«به نام خدا»	
لطفاً پاسخ هر سؤال را در محل مشخص‌شده بنویسید. زمان پاسخ‌گویی: ۱۲۰ دقیقه در سؤالات محاسباتی، نیازی به انجام محاسبات نهایی نیست	دانشگاه صنعتی امیرکبیر (پلی‌تکنیک تهران) امتحان پایان ترم داده‌کاوی- نیمسال دوم ۱۴۰۲-۱۴۰۱
نام و نام خانوادگی:	شماره‌ی دانشجویی:

نمره سؤال ۱	نمره سؤال ۲	نمره سؤال ۳	نمره سؤال ۴	نمره سؤال ۵	نمره سؤال ۶	جمع نمرات

۱. (۳ نمره) شرکتی قصد دارد از بین تعداد زیادی ایمیل که دریافت می‌کند، ایمیل‌های اسپم را از غیراسپم شناسایی کند. بدین منظور **1000** ایمیل را بررسی و آن‌ها را به دو دسته اسپم و غیراسپم تقسیم کرده است. از این **1000** ایمیل، **300** ایمیل در کلاس اسپم قرار گرفته‌اند.

الف) بر اساس بررسی‌های انجام‌گرفته، کلمه **buy** در **75** ایمیل آمده که از این تعداد، **70** ایمیل، اسپم بوده است. احتمال این‌که ایمیلی حاوی کلمه **buy**، اسپم باشد، چقدر است؟

پاسخ. (۱ نمره)

قرارداد: Spam: S, buy:b, computer:c, won:w, faculty:f, meeting:m

$$P(S|b) = \frac{P(S \cap b)}{P(b)} = \frac{70}{75},$$

ب) فرض کنید احتمال این‌که یک ایمیل اسپم شامل هریک از لغات **computer**، **won**، **faculty** و **meeting** باشد، به‌ترتیب برابر با **0.1**، **0.85**، **0.01** و **0.05** است. یک ایمیل در کلاس اسپم، با چه احتمالی حاوی سه لغت **computer**، **buy** و **meeting** بوده، ولی شامل **won** و **faculty** نمی‌باشد؟

پاسخ. (۱ نمره)

$$P(b|S) = \frac{P(b)P(S|b)}{P(S)} = \frac{(0.075)\left(\frac{70}{75}\right)}{0.3} = \frac{0.7}{3},$$

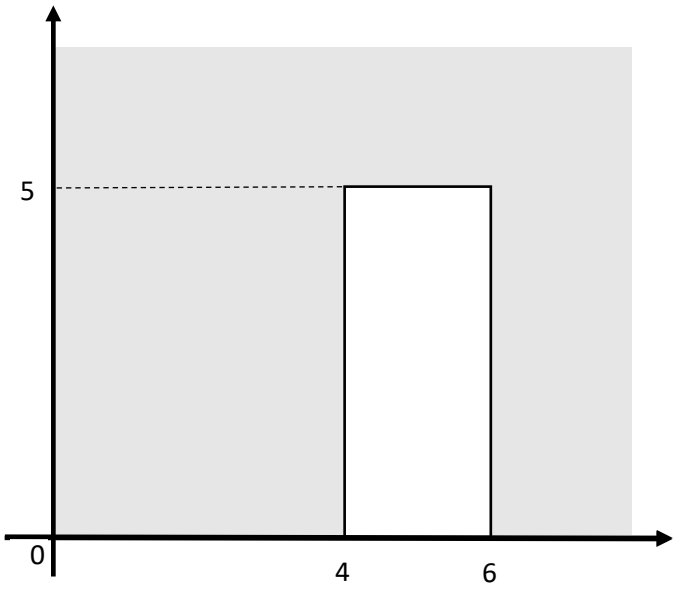
$$P(c, b, m, \sim w, \sim f|S) = P(c|S)P(b|S)P(m|S)P(\sim w|S)P(\sim f|S) = (0.1)\left(\frac{0.7}{3}\right)(0.05)(1-0.85)(1-0.01) = a,$$

ج) اگر بدانیم هر ایمیل با احتمال **0.02** شامل همان ترکیب لغات قسمت (ب) است، آن‌گاه با چه احتمالی یک ایمیل با این ترکیب لغات با کلاس نامشخص، اسپم خواهد بود؟

پاسخ. (۱ نمره)

$$P(S|c, b, m, \sim w, \sim f) = \frac{P(S)P(c, b, m, \sim w, \sim f|S)}{P(c, b, m, \sim w, \sim f)} = \frac{(0.3)(a)}{0.02}$$

۲. (۴ نمره) یک شبکه عصبی با مشخص‌کردن کامل وزن‌ها و بایاس‌ها ارائه دهید که نقاط ربع اول مختصات را به‌صورت زیر در دو کلاس سفید و رنگی دسته‌بندی کند (ورودی شبکه $\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \in \mathbb{R}^2$ می‌باشد که $x_1 \geq 0$ و $x_2 \geq 0$)



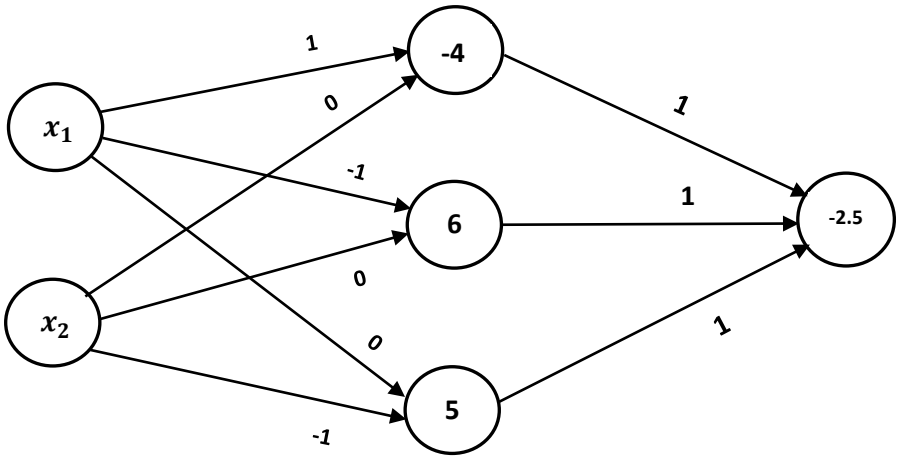
پاسخ) ناحیه رنگی را با برچسب **0** و ناحیه سفیدرنگ را با برچسب **1** درنظر می‌گیریم. معادله هریک از خطوط را به‌دست می‌آوریم

$$x_1 - 4 = 0,$$
$$x_1 - 6 = 0,$$
$$x_2 - 5 = 0,$$

برای این‌که نقطه‌ای در ناحیه سفیدرنگ قرار بگیرد، باید در سه شرط زیر صدق کند:

$$x_1 - 4 \geq 0, \quad -x_1 + 6 \geq 0, \quad -x_2 + 5 \geq 0$$

شبکه عصبی پرسپترون را با یک لایه مخفی درنظر گرفته که تعداد نورون‌های لایه مخفی برابر با تعداد خطوط یعنی سه است. تابع فعال‌ساز در هر نود را نیز تابع پله‌ای قرار می‌دهیم که اگر ورودی به آن از صفر بزرگتر باشد، ***1*** و در غیر اینصورت ***0*** را نظیر می‌کند. بایاس ورودی به هر نورون نیز داخل آن نوشته شده است.



۳. (۳ نمره) ماتریس درهم‌ریختگی زیر مربوط به خروجی یک روش دسته‌بندی با چهار کلاس A، B، C و D می‌باشد. معیار ارزیابی F_1 را برای این روش نسبت به کلاس B محاسبه کنید.

		Actual Classes			
		A	B	C	A
Predicted Classes	A	50	3	0	0
	B	26	8	0	1
	C	20	2	4	0
	D	12	0	0	1

پاسخ.

$$TP_B = 8,\quad FP_B = 27,\quad TN_B = 87,\quad FN_B = 5, \qquad \text{(2 نمره)}$$

$$Recall_B = \frac{TP_B}{TP_B + FN_B} = \frac{8}{8+5} = \frac{8}{13},$$

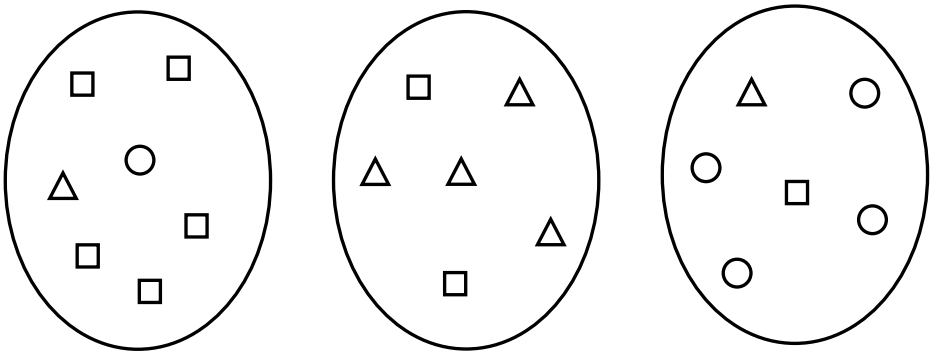
$$Precision_B = \frac{TP_B}{TP_B + FP_B} = \frac{8}{8+27} = \frac{8}{35}$$

$$F_1 = \frac{2\left(\frac{8}{13}\right)\left(\frac{8}{35}\right)}{\frac{8}{13}+\frac{8}{35}}$$

۱ نمره

۴. (۳ نمره) الف) مجموعه داده‌ای با دو برچسب مثبت و منفی داریم که در آن تعداد داده‌های با برچسب منفی با تعداد داده‌های با برچسب مثبت برابر است. اگر این داده‌ها توسط الگوریتمی به دو خوشه افراز شوند به‌طوری‌که $\frac{1}{4}$ داده‌ها در یک خوشه و $\frac{3}{4}$ داده‌ها در خوشه دیگر قرار بگیرند، کم‌ترین و بیش‌ترین مقداری که خلوص (purity) این خوشه‌بندی می‌تواند داشته باشد چیست؟

ب) داده‌های زیر که هرکدام متعلق به یکی از کلاس‌های دایره، مربع یا مثلث است را با یکی از روش‌های خوشه‌بندی به صورت زیر در سه خوشه تفکیک کرده‌ایم:



آنتروپی خوشه‌بندی فوق را محاسبه کنید.

پاسخ.

الف) (۲ نمره) فرض کنیم تعداد داده‌ها برابر با $8k$ باشد و خوشه C1 را با $\frac{1}{4}$ داده‌ها یعنی $2k$ و خوشه C2 را با $\frac{3}{4}$ داده‌ها یعنی $6k$ درنظر می‌گیریم. کم‌ترین میزان خلوص زمانی حاصل می‌شود که در هر خوشه نیمی از داده‌ها برچسب مثبت و نیمی برچسب منفی داشته باشند:

$$purity = \frac{2k}{8k}\max\left\{\frac{k}{2k},\frac{k}{2k}\right\} + \frac{6k}{8k}\max\left\{\frac{3k}{6k},\frac{3k}{6k}\right\} = \frac{1}{2}$$

و بیش‌ترین مقدار خلوص زمانی به‌دست می‌آید که داده‌های خوشه C1 همگی هم‌برچسب بوده و سایر داده‌ها در C2 قرار بگیرند:

$$purity = \frac{2k}{8k}\max\left\{\frac{2k}{2k},0\right\} + \frac{6k}{8k}\max\left\{\frac{2k}{6k},\frac{4k}{6k}\right\} = \frac{1}{4} + \frac{2}{4} = \frac{3}{4}$$

ب) (۱ نمره)

$$e_1 = -\frac{5}{7}\log\frac{5}{7} - \frac{1}{7}\log\frac{1}{7} - \frac{1}{7}\log\frac{1}{7}$$

$$e_2 = -\frac{2}{6}\log\frac{2}{6} - \frac{4}{6}\log\frac{4}{6}$$

$$e_3 = -\frac{1}{6}\log\frac{1}{6} - \frac{1}{6}\log\frac{1}{6} - \frac{4}{6}\log\frac{4}{6}$$

$$entropy = \frac{7}{19}e_1 + \frac{6}{19}e_2 + \frac{6}{19}e_3$$

۵. (۴ نمره) مجموعه نقاط $\{1, 4, 9, 16, 25\}$ در فضای یک‌بعدی را با الگوریتم سلسله‌مراتبی **Complete linkage** با رویکرد تجمعی به دو خوشه افراز کنید.

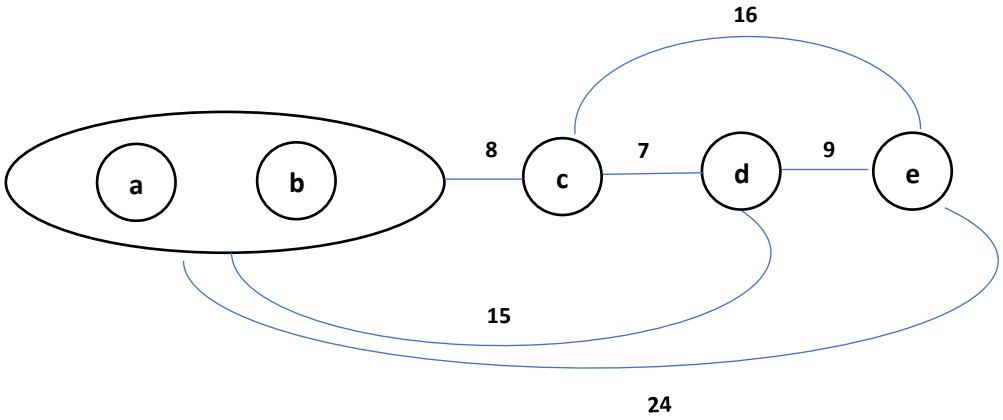
پاسخ) قرار می‌دهیم:

$$a := 1, \quad b := 4, \quad c := 9, \quad d := 16, \quad e := 25$$

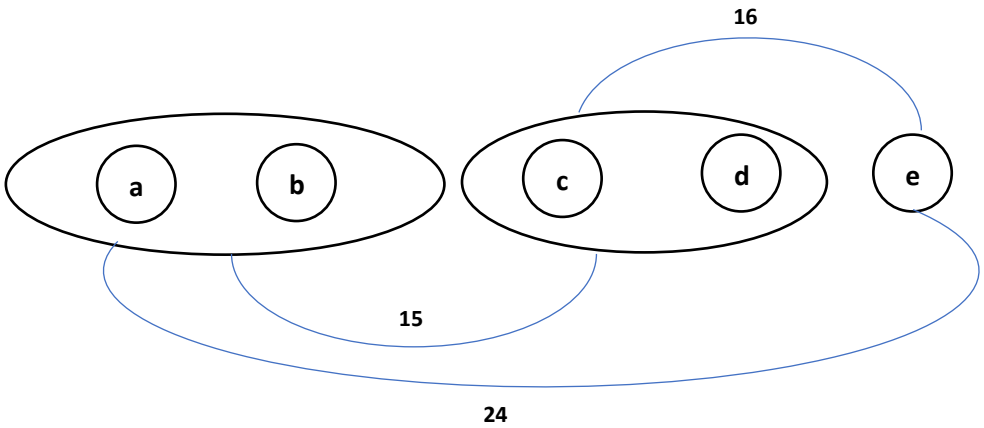
ماتریس فاصله این نقاط را به‌دست می‌آوریم: (۱ نمره)

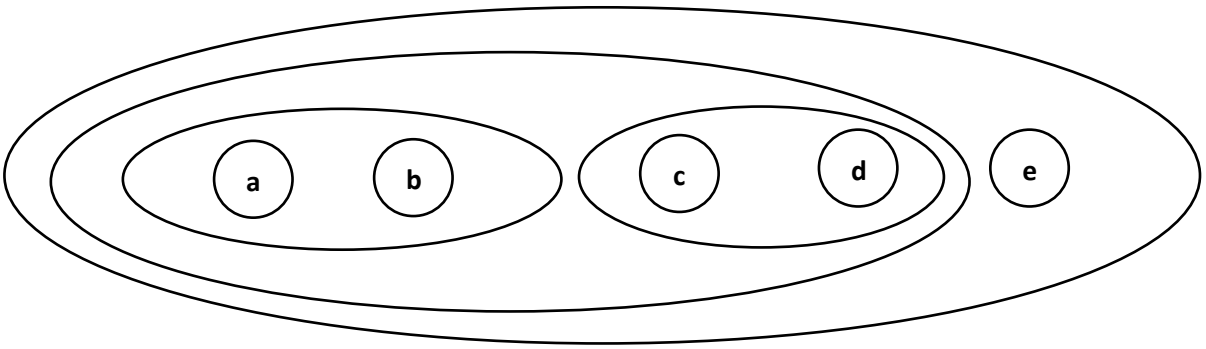
$$D = \begin{bmatrix} 0 & 3 & 8 & 15 & 24 \\ 3 & 0 & 5 & 12 & 21 \\ 8 & 5 & 0 & 7 & 16 \\ 15 & 12 & 7 & 0 & 9 \\ 24 & 21 & 16 & 9 & 0 \end{bmatrix}$$

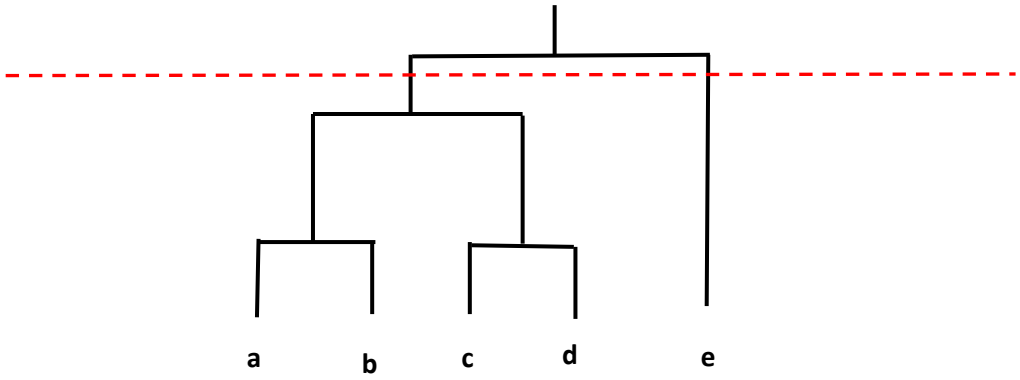
ابتدا هر داده را به عنوان یک خوشه درنظرگرفته و دو خوشه‌ای که کم‌ترین فاصله را دارند یافته و با هم ادغام می‌کنیم و با متر ماکزیمم فاصله، فاصله بین خوشه‌ها را بروزرسانی می‌کنیم: (۰.۷۵ نمره)



همین روند را ادامه می‌دهیم و مجددا دو خوشه‌ای که با ماکزیمم فاصله، کم‌ترین فاصله را دارند ادغام می‌کنیم: (۰.۷۵ نمره)



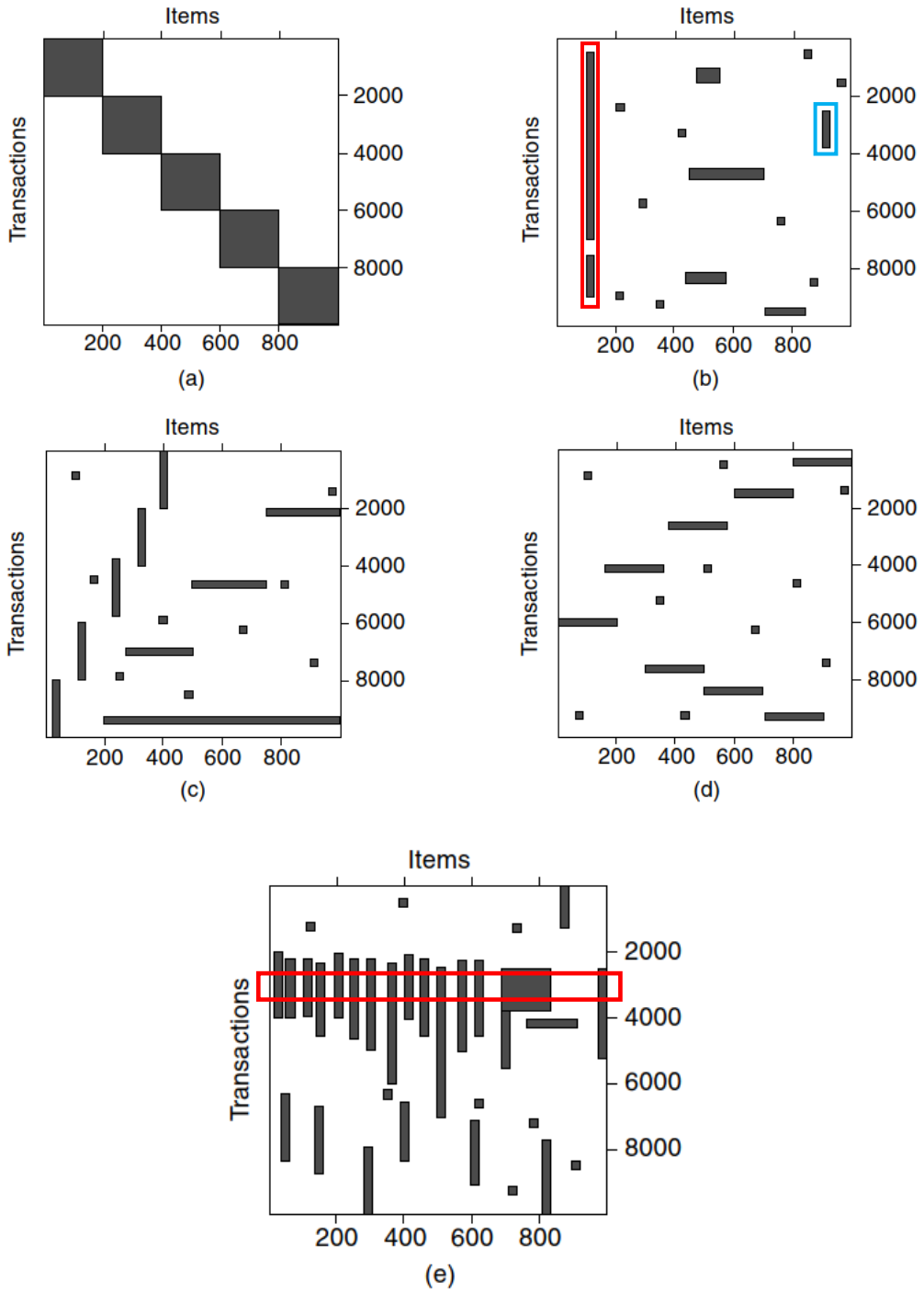




(۱ نمره)

(۰.۵ نمره) افراز به دو خوشه: $\{a, b, c, d\}, \{e\}$

۶. (۳ نمره) مجموعه داده‌های زیر را درنظربگیرید که هرکدام متشکل از **10000** تراکنش و **1000** آیتم می‌باشد. سلول‌های سیاه‌رنگ، نشان‌دهنده وجود آیتم مربوطه و سلول‌های سفیدرنگ، نشان‌دهنده عدم وجود آیتم مربوطه در تراکنش موردنظر است. روش **Apriori** را با $minsup = 10\%$ برای استخراج الگوهای پرتکرار این مجموعه داده‌ها به‌کار گرفته‌ایم.($minsup = 10\%$ بدین مفهوم است که الگویی پرتکرار است که حداقل در **1000** تراکنش ظاهر شده باشد)



(۰.۵ نمره) درک کلی سؤال

الف) کدام‌یک از مجموعه داده فوق، دارای بیش‌ترین تعداد از الگوهای پرتکرار است؟

پاسخ. (۰.۵ نمره) مجموعه داده **e**، زیرا اگر به عنوان مثال مجموعه آیتم‌های پرتکرار داخل مستطیل قرمزرنگ را درنظر بگیریم، یک مجموعه پرتکرار با طول حداقل **500** داریم که با درنظرگرفتن زیرمجموعه‌های ناتهی آن، حداقل $2^{500} - 1$ الگوی پرتکرار می‌توان ساخت، ولی تعداد الگوهای پرتکرار در سایر مجموعه داده‌ها کم‌تر است. به عنوان نمونه در مجموعه داده **a**، تعداد الگوهای پرتکرار برابر با $5 \times (2^{200} - 1)$ است که به مراتب کمتر است.

ب) کدام‌یک از مجموعه داده فوق، دارای کم‌ترین تعداد از الگوهای پرتکرار است؟

پاسخ. (۰.۵ نمره) مجموعه داده **d** که با توجه به **minsup**، الگوی پرتکرار ندارد.

ج) کدام‌یک از مجموعه داده فوق، الگوی پرتکرار با بزرگ‌ترین طول را تولید می‌کند؟

پاسخ. (۰.۵ نمره) مجموعه داده **e**، زیرا با توجه به توضیحات قسمت الف، یک مجموعه پرتکرار با طول حداقل **500** دارد.

د) کدام‌یک از مجموعه داده فوق، الگوهای پرتکرار با ماکزیمم support را تولید می‌کند؟

پاسخ. (۰.۵ نمره) مجموعه داده **b**، زیرا مجموعه الگوی پرتکراری که بین آیتم‌های **0** تا **200** قرار دارد، با توجه به مستطیل قرمزرنگ دارای **support** حداقل **80%** است که از **support** الگوهای پرتکرار سایر مجموعه داده‌ها بیشتر است.

ه) در الگوهای پرتکرار کدام مجموعه داده، هم الگوی پرتکرار با support کمتر از 20% و هم الگوی پرتکرار با support بیشتر از 70% وجود دارد؟

پاسخ. (۰.۵ نمره) مجموعه داده **b**، الگوهای پرتکراری که در داخل مستطیل‌های سبزرنگ و قرمزرنگ قرار گرفته‌اند، به ترتیب **support** کمتر از **20%** و بیشتر از **70%** دارند.