داده‌ها را به سمت خودش می‌کشد و منحرف می‌شود و در نتایج نهایی اثرات بدی می‌گذارد.

مثلاً اگر ۷ نقطه ۲۵ و ۱۰ و ۹ و ۷ و ۳ و ۲ و ۱ در یک فضای یک بعدی داشته باشیم اگر داده‌ی پرت ۲۵ را حذف کنیم باید به دو خوشه {۳، ۲، ۱} و {۱۰، ۹، ۷} تقسیم شوند نقاط. اما در این مثال الگوریتم k mean خوشه‌ها را با k = ۲ شروع می‌کند و ابتدا افراز دوتایی {۳، ۲، ۱} و {۲۵، ۱۰، ۹، ۷} خواهیم داشت. حال اختلاف هر نقطه با مرکز هر خوشه را به دست آورده:

$$\mu_1 = 2 \quad , \quad \mu_2 = \frac{7+9+10+25}{4} = 12.75$$

اختلاف درون خوشه‌ای

$$= (1-2)^2 + (2-2)^2 + (3-2)^2 + \underbrace{(7-12.75)^2}_{33.06} + \underbrace{(9-12.75)^2}_{14.06} + \underbrace{(10-12.75)^2}_{7.56} + \underbrace{(25-12.75)^2}_{150.06} = 206.74$$

ولی نقطه ۷ به مرکز خوشه اولی نزدیک‌تر است تا دومی. $7-2=5$ ولی $12.75-7=5.75$ پس نقطه ۷ وارد خوشه اول می‌شود

{۷، ۳، ۲، ۱} و {۲۵، ۱۰، ۹}

$m_1 = \frac{1+2+3+7}{4} = 3.25$ $\qquad$ $m_2 = \frac{9+10+25}{3} = 14.67$

در این حالت تمام نقاط از یک خوشه خود کمترین فاصله را دارند و نقطه‌ای جابه‌جا نمی‌شود و در خوشه خود باقی می‌ماند.
پس مشاهده کردیم با وجود داده‌ی پرت ۲۵ مقدار ۷ در خوشه اول قرار نگیرد در حالی که با تخمین داده پرت به صورت دقیق‌تر لازم بود مقدار ۷ در خوشه دوم باشد. پس زمانی که داده‌های پرت وجود داشته باشند انتخاب تصادفی مراکز می‌تواند ضعیف عمل کند و با تکرار روند الگوریتم در نهایت نتیجه درست نباشد.

سوال ۴:

الف) kmeans به داده‌های پرت حساس است و از آنجا که فاصله‌ی میانگین داده‌ها با اکثر داده‌ها بسیار زیاد است هنگام قرار گرفتن داده‌های پرت یک خوشه باعث انحراف و تغییر مقدار میانگین خوشه می‌شوند. اما الگوریتم kmedoids به جای در نظر گرفتن مقدار میانگین اشیا موجود در خوشه به عنوان یک نقطه مرجع از اشیا واقعی برای نمایش خوشه‌ها استفاده می‌کند و باقی اشیا با توجه به شباهت خود با شیء نماینده به هر خوشه اختصاص داده می‌شوند و با توجه به این ویژگی kmedoid نسبت به نویز مقاوم‌تر است اما پیچیدگی هر تکرار در الگوریتم kmedoid برابر با $O(k(n-k)^2)$ است اما پیچیدگی kmean از $O(n\times k\times I\times d)$ است که پیچیدگی در هر تکرار از $O(n\times k\times d)$ است. که برای مقادیر بزرگ n و k محاسبات هزینه‌بر است.
پس kmedoid برای دیتاست‌های خیلی بزرگ به دلیل پیچیدگی زیادش به خوبی عمل نمی‌کند ولی kmeans در این مورد بهتر عمل می‌کند.
از طرف دیگر kmeans روی چگالی و اندازه و شکل خوشه‌ها محدودیت دارد. اگر سایز خوشه‌ها از یکدیگر خیلی متفاوت باشند و یا چگالی خوشه‌ها خیلی متفاوت باشد و اگر شکل غیر نمونه‌های کروی ظاهر باشد kmeans به خوبی عمل نمی‌کند اما kmedoid این محدودیت‌ها را ندارد.

ب): اگر خوشه‌بندی سلسله مراتبی single link باشد می‌توان اشکال غیر بیضوی را به خوبی خوشه‌بندی کند اما در برابر نویز حساس است و به خوبی عمل نمی‌کند. اما اگر complete link باشد در برابر نویز حساسیت کمتری دارد ولی با نمونه‌های بزرگ کروی برخورد می‌کند و دایره‌های بزرگ را می‌شکند و تکه تکه می‌کند ولی روش group average حساسیت کمتری به نویز دارد و کروی با نمونه‌ها برخورد می‌کند. پس هر کدام از این حالات حداقل یکی از این مشکلات را دارند که در برابر نویز حساس باشند یا قابلیت مدیریت کردن داده‌های با سایز متفاوت و یا غیر کروی را نداشته باشند.
از طرف دیگر محاسبه فاصله در ماتریس مجاورت از $O(n^2)$ است که $n-1$ مرحله هم صورت می‌گیرد پس در کل روش از $O(n^3)$ است که پیچیدگی نسبتاً بالایی است. که پیچیدگی kmeans و kmedoids کمتر از این مقدار است.
از طرف دیگر روش سلسله مراتبی هیوریستیک عمل کرده و به عقب برنمی‌گردد پس نمی‌توان خطای ادغام و انشعاب آن را اصلاح کرد. در حالی که kmeans و kmedoids در این مورد بهتر عمل می‌کنند و مشکلی ندارند.