

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

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

	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

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

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 a_1, a_2, a_3 با مقادیر ممکن زیر:

$$a_1 \in \{on, off\}, a_2 \in \{blue, red\}, a_3 \in \{light, heavy\}$$

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.

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

۴. [FpGrowth] مجموعه داده‌های تراکنشی D را در نظر بگیرید. فرض کنید min support برابر با ۴۰٪ است.

آیتم‌ها به تعداد زیر در این دیتاست خریداری شده اند:

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

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 }

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 آیت‌م‌های پرتکرار آن را پیدا کنید.

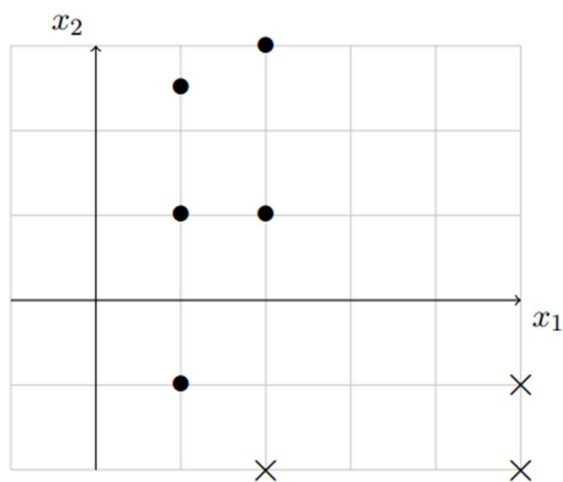
۵. [Logistic Regression]

چهارگزینه‌ای (اول) در کدام یک از موقعیت‌های زیر استفاده از رگرسیون لجستیک مناسب است؟

- 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.

چهارگزینه‌ای (دوم) هدف از تابع سیگموئید 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.



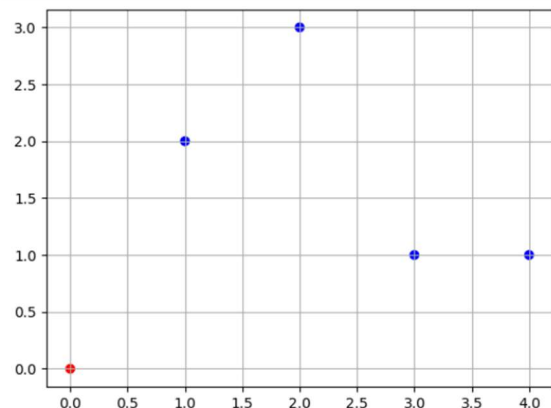
۶. [Validation] مجموعه داده در زیر رسم شده است، با نقاط با برجسب مثبت به‌صورت نقاط توپر (•) و نقاط با برجسب منفی به‌صورت علامت‌های ضربدر (X):

در صورت تساوی در فاصله، نقطه‌ای را که مختصه x_1 کمتری دارد انتخاب کنید و اگر هنوز تساوی وجود داشت، نقطه‌ای را که مختصه x_2 کمتری دارد انتخاب کنید.

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

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

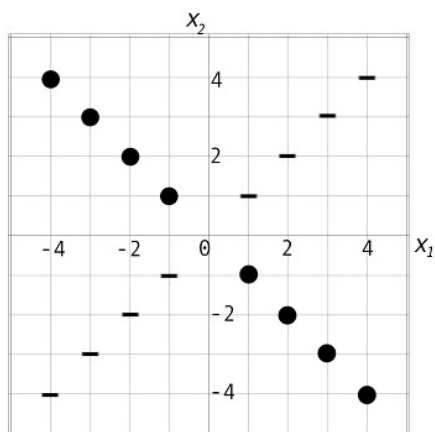
۷. [kmeans++] فرض کنید پنج داده داریم: (0, 0)، (1, 2)، (2, 3)، (3, 1)، (4, 1). تعداد خوشه‌ها را برابر ۳ در نظر بگیرید. مرکز خوشه اول به صورت تصادفی به عنوان (0,0) انتخاب شده است. این داده‌ها در شکل زیر نشان داده شده‌اند.



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

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، هر کدام برای هر نقطه داده)

[امتیازی] **[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 $(x_1, x_2) = (1, 1)$ وجود دارد. دو کلاس داریم: دایره‌ها و خط‌ها. هدف شما ساخت یک درخت تصمیم برای انجام طبقه‌بندی است.



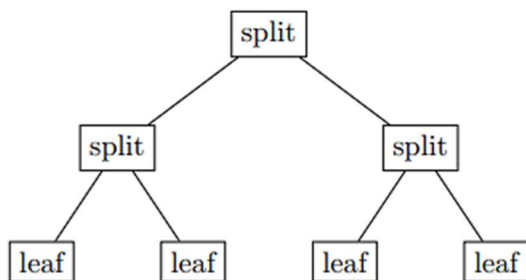
الف) آنتروپی برای کل مجموعه داده‌ها چقدر است؟

ب) آنتروپی میانگین وزنی دو گره فرزند در تقسیم فیچر x_1 در موقعیت $x_1=0$ برای این مجموعه داده‌ها چقدر است؟

ج) آنتروپی میانگین وزنی دو گره فرزند در تقسیم فیچر x_1 در موقعیت $x_1=1.5$ برای این مجموعه داده‌ها چقدر است؟

د) حال هر تقسیم‌بندی ممکن را در بعد z 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.

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.



موفق باشید!!