به نام خدا

تمرین سوم داده کاوی

حدیث غفوری 9825413

سوال 1

class A: point p1,p2,p3

class B: points p4,p5,p6,p7,p8

recall = R(i,j) = $\frac{n_{ij}}{n_i}$

recall of class i with respect to cluster j.

precision = P(i,j) = $\frac{n_{ij}}{n_j}$

precision of class i with respect to cluster j.

F-measure = F(i,j) = $\frac{2*R(i,j)*P(i,j)}{R(i,j)+P(i,j)}$

cluster1 = {p1,p2,p3, p4,p5,p6,p7,p8}

class A:

R(A,1) = $\frac{n_{A1}}{n_A} = \frac{3}{3} = 1$

P(A,1) = $\frac{n_{A1}}{n_1} = \frac{3}{8} = 0.375$

F(A,1) = $\frac{2*1*0.375}{1+0.375} = 0.55$

class B:

R(B,1) = $\frac{n_{B1}}{n_B} = \frac{5}{5} = 1$

P(B,1) = $\frac{n_{B1}}{n_1} = \frac{5}{8} = 0.625$

F(B,1) = $\frac{2*1*0.625}{1+0.625} = 0.77$

cluster2 = {p1,p2, p4,p5 }

class A:

R(A,2) = $\frac{n_{A2}}{n_A} = \frac{2}{3} = 0.667$

P(A,2) = $\frac{n_{A2}}{n_2} = \frac{2}{4} = 0.5$

F(A,2) =0.57

class B:

R(B,2) = $\frac{n_{B2}}{n_B} = \frac{2}{5} = 0.4$

P(B,2) = $\frac{n_{B2}}{n_2} = \frac{2}{4} = 0.5$

F(B,2) = 0.44

cluster3 = {p3,p6, p7,p8 }

class A:

R(A,3) = $\frac{n_{A3}}{n_A} = \frac{1}{3} = 0.333$

P(A,3) = $\frac{n_{A3}}{n_3} = \frac{1}{4} = 0.25$

F(A,3) =0.29

class B:

R(B,3) = $\frac{n_{B3}}{n_B} = \frac{3}{5} = 0.6$

P(B,3) = $\frac{n_{B3}}{n_3} = \frac{3}{4} = 0.75$

F(B,3) = 0.67

cluster4 = {p1,p2 }

class A:

R(A,4) = $\frac{n_{A4}}{n_A} = \frac{2}{3} = 0.667$

P(A,4) = $\frac{n_{A4}}{n_4} = \frac{2}{2} = 1$

F(A,4) =0.8

class B:

R(B,4) = $\frac{n_{B4}}{n_B} = \frac{0}{5} = 0$

P(B,4) = $\frac{n_{B4}}{n_4} = \frac{0}{2} = 0$

F(B,4) = 0

cluster5 = {p4,p5 }

class A:

R(A,5) = 0

P(A,5) = 0

F(A,5) =0

class B:

R(B,5) = $\frac{n_{B5}}{n_B} = \frac{2}{5} = 0.4$

P(B,5) = $\frac{n_{B5}}{n_5} = \frac{2}{2} = 1$

F(B,5) = 0.57

cluster6 = {p3,p6 }

class A:

R(A,6) = $\frac{n_{A6}}{n_A} = \frac{1}{3} = 0.33$

P(A,6) = $\frac{n_{A6}}{n_6} = \frac{1}{2} = 0.5$

F(A,6) =0.4

class B:

R(B,6) = $\frac{n_{B6}}{n_B} = \frac{1}{5} = 0.2$

P(B,6) = $\frac{n_{B6}}{n_6} = \frac{1}{2} = 0.5$

F(B,6) = 0.29

cluster7 = {p7,p8 }

class A:

R(A,7) = $\frac{n_{A7}}{n_A} = \frac{0}{3} = 0$

P(A,7) = $\frac{n_{A7}}{n_7} = \frac{0}{2} = 0$

F(A,7) =0

class B:

R(B,7) = $\frac{n_{B7}}{n_B} = \frac{2}{5} = 0.4$

P(B,7) = $\frac{n_{B7}}{n_7} = \frac{2}{2} = 1$

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

overall clustering:

جواب

F = 3/8 * F(A) + 5/8 *F(B) = 0.78

سوال 2

این امکان وجود دارد که همه مقادیر k ممکن تنها منجر به یک خوشه غیر خالی شوند زیرا داده ها ذاتاً برای خوشه بندی مناسب نیستند. به عنوان مثال، اگر مجموعه داده شامل 100 رکورد باشد که تقریباً یکسان هستند و تغییرات بسیار کمی دارند، ممکن است K-means نتواند هیچ خوشه معنی‌داری را شناسایی کند.

در چنین مواردی، استفاده از نسخه افزایشی K-means نیز نتیجه متفاوتی به همراه نخواهد داشت زیرا از همان داده ها استفاده می شود. نسخه افزایشی K-means مراکز خوشه را به طور مکرر به روز می کند، اما اگر ساختار واضحی در داده ها وجود نداشته باشد، این به روز رسانی تکراری نتیجه خوشه بندی را به طور قابل توجهی تغییر نمی دهد.

همچنین ممکن است انتخاب تصادفی اولیه مراکز خوشه در K-means و نسخه افزایشی الگوریتم K-means به دلیل ماهیت داده ها نتایج مشابهی ایجاد کند. به عنوان مثال، اگر نقاط داده در یک منطقه از فضای ویژگی به طور نزدیک به یکدیگر خوشه بندی شوند، هر دو الگوریتم K-means و K-means افزایشی ممکن است در نهایت مجموعه واحدی از مراکز خوشه را در هر تکرار انتخاب کنند.

برای اطمینان از اینکه هیچ ساختار خوشه‌بندی بالقوه‌ای را در داده‌ها از دست نمی‌دهیم، ممکن است لازم باشد الگوریتم‌ها یا تکنیک‌های خوشه‌بندی دیگری را امتحان کنیم که برای شناسایی الگوها یا ساختارهای خاص در داده‌ها مناسب‌تر هستند. علاوه بر این، تکنیک های visualization نیز می‌تواند بینشی در مورد ساختار زیربنایی داده‌ها ارائه دهد.

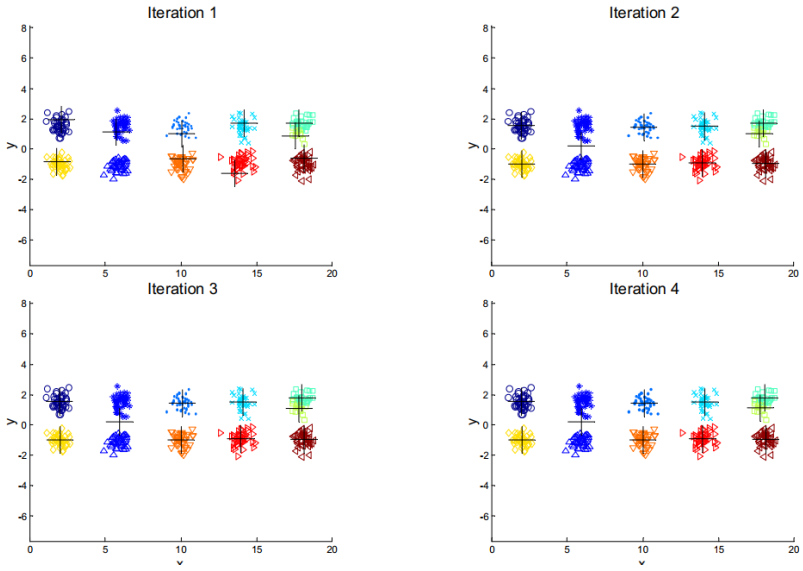
خوشه‌بندی Single Link و DBSCAN نمونه‌هایی از الگوریتم‌های خوشه‌بندی هستند که می‌توانند مجموعه‌های داده با چگالی غیریکنواخت یا خوشه‌هایی با اشکال مختلف را مدیریت کنند. این الگوریتم‌ها نیازی به تعیین تعداد خوشه‌ها از قبل ندارند، که می‌تواند در هنگام برخورد با ساختارهای داده ناشناخته یک مزیت باشد.

خوشه‌بندی Single Link برای مجموعه‌های داده‌ای که شامل خوشه‌های با شکل نامنظم یا دراز هستند، مناسب است، زیرا می‌تواند ارتباط درون و بین خوشه‌ها را ثبت کند

در شرایطی که K-means به دلیل ساختارها و الگوهای پیچیده داده، خوشه‌های معنی‌دار را شناسایی نمی‌کند، هم خوشه‌بندی Single Link و هم DBSCAN می‌توانند جایگزین‌های موثری برای کشف ساختار در داده‌ها باشند. با این حال، مهم است که در نظر داشته باشیم که این الگوریتم‌ها با پارامترها و الزامات تنظیم خاص خود ارائه می‌شوند و ممکن است همیشه نتایج بهینه را ارائه نکنند.

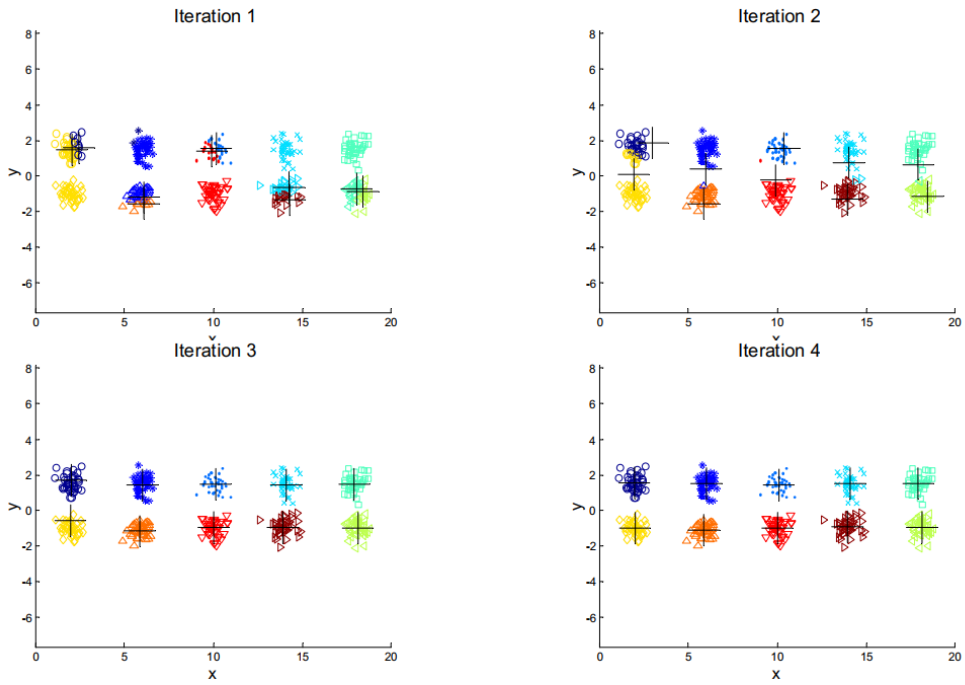
سوال 3

کلاسترینگ با برخی از جفت‌های خوشه شروع می‌شود که دارای سه مرکز اولیه هستند، در حالی که برخی دیگر فقط یک مرکز دارند. که باعث تشخیص اشتباه کلاسترها میشود.



درصورتی که اگر نقاط شروع به شکل زیر باشند، کلاسترینگ به درستی انجام میشود.

شروع با دو مرکز اولیه در یک خوشه از هر جفت خوشه که نتیجه حاصل ۱۰ کلاستر است که به درستی تشخیص داده میشود.



فرض کنید یک مجموعه داده با دو خوشه داریم که هر یک شامل 100 نقطه در دو بعد است و مراکز این خوشه ها به شرح زیر است:

- مرکز خوشه 1: (0، 0)

- مرکز خوشه 2: (10، 0)

اگر الگوریتم K-means را با k=2 اجرا کنیم و مراکز اولیه را به طور تصادفی انتخاب کنیم، ممکن است در نهایت هر دو مرکز اولیه به یک خوشه نسبت داده شوند. به عنوان مثال، فرض کنید که مراکز اولیه به صورت زیر انتخاب می شوند:

- مرکز 1: (1، 0)

- مرکز 2: (5، 0)

هنگامی که ما شروع به تکرار الگوریتم K-means می کنیم، این احتمال وجود دارد که مرکز 1 به خوشه 1 و مرکز 2 به خوشه 2 اختصاص یابد.

با این حال، هنوز این شانس وجود دارد که هر دو مرکز اولیه به خوشه 1 یا خوشه 2 اختصاص داده شوند، که منجر به خوشه بندی ضعیف می شود.

برای جلوگیری از این مشکل، می‌توانیم الگوریتم K-means را با مراکز اولیه تصادفی مختلف چندین بار اجرا کنیم و خوشه‌بندی نهایی را انتخاب کنیم که کمترین مجموع مربع فاصله بین نقاط و مراکز خوشه‌ای اختصاص داده شده را داشته باشد. با این حال، حتی با اجراهای متعدد، ممکن است همچنان با خوشه بندی ضعیفی مواجه شویم. به عنوان مثال، در یک اجرا، هر دو مرکز اولیه ممکن است به خوشه 1 و در اجرای دیگر، هر دو مرکز اولیه ممکن است به خوشه 2 اختصاص داده شوند. در چنین مواردی، ممکن است نیاز به استفاده از الگوریتم ها یا تکنیک های دیگر خوشه بندی برای به دست آوردن نتیجه بهتر داشته باشیم.

سوال 4

a.

الگوریتم k-medoids در حضور نویز و نقاط پرت (outliers) از k-means قوی تر است، زیرا یک medoid کمتر از میانگین تحت تأثیر مقادیر پرت یا سایر مقادیر شدید قرار می گیرد. با این حال، پردازش آن هزینه بیشتری نسبت به روش k-means دارد.

b.

هر دو k-means و k-medoid خوشه بندی مبتنی بر پارتیشن بندی را انجام می دهند. مزیت چنین رویکردهای پارتیشن بندی این است که می توانند مراحل خوشه بندی قبلی را (با جابجایی مکرر) خنثی کنند(undo)، برخلاف روش های سلسله مراتبی، که نمی توانند پس از اجرای تقسیم یا ادغام، تنظیمات را انجام دهند. این ضعف روش‌های سلسله مراتبی می‌تواند کیفیت خوشه‌بندی حاصل از آن‌ها را تحت تاثیر قراردهد.

روش‌های مبتنی بر پارتیشن‌بندی برای پیدا کردن خوشه‌های کروی شکل به خوبی کار می‌کنند. کیفیت خوشه بندی به دست آمده، به طور کلی، برای پایگاه های داده کوچک تا متوسط خوب است. نیاز آنها به دانستن تعداد خوشه ها از قبل می تواند به عنوان یک نقطه ضعف در نظر گرفته شود. روش های خوشه بندی سلسله مراتبی می توانند تعداد خوشه ها را به صورت خودکار تعیین کنند. با این حال، آنها مقیاس بندی دشواری دارند زیرا هر تصمیم برای ادغام یا تقسیم ممکن است به بررسی و ارزیابی تعداد زیادی از ابجکت ها یا خوشه ها نیاز داشته باشد.

روش‌های سلسله مراتبی، با این حال، می‌توانند با سایر رویکردهای خوشه‌بندی، برای خوشه‌بندی بهبودیافته، مانند BIRCH، ROCK، و Chameleon ادغام شوند.