داده کاوی

نمونه سوالات امتحان میان ترم

۱. الف) مجموعه نقاط P = {a, b, c, …, j} با این مشخصات را در نظر بگیرید:

Point	X
a	1
b	2
c	3
d	5
e	6
f	9
g	9
h	10
i	10
j	12

می‌خواهیم با استفاده از گسسته‌سازی بر روی صفت X، این داده‌ها را در ۳ bin قرار دهیم. مرزهای binها را در صورتی که از هر کدام از این روش‌ها برای گسسته‌سازی استفاده کنیم تعیین کنید:

i. Equal width binning

ii. Equal depth binning

ب) با توجه به مقادیر صفت X داده شده در بخش (الف)، به سوالات زیر پاسخ دهید:

i. با استفاده از روش هنجارسازی min-max و با نگاشت به بازه [0.0, 1.0]، عدد ۵ به چه مقداری نگاشته می‌شود؟

ii. با استفاده از روش هنجارسازی z-score، عدد ۵ به چه مقداری نگاشته می‌شود؟ (σ=3.8)

iii. با استفاده از روش هنجارسازی با decimal scaling، عدد ۵ به چه مقداری نگاشته می‌شود؟

۲. مجموعه تراکنش‌های مقابل را در نظر بگیرید. فرض کنید min-sup = 2.

Tid	Items
1	A, B
2	A, B, D
3	B, D, E
4	B, C, D, E
5	A, B, C

الف) با استفاده از الگوریتم Apriori همه itemsetهای مکرر را پیدا کنید. گام‌های الگوریتم را در پاسخ بیاورید.

ب) از بین این itemsetهای مکرر، کدام‌ها itemset مکرر بسته (closed) هستند؟

پ) مجموعه الگوهای ماکزیمم (max-pattern) کدام است؟

علاوه بر اطلاعات تراکنش‌ها، اطلاعات قیمت اقلام نیز به این صورت در دست است. می‌خواهیم محدودیت کاوش را پیچیده‌تر کنیم. در هر یک از قسمت‌های (ت) و (ث)، توضیح دهید آیا می‌توان محدودیت جدید را در هنگام کاوش الگوهای مکرر با استفاده از Apriori وارد کرد؟

Item	Price
A	100
B	200
C	300
D	400
E	500

- اگر بله، چگونه؟ چه گام‌هایی از پاسخ بخش (الف) تغییر می‌کنند؟ نتیجه را بدست آورید.
- اگر خیر چرا؟ چه روش کاوش itemsetهای مکرری می‌شناسید که بتواند این محدودیت را در فرایند کاوش وارد کند؟ این روش را بر روی مجموعه داده‌ها اعمال کنید و نتیجه را بدست آورید.

ت) محدودیت جدید: min-sup = 2 ∧ min(I.price) <=200

ث) محدودیت جدید: min-sup = 2 ∧ avg(I.price) >= 300

۳. دانشگاه تهران می‌خواهد یک پایگاه داده تحلیلی (data warehouse) برای نگهداری سابقه دانشجویان با این اطلاعات ایجاد کند: دانشجو، رشته تحصیلی، درس، دانشکده و نمره و می‌خواهد امکان محاسبه معدل دانشجو و معدل یک رشته تحصیلی را داشته باشد.

الف) شمای ستاره‌ای را رسم کنید. کلیه فرضیات خود درباره سطوح هر بعد و معیارهای سنجش را بیان کنید.

ب) می‌خواهیم از cuboid پایه شروع کنیم و ۱۰ دانشجوی برتر هر دانشکده در پردیس دانشکده‌های فنی دانشگاه تهران را بر پایه معدل در پاییز ۱۳۹۹ پیدا کنیم. با توجه به طراحی شما از چه اعمال OLAP ای برای این جستجو باید استفاده کرد؟

۴. فرض کنید cuboid پایه یک data cube فقط دو سلول دارد: (a1, a2,…,a20) و (b1,b2,…,b20) که در آن ai = bi اگر i mod 3 = 0 و در غیر این صورت ai ≠ bi.

الف) در این data cube چند aggregate cell غیرتهی داریم؟

ب) در این data cube چند aggregate cell بسته داریم؟

پ) اگر شرط minimum support = 2 را اضافه کنیم، تعداد aggregate cell های غیرتهی چقدر خواهد بود؟

۵. جدول زیر اطلاعات تراکنش‌هایی که شامل خرید شیر و نان هستند را از بین کل تراکنش‌ها نشان می‌دهد:

-	Milk	Not Milk	Sum (row)
Bread	100	800	900
Not Bread	200	8900	9100
Sum (col)	300	9700	10000

با توجه به این اطلاعات دو معیار lift و all-confidence را برای خریدن شیر و نان محاسبه کنید. آیا خریدن شیر و نان با هم همبستگی دارند؟ توضیح دهید.

$$lift(A,B) = \frac{P(A \cup B)}{P(A)P(B)}, all\text{-}confidence(A,B) = \min\big(P(A|B), P(B|A)\big)$$

موفق باشید.