حل تمارین تئوری تکلیف چهارم

سوال اول :مشابه انچه در ویدیوی یوتیوبی که در گروه درس گذاشته شد،حل میشود.

cluster #1 = *{p1, p2, p3, p4, p5, p6, p7, p8}*:
Class = A:
$R(A, 1) = 3/3 = 1,$
$P(A, 1) = 3/8 = 0.375$
$F(A, 1) = 2 \times 1 \times 0.375/(1 + 0.375) = 0.55$
Class = B:
$R(B, 1) = 5/5 = 1,$
$P(A, 1) = 5/8 = 0.625,$
$F(A, 1) = 0.77$

cluster #2= *{*p1,p2,p4,p5*}*
Class = A:
$R(A, 2) = 2/3,$
$P(A, 2) = 2/4,$
$F(A, 2) = 0.57$
Class = B:
$R(B, 2) = 2/5,$
$P(B, 2) = 2/4,$
$F(B, 2) = 0.44$

cluster #3= *{*p3, p6, p7, p8*}*
Class = A:
$R(A, 3) = 1/3,$
$P(A, 3) = 1/4,$
$F(A, 3) = 0.29$
Class =B:
$R(B, 3) = 3/5,$
$P(B, 3) = 3/4,$
$F(B, 3) = 0.67$

cluster #4=*{*p1, p2*}*
Class = A:
$R(A, 4) = 2/3,$
$P(A, 4) = 2/2,$
$F(A, 4) = 0.8$
Class =B:
$R(B, 4) = 0/5,$

$P(B, 4) = 0/2,$
$F(B, 4) = 0$

cluster #5 = {p4, p5}
Class = A:
$R(A, 5) = 0,$
$P(A, 5) = 0,$
$F(A, 5) = 0$
Class =B:
$R(B, 5) = 2/5,$
$P(B, 5) = 2/2,$
$F(B, 5) = 0.57$

cluster #6 = {p3, p6}
Class = A:
$R(A, 6) = 1/3,$
$P(A, 6) = 1/2,$
$F(A, 6) = 0.4$
Class =B:
$R(B, 6) = 1/5,$
$P(B, 6) = 1/2,$
$F(B, 6) = 0.29$
cluster #7 = {p7, p8}
Class = A:
$R(A, 7) = 0,$
$P(A, 7) = 1,$
$F(A, 7) = 0$
Class = B:
$R(B, 7) = 2/5,$
$P(B, 7) = 2/2,$
$F(B, 7) = 0.57$
Class A: $F(A) = \max\{F(A, j)\} = \max\{0.55, 0.57, 0.29, 0.8, 0, 0.4, 0\} = 0.8$
Class B: $F(B) = \max\{F(B, j)\} = \max\{0.77, 0.44, 0.67, 0, 0.57, 0.29, 0.57\} = 0.77$
Clustering: $F = 3/8 * F(A) + 5/8 * F(B) = 0.78$

سوال دوم:

الگوریتم Link Singleدر خوشه بندی از فاصله بین داده ها برای تشکیل خوشه استفاده می کند و داده هایی که فاصله کمتری با یکدیگر دارند را در یک خوشه قرار می دهد. بنابراین، این الگوریتم می تواند بهتر با داده های پرت کنار بیاید و خوشه بندی بهتری را ارائه دهد.
در مورد الگوریتم ،DBSCANاین الگوریتم بر اساس تراکم داده ها خوشه بندی را انجام می دهد. به عبارت دیگر، این الگوریتم به داده هایی که در نزدیکی هم قرار دارند و تراکم بیشتری دارند، به عنوان یک خوشه تشخیص می دهد.

بنابراین، در مجموعه داده هایی که داده های پرت وجود دارد، این الگوریتم می تواند بهتر با آنها کنار بیاید وخوشه بندی بهتری ارائه دهد.پس به طور معیارهای مختلفی نظیر سطح چگالی یک دیتاست تعیین کننده در روش بهینه هستند. لذا باید با اشنایی با خود دیتاست که توسط تجربه متخصصان ان حوزه بدست می اید و بهره گیری از روش درست مهندسی داده به انتخاب روش مناسب پرداخت.

سوال سوم:طبعا بهره گیری از انتخاب رندم اولین نقاط معیار مناسبی نبوده و گاهی ممکن است با چند بار اجرای الگوریتم نیز به جواب درستی نرسید.دلیل های متفاوتی میتواند برای ان وجود داشته باشد.در حالتی ممکن است مشخصات دیتاست(چگالی ،رنج فاصله هایی که داده ها از هم دارند،نحوه تراکم حول فضای موجود و...) باعث این اتفاق شود. در این جا برحسب مشکلی که وجود دارد باید از الگوریتم های دیگری استفاده شود.

در حالت دیگر ممکن است مکانیزمی که منجر به تغییر مراکز کلاستر در iteration های متوالی میشود ، مناسب نباشد.برای مثال ممکن است از مکانیزمی میانه طور استفاده کند که فقط منجر به تشخیص کلاستر هایی با اشکال خاص میشود.

سوال چهارم:

سوال (4
الف)
مزایایk-means:
از نظر محاسباتی کارآمد است و روی مجموعه داده های بزرگ به خوبی کار می کند.
پیاده سازی و تفسیر آن آسان است .برای مواردی که خوشه ها کروی هستند یا شکل مشابهی دارند مناسب است.

سوال چهارم :

سوال 4)

الف)مزایای k-means: از نظر محاسباتی کارآمد است و روی مجموعه داده های بزرگ به خوبی کار می کند . پیاده سازی و تفسیر آن آسان است .برای مواردی که خوشه ها کروی هستند یا شکل مشابهی دارند مناسب است.

ضعف های K-means: به نقاط پرت حساس است و می تواند تحت تأثیر قرارگیری اولیه مرکزها قرار گیرد . فرض میکند که خوشهها واریانس مساوی دارند، که ممکن است در برخی از مجموعههای داده درست نباشد .وقتی خوشه ها اشکال غیر محدب داشته باشند خوب کار نمی کند.

K-medoids مزایای: نسبت به نقاط پرت مقاوم است و می تواند خوشه های غیر کروی را مدیریت کند .می تواند داده های پیوسته و مقوله ای را مدیریت کند .نتایجی را ایجاد می کند که قابل تفسیرتر از k-meansهستند.

K-medoids ضعف ها: این می تواند از نظر محاسباتی گران باشد، به خصوص برای مجموعه داده های بزرگ .به انتخاب متریک فاصله حساس است .ممکن است به حداقل محلی و نه حداقل جهانی همگرا شود. در حالت کلی K-means کارایی بیشتری دارد ولی عملکرد آن تحت تاثیر منفی داده های نویز و پرت قرار میگیرد.

ب) می تواند کاری را که انجام شده است خنثی کند (با حرکت دادن دادهها در اطراف خوشه ها) کیفیت در آن خوب است به طور کلی، باید تعداد خوشه ها مشخص باشد . فقط خوشه های کروی شکل را پیدا کنید.

سلسله مراتبی: نمی توان کاری را که انجام شد بازگرداند - کیفیت ممکن است ضعیف باشد، به تعداد نیاز ندارد از خوشه هایی که باید شناخته شوند .کارآمدتر و موازی تر(تقسیم کنید)، ممکن است فقط دلخواه باشد خوشه های شکل دار