«به نام خدا»	
لطفاً پاسخ هر سؤال را در محل مشخص‌شده بنویسید. زمان پاسخ‌گویی: ۱۲۰ دقیقه	دانشگاه صنعتی امیرکبیر (پلی‌تکنیک تهران) امتحان میان 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	جمع نمرات

۱. (۴ نمره) الف) دو روش برای مدیریت مقادیر از دست‌رفته یک مجموعه داده نام ببرید، هرکدام را به‌طور مختصر توضیح دهید و زمان مناسب استفاده از این روش‌ها را بیان کنید.

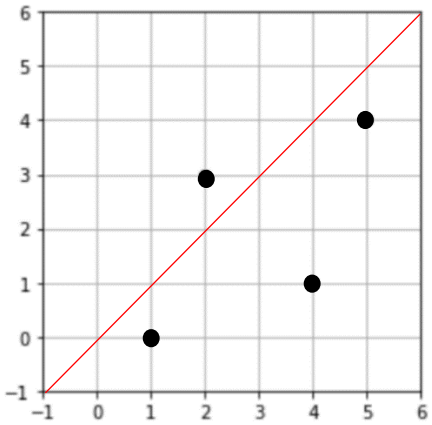
ب) مفهوم نرمال‌سازی داده را با ذکر یک مثال توضیح دهید.

پاسخ. توضیحات مطابق جزوه

الف) (۲ نمره)

ب) (۲ نمره)

۲. (۳ نمره) نقاط زیر در $\mathbb{R}^2$ را درنظربگیرید.



الف) نشان دهید ماتریس کوواریانس این داده‌ها برابر است با

$$S = \frac{1}{3}\begin{bmatrix} 10 & 6 \\ 6 & 10 \end{bmatrix}$$

ب) اگر $u = \begin{bmatrix} \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} \end{bmatrix}$ بردار ویژه یکه متناظر با بزرگ‌ترین مقدار ویژه ماتریس S باشد، آن‌گاه تصویر یک بعدی داده‌های فوق با بیش‌ترین واریانس را بیابید. هم‌چنین روی شکل بالا، خطی که داده‌ها روی آن تصویر شده است را ترسیم کنید.

پاسخ. الف) (۱ نمره) در اینجا چهار داده داریم که هرکدام دو ویژگی دارد، ماتریس کوواریانس برابر است با

$$S = \begin{bmatrix} Cov(z_1, z_1) & Cov(z_1, z_2) \\ Cov(z_2, z_1) & Cov(z_2, z_2) \end{bmatrix}, \qquad Cov(z_i, z_j) = \frac{1}{4-1}\sum_{k=1}^{4}(z_{ki} - \overline{z_I})(z_{kj} - \overline{z_J})$$

که z_{k1} و z_{k2}، مقادیر ویژگی اول و ویژگی دوم برای داده kام می‌باشند. هم‌چنین $\overline{z_1}$ میانگین ویژگی اول و $\overline{z_2}$ میانگین ویژگی دوم برای کل داده‌ها است.

$$X_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix},\ X_2 = \begin{bmatrix} 2 \\ 3 \end{bmatrix},\ X_3 = \begin{bmatrix} 4 \\ 1 \end{bmatrix},\ X_4 = \begin{bmatrix} 5 \\ 4 \end{bmatrix}$$

$$\overline{z_1} = \frac{1+2+4+5}{4} = 3,\ \overline{z_2} = \frac{0+3+1+4}{4} = 2$$

$$Cov(z_1, z_1) = \frac{1}{3}\left[(1-3)^2 + (2-3)^2 + (4-3)^2 + (5-3)^2\right] = \frac{10}{3}$$

$$Cov(z_1, z_2) = Cov(z_2, z_1) = \frac{1}{3}\left[(1-3)(0-2) + (2-3)(3-2) + (4-3)(1-2) + (5-3)(4-2)\right] = \frac{6}{3}$$

$$Cov(z_2, z_2) = \frac{1}{3}\left[(0-2)^2 + (3-2)^2 + (1-2)^2 + (4-2)^2\right] = \frac{10}{3}$$

ب) (۱ نمره) بردار u، مؤلفه اصلی می‌باشد و تصویر یک‌بعدی داده‌ها با بیش‌ترین واریانس برابر خواهد بود با

$$u^T X_1 = \frac{\sqrt{2}}{2},\quad u^T X_2 = \frac{5\sqrt{2}}{2},\quad u^T X_3 = \frac{5\sqrt{2}}{2},\quad u^T X_4 = \frac{9\sqrt{2}}{2},$$

(۱ نمره) خطی که داده‌ها روی آن تصویر می‌شود، خطی است که از مبدأ می‌گذرد و در راستای u می‌باشد، بنابراین خط $y = x$ خط مورد نظر است.

۳. (۵ نمره) جدول داده‌های زیر را درنظربگیرید که پاس‌شدن در درسی را با توجه به دو ویژگی GPA(Grade Point Average) و مطالعه‌داشتن یا نداشتن بیان می‌کند. با معیار بهره اطلاعاتی، درخت تصمیم این داده‌ها را رسم کنید. در محاسبات خود $\log_2 3$ را برابر با 1.6 درنظر بگیرید.

GPA	Studied	Passed
L	F	F
L	T	T
M	F	F
M	T	T
H	F	T
H	T	T

پاسخ. توجه داریم که $\log_2 3 = 1.6$ فرض شده است.

$$H(X) = -\frac{2}{6}\log_2\frac{2}{6} - \frac{4}{6}\log_2\frac{4}{6} = \frac{14}{15}$$

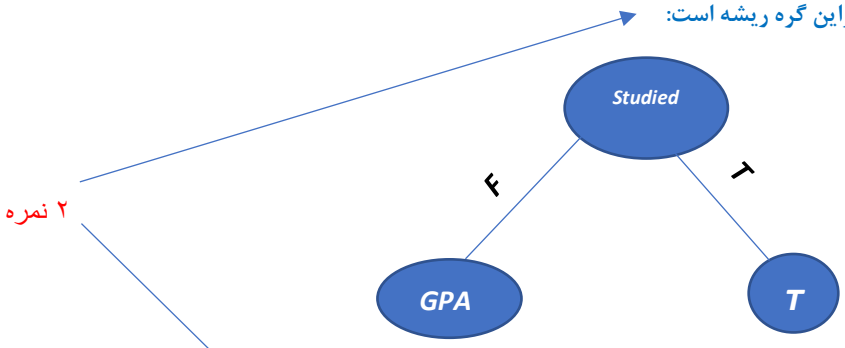
GPA	Passed		Studied	Passed
L	F		F	F
L	T		F	F
M	F		F	T
M	T		T	T
H	T		T	T
H	T		T	T

۲ نمره

$$IG(X; GPA) = H(X) - \sum_{i=1}^{r}\frac{|C_i|}{|X|}H(C_i) = \frac{14}{15} - \frac{2}{6}\left(-\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}\right) - \frac{2}{6}\left(-\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}\right) - \frac{2}{6}(0) = \frac{14}{15} - \frac{4}{6} = \frac{4}{15}$$

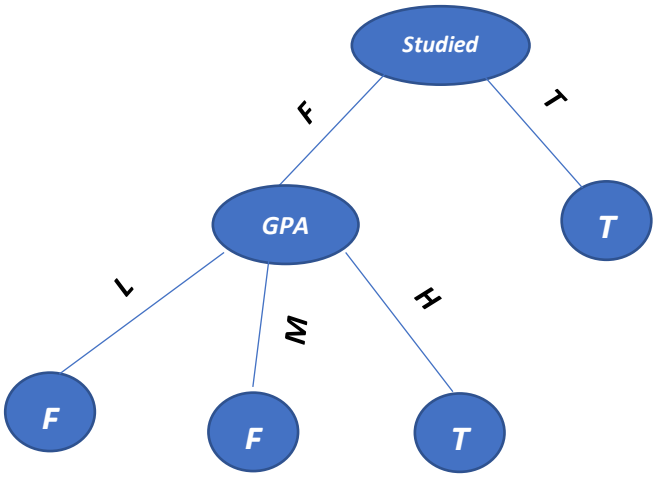
$$IG(X; Studied) = H(X) - \sum_{i=1}^{r}\frac{|C_i|}{|X|}H(C_i) = \frac{14}{15} - \frac{3}{6}\left(-\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3}\right) - \frac{3}{6}(0) = \frac{7}{15}$$

ویژگی ***Studied***، بهره اطلاعاتی بیشتری دارد و بنابراین گره ریشه است:



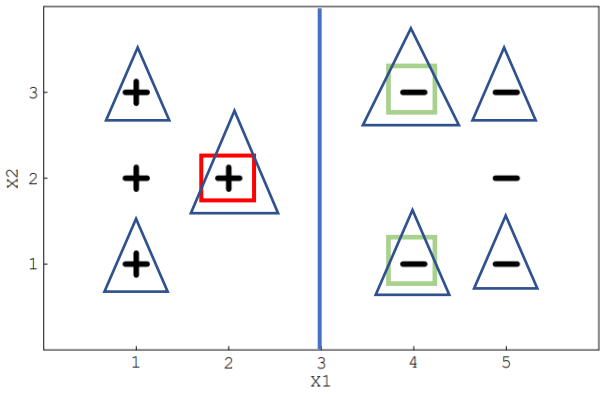
۲ نمره

GPA	Passed		Studied
L	F		F
M	F		F
H	T		F



رسم درخت، ۱ نمره

۴. (۴ نمره) داده‌های زیر را درنظربگیرید.



الف) اگر از **SVM** برای دسته‌بندی داده‌های فوق استفاده کنیم، به نظر شما چه خطی به عنوان جداساز ارائه می‌گردد؟ روی شکل رسم کنید.

ب) حداکثر تعداد داده‌ای که می‌توانیم از مجموعه داده فوق حذف کنیم که در صورت به‌کارگیری مجدد **SVM**، جداساز تغییر نکند، چند داده است؟ پاسخ خود را توضیح دهید.

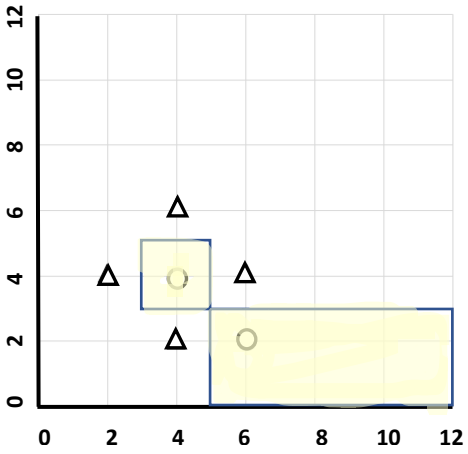
ج) کدام داده‌ها را از مجموعه داده فوق حذف کنیم که در صورت به‌کارگیری مجدد **SVM**، جداساز تغییر کند؟

پاسخ. الف) (۱ نمره) خط آبی رنگ

ب) (۱/۵ نمره) حداکثر ۷ داده است (به‌عنوان مثال داده‌هایی که در مثلث قرار گرفته‌اند)، اگر بیشتر از ۷ داده مثلا ۸ داده حذف شود، فقط یک داده باقی می‌ماند و دسته‌بندی بی‌معنی خواهد بود.

ج) (۱/۵ نمره) مجموعه داده‌های مختلفی را می‌توان در این قسمت درنظر گرفت. به‌عنوان مثال می‌توان تک داده‌ای که در مربع قرمز رنگ قرار گرفته است را درنظرگرفت و یا دو داده‌ای که در مربع‌های سبزرنگ قرار گرفته‌اند را مد نظر قرار داد و حذف کرد.

۵. (۴ نمره) مجموعه داده زیر با دو برچسب ○و △ درنظربگیرید.



الف) دسته‌بندی KNN با $K = 1$ را توضیح دهید و تحلیل کنید با به‌کارگیری آن روی مجموعه داده فوق، مرزهای جداکننده دو کلاس به چه صورت خواهد بود؟

ب) اگر از دسته‌بندی قسمت الف استفاده کنیم، برچسب داده $(8, 1)$ چه خواهد بود؟

پاسخ. الف) (۲ نمره) برای هرداده ○ و هر داده △، عمودمنصف خط واصل بین این دو داده را درنظر می‌گیریم و بدین‌ترتیب ناحیه را به دو زیرناحیه ○ و △ تقسیم می‌کنیم. اگر برای همه داده‌ها این کا را انجام دهیم، آنگاه در شکل فوق نواحی زرد رنگ، برچسب ○و ناحیه سفید برچسب △ارائه می‌دهند.

ب) (۲ نمره) داده $(8, 1)$ در ناحیه زردرنگ است و برچسب آن ○ خواهد بود.