۱. **[درخت تصمیم]** داده‌های زیر برای ساخت یک درخت تصمیم‌گیری استفاده می‌شود تا پیش‌بینی شود که آیا افراد در یک شرکت استخدام می‌شوند (Y) یا نه (N).

درخت تصمیم‌گیری مربوط به این مجموعه داده را با استفاده از الگوریتم ID3 رسم کنید. مراحل و محاسبات را بنویسید و درخت تصمیم‌گیری را در پایان رسم کنید.

نکته: log(3/4)=-0.415 و log(1/4)=-2 و log(2/5)=-1.32 و log (3/5)=-0.737

Student ID	ML grade	GPA	Internship	(Output) Hired?
1	L	H	Y	Y
2	L	L	N	N
3	L	L	Y	N
4	L	L	N	N
5	H	H	Y	Y
6	H	L	Y	Y
7	H	H	N	Y
8	H	L	N	Y

H(Hired) = -5/8 log (5/8) – 3/8 log (3/8) = 0.95

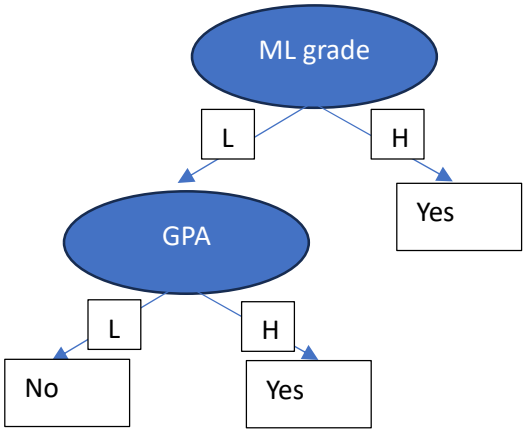
H(Hired | ML grade) =- 4/8 * (3/4 log (3/4) + 1/4 log (1/4)) -4/8 * 0 = 0.40

H(Hired | GPA) = -3/8 * 0 - 5/8 * (2/5 log (2/5) + 3/5 log (3/5)) = 0.60

H(Hired | Internship) = -4/8 * (1/4 log(1/4) +3/4 log(3/4)) + 4/8 * 1 = 0.90

So we select “ML grade” for the root.

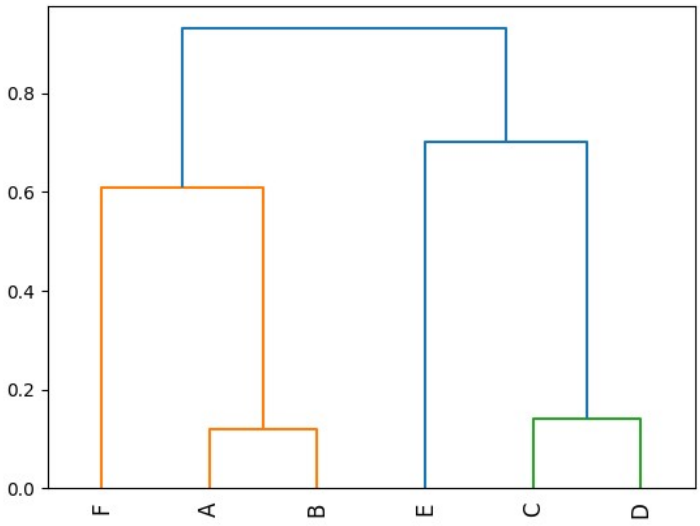
We select “GPA” for the next split (ML grade = L) and we reach pure nodes:



۲. **[خوشه بندی]** جدول روبرو یک ماتریس فاصله برای ۶ داده است.

	A	B	C	D	E	F
A	0					
B	0.12	0				
C	0.51	0.25	0			
D	0.84	0.16	0.14	0		
E	0.28	0.77	0.70	0.45	0	
F	0.34	0.61	0.93	0.20	0.67	0

با استفاده از روش complete linkage، داده‌ها را خوشه‌بندی کنید. Dendrogram نهایی را رسم کنید.



---

۳. **[Naïve Bayes]** مسئله زیر را در نظر بگیرید، که در آن دو کلاس داریم: Tainted و Clean. به همراه سه ویژگی باینری a1,a2,a3 با مقادیر ممکن زیر:

a1 ∈ {on, off}, a2 ∈ {blue,red}, a3 ∈ {light, heavy}

شش داده آموزشی به صورت زیر به ما داده شده اند:

Tainted: (on, blue, light) (off,red, light) (on,red, heavy)
Clean: (off,red, heavy) (off, blue, light) (on, blue, heavy)

الف) احتمالات اولیه (prior probabilities) را برای دو کلاس محاسبه کنید.

P(Y=Tainted) = 3/6, p(Y=clean) = 3/6

ب) یک نمونه جدید (on, red, light) را با استفاده از طبقه‌بندی‌کننده‌ای naïve bayes، طبقه‌بندی کنید. محاسبات را در پاسخنامه بنویسید.

$$P(Y = tainted \mid on, red, light)$$
$$\propto P(tainted) * P(on|tainted) * p(red|tainted) * p(light\ |tainted)$$
$$= \frac{1}{2} * \frac{2}{3} * \frac{2}{3} * \frac{2}{3}$$

$$P(Y = clean \mid on, red, light) \propto P(clean) * P(on|clean) * p(red|clean) * p(light\ |clean)$$
$$= \frac{1}{2} * \frac{1}{3} * \frac{1}{3} * \frac{1}{3}$$

Classified as “Tainted”

---

۴. **[FpGrowth]** مجموعه داده های تراکنشی D را در نظر بگیرید. فرض کنید min supportبرابر با ۴۰٪ است.

آیتم ها به تعداد زیر در این دیتاست خریداری شده اند:

{I1:7, I2:6, I3:5, I4:3, I5:6, I6:4, I7:2, I8:1}

الف) الگوریتم FP-growth را برای تولید مجموعه‌های آیتم پرتکرار در D اعمال کنید. درخت FP را نشان دهید. فقط درخت نهایی را رسم کنید.

ب) فقط برای آیتم I5، Conditional pattern base و آیتم های پرتکرارآن را پیدا کنید.

TID	items_bought
T100	{ I6, I1, I3}
T200	{ I1, I2, I4, I5, I3}
T300	{ I3, I2, I5}
T400	{ I6, I7}
T500	{ I1, I3, I2, I4, I5}
T600	{ I1, I3, I6}
T700	{ I1, I2, I5, I7}
T800	{ I2, I8, I5, I1}
T900	{ I4, I6}
T1000	{ I1, I2, I5 }

---

۵. **[Logistic Regression]**

چهارگزینه ای اول) در کدام یک از موقعیت‌های زیر استفاده از رگرسیون لجستیک مناسب است؟

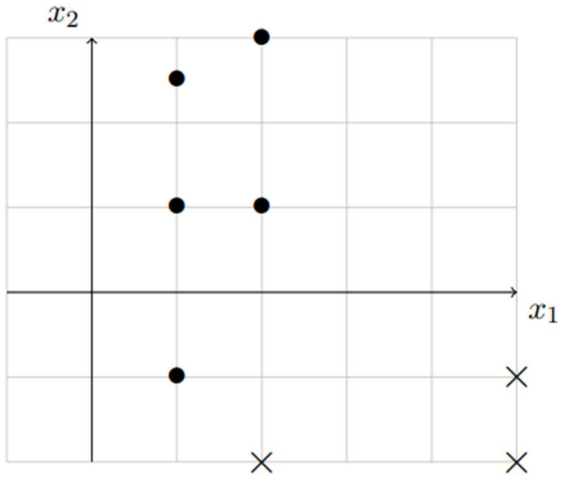
logistic regression is suitable for “classification” tasks

الف) پیش‌بینی اینکه آیا یک تراکنش کارت اعتباری کلاهبردارانه است یا خیر، بر اساس برخی ویژگی‌ها.
ب) پیش‌بینی تعداد خودروهایی که در ساعات اوج از یک تقاطع خاص عبور می‌کنند.
ج) پیش‌بینی درآمد سالانه یک فرد بر اساس تحصیلات و سابقه شغلی او.
د) پیش‌بینی قیمت یک سهام بر اساس داده‌های تاریخی.

چهارگزینه ای دوم) هدف از تابع سیگموید در رگرسیون لجستیک چیست؟

الف) ورودی پیوسته (continuous) را به داده‌های دسته‌ای (categorical) تبدیل می‌کند.
ب) ورودی را استاندارد (standardize) می‌کند تا میانگین صفر و واریانس ۱ داشته باشد.
ج) خروجی را به احتمال تبدیل می‌کند.

۶. **[Validation]** مجموعه داده در زیر رسم شده است، با نقاط با برچسب مثبت به‌صورت نقاط توپر (•) و نقاط با برچسب منفی به‌صورت علامت‌های ضربدر (X):



در صورت تساوی در فاصله، نقطه‌ای را که مختصه x1 کمتری دارد انتخاب کنید و اگر هنوز تساوی وجود داشت، نقطه‌ای را که مختصه x2 کمتری دارد انتخاب کنید.

الف) دقت Leave-One-Out Cross Validation برای الگوریتم ۱-نزدیک‌ترین همسایه (1-NN) روی این مجموعه داده محاسبه کنید.

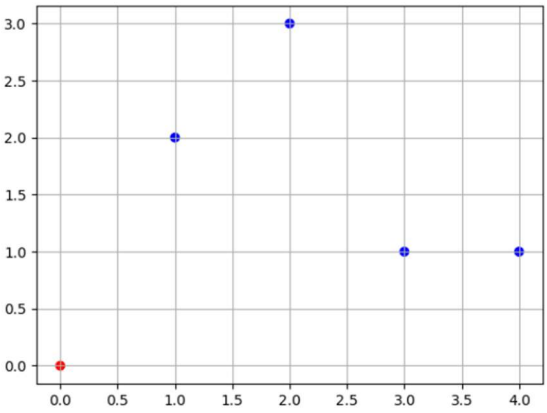
6/8. نقاط (2,-2) و (1,-1) رو اشتباه میکند

ب) دقت Leave-One-Out Cross Validation الگوریتم ۳-نزدیک‌ترین همسایه (3-NN) را روی این مجموعه داده محاسبه کنید.

7/8. نقطه (2,-2) رو اشتباه میکند

---

۷. **[kmeans++]** فرض کنید پنج داده داریم: (0, 0) ، (1, 2)، (2, 3)، (3, 1)، (4, 1). تعداد خوشه‌ها را برابر ۳ درنظر بگیرید. مرکز خوشه اول به صورت تصادفی به عنوان (0,0) انتخاب شده است. این داده‌ها در شکل زیر نشان داده شده‌اند.



الف) احتمال انتخاب هر نقطه داده به عنوان مرکز برای خوشه ۲ چیست؟ (پاسخ باید شامل ۵ احتمال باشد، هرکدام برای هر نقطه داده)

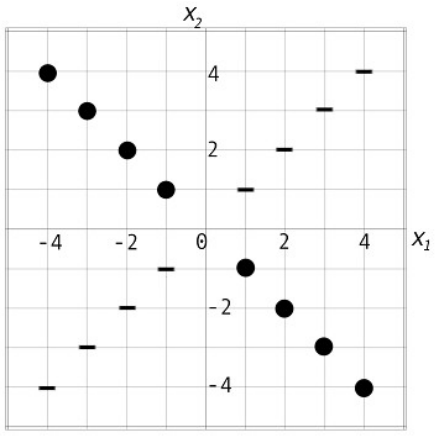
(0, 0): 0
(1, 2): 0.111
(2, 3): 0.289
(3, 1): 0.222
(4, 1): 0.378

(4,1) will be selected with highest probability

ب) فرض کنید مرکز خوشه ۲ به عنوان محتمل‌ترین مرکز که در سوال قبلی محاسبه کرده‌اید، انتخاب شده است. حالا احتمال انتخاب هر نقطه داده به عنوان مرکز برای خوشه ۳ چیست؟ (پاسخ باید شامل ۵ احتمال باشد، هرکدام برای هر نقطه داده)

(0, 0): 0
(1, 2): 0.357
(2, 3): 0.571
(3, 1): 0.071
(4, 1): 0

۸. **[درخت تصمیم]** شما ۱۶ نقطه داده در مجموعه آموزشی خود دارید که در نمودار زیر نشان داده شده‌اند. این نقاط همگی مختصات صحیح دارند؛ برای مثال، یک نقطه در موقعیت (x1, x2) = (۱, ۱) وجود دارد. دو کلاس داریم: دایره‌ها و خطها. هدف شما ساخت یک درخت تصمیم برای انجام طبقه‌بندی است.

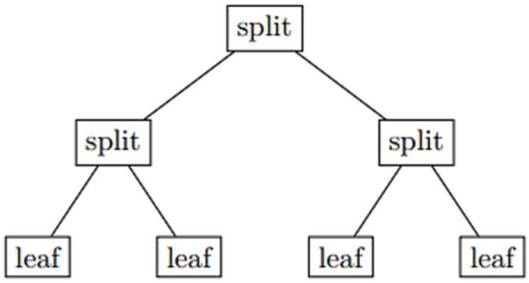


الف) آنتروپی برای کل مجموعه داده‌ها چقدر است؟ 1

ب) آنتروپی میانگین وزنی دو گره فرزند در تقسیم فیچر x1 در موقعیت x1=0 برای این مجموعه داده‌ها چقدر است؟1

ج) آنتروپی میانگین وزنی دو گره فرزند در تقسیم فیچر x1 در موقعیت x1=1.5 برای این مجموعه داده‌ها چقدر است؟1

د) حال هر تقسیم‌بندی ممکن را در بعد j در موقعیت s در نظر بگیرید. کدام‌یک از این تقسیم‌بندی‌ها برای این مجموعه داده‌ها کمترین آنتروپی میانگین وزنی را دارد؟ اگر چندین تقسیم‌بندی با کمترین آنتروپی میانگین وزنی وجود داشته باشد، همه‌ی این تقسیم‌بندی‌ها را توصیف کنید.



هیچ فرقی بین تقسیم بندی ها وجود ندارد. همگی منجر به آنتروپی وزن دار ۱ برای فرزندان میشود. تمام این تقسیم‌ها در هر دو گره فرزند، تعداد یکسانی از هر کلاس خواهند داشت.

ه) آیا روش id3 تضمین می‌کند که درختی با ساختار ترسیم‌شده در روبرو پیدا شود که بالاترین دقت را بر روی داده‌های آموزشی داشته باشد؟ حتماً جواب خود را توجیه کنید.

خیر هیچ تضمینی برای id3 وجود ندارد. چون تمام حالت های split منجر به جواب یکسان میشود هیچ تضمینی وجود ندارد که id3 در x1=0 و سپس در x2=0 تقسیم بندی را انجام دهد.

موفق باشید!!