به نام خدا

سری چهارم تمارین درس داده کاوی

استاد درس:

دکتر مریم امیر مزلقانی

نیم سال اول ۱۴۰۲-۱۴۰۳

راه ارتباطی:

Aut.DataMining.Fall@gmail.com

توضیحات:

۱. این تمرین شامل دو بخش عملی و تئوری هست و پاسخ به هر دو بخش الزامی است.

۲. تمرین عملی در قالب یک نوت بوک آماده شده است و دیتای لازم برای این تمرین در پوشه *Practical* موجود است.

۳. در نوت بوک تمرین عملی حتما هر خواسته را در بلوک مربوط به خودش انجام دهید.

۴. حین حل تمرین عملی شما نیازمند به ساخت ۴ فایل جدید هستید (۱ فایل csv و ۳ فایل png)، به نحوه نامگذاری این فایل ها دقت کنید.

۵. **ملاک اصلی انجام تمارین عملی، گزارش است** و ارسال کد بدون pdf گزارش فاقد ارزش است. لذا برای این بخش یک فایل گزارش تهیه کنید و در آن برای هر بخش از تمرین عملی، توضیحات مربوط به آن را ذکر کنید.

۶. خوانا و مرتب بودن پاسخ های شما در نمره‌تان تاثیر مثبت خواهد داشت.

۷. مطابق قوانین دانشگاه هر نوع کپی برداری و اشتراک کار دانشجویان غیرمجاز بوده و برخورد خواهد شد. استفاده از کدها و توضیحات اینترنت به منظور یادگیری **الزاما با ذکر منبع** بلامانع است.

۸. فایل های ایجاد شده در تمرین عملی + نوت بوک + گزارش + پاسخ تمارین تئوری را به صورت زیپ در آورده و با فرمت **StudentID_DM01.zip** در سامانه کورسز آپلود نمایید.

۹. شایان ذکراست هر روز تاخیر باعث کسر **۲۰ ٪ نمره** خواهد شد.

بخش تئوری:

سوال ۱.

مجموعه داده‌ای دارای ۵ تراکنش است که در جدول زیر آمده‌اند. با در نظر گرفتن آستانه پشتیبانی برابر با ۶۰٪ و آستانه اطمینان ۸۰٪ به سوالات زیر پاسخ دهید.

TID	*Item_bought*
T100	*{M, O, N, K, E, Y}*
T200	*{D, O, N, K, E, Y}*
T300	*{M, A, K, E}*
T400	*{M, U, C, K, Y}*
T500	*{C, O, O, K, I, K}*

الف) الگوریتم Apriori را بر روی تراکنش های داده شده اجرا کنید. تمامی مراحل تولید آیتم های کاندید را نشان دهید و در نهایت مجموعه آیتم های پرتکرار را بدست آورید.

ب) تمامی قواعد انجمنی قابل تولید از مجموعه آیتم‌ها را نوشته، آنهایی که مطمئن هستند را مشخص کرده و براساس میزان اطمینان مرتب کنید.

ج) با استفاده از الگوریتم FP-Growth مجموعه آیتم های پرتکرار را بدست آورید.

د) بهینگی الگوریتم های FP-Growth و Apriori را با هم مقایسه کنید.

سوال ۲.

مختصات ۱۵ نقطه به شما داده شده است. با استفاده از الگوریتم K-means و فاصله اقلیدسی، نقاط داده شده را در ۳ دسته خوشه بندی کنید. در ابتدا نقاط A2 و A7 و A15 را به عنوان مراکز دسته ها در نظر بگیرید و الگوریتم را تا جایی ادامه دهید که مراکز نهایی بدست آیند.

A1 (2, 10) – A2 (2, 6) – A3 (11, 11) – A4 (6, 9) – A5 (6, 4) – A6 (1, 2) – A7 (5, 10) A8 (4, 9) – A9 (10, 12) – A10 (7, 5) – A11 (9, 11) – A12 (4, 6) – A13 (3, 10) – A14 (3, 8) – A15 (6, 11)

سوال ۳.

مجموعه ای از برچسب خوشه ها و ماتریس مشابهت در جداول زیر داده شده است. مقدار correlation بین ماتریس مشابهت و ماتریس incidence را محاسبه کنید.

Point	*P1*	*P2*	*P3*	*P4*
P1	*1*	*0.8*	*0.65*	*0.55*
P2	*0.8*	*1*	*0.7*	*0.6*
P3	*0.65*	*0.7*	*1*	*0.9*
P4	*0.55*	*0.6*	*0.9*	*1*

Point	*Cluster Label*
P1	*1*
P2	*1*
P3	*2*
P4	*2*

سوال ۴.

الف) در مورد kmeans افزایشی تحقیق کنید.

ب) فرض کنید یک مجموعه داده با ۲۵۰ رکورد به شما داده میشود و از شما میخواهیم که دادهها رو خوشه بندی کنید. شما در ابتدا از kmeans استفاده میکنید اما برای تمامی مقادیر k بین ۱ تا ۲۵۰ الگوریتم مورد استفاده تنها یک خوشه غیر خالی برمیگرداند، سپس از شما میخواهیم از kmeans افزایشی استفاده کنید ولی باز هم همان نتیجه به دست میآید. اول توضیح دهید این چطور ممکن هست؟ آیا میتوان با DBSCAN چنین مشکلی را حل کرد؟

سوال ۵.

به سوالات زیر پاسخ دهید:

الف) موارد زیر در مورد الگوریتم Dbscan را بررسی کنید؟ درست یا غلط بودن هر یک

را ذکر کنید.

۱. در برابر موارد پرت مقاوم است.
۲. برای اینکه نقاط داده در یک خوشه قرار گیرند، باید در آستانه فاصله تا یک نقطه مرکزی باشند.
۳. دارای مفروضات قوی برای توزیع نقاط داده در فضای داده است.
۴. نیازی به دانستن تعداد خوشه‌ها در این الگوریتم نیست.

سوال ۶.

با توجه به ماتریس شباهت زیر خوشه بندی سلسله مراتبی را با لینک تک و کامل (min, max)را انجام دهید. نتایج خود را با کشیدن یک دندروگرام نمایش دهید.

	p1	p2	p3	p4	p5	p6
p1	0.0000	0.2357	0.2218	0.3688	0.3421	0.2347
p2	0.2357	0.0000	0.1483	0.2042	0.1388	0.2540
p3	0.2218	0.1483	0.0000	0.1513	0.2843	0.1100
p4	0.3688	0.2042	0.1513	0.0000	0.2932	0.2216
p5	0.3421	0.1388	0.2843	0.2932	0.0000	0.3921
p6	0.2347	0.2540	0.1100	0.2216	0.3921	0.0000