۱- موارد زیر را تعریف کنید؟

داده پرت:

داده‌ای که به طور قابل ملاحظه‌ای از سایر داده‌های دیگر (هم رده) فاصله دارد.

تحلیل توسعه:

تحلیل توسعه داده اغلب به یافتن مدل هایی برای اشیایی که در طول زمان رفتارشان را عوض می کنند گفته می شود. مثلا پیش بینی قیمت یک کالا در یک بازار بورس.

نمودار ROC:

نموداری جهت نمایش کارایی یک رده‌بند دو رده‌ای که با تغییر سطح پذیرش یک رده، TP=های گوناگون را در مقابل FP=های گوناگون نمایش می‌دهد.

Cross Validation:

نوعی روش ارزیابی است که در آن مجموعه داده به n بخش مجزا و بدون همپوشانی افراز شده و در n مرحله پشت سر هم، هر بار یک بخش از این n بخش به عنوان مجموعه آزمایشی و بقیه به عنوان مجموعه آموزشی در نظر گرفته می‌شوند.

مکعب داده:

یک روش نمایش فشرده داده‌های یک انبار داده است که در آن داده‌ها بر اساس یک معیار (مثلا در ساده‌ترین حالت فراوانی) و یک یا چند فیلد به صورت یک آرایه یک یا چند بعدی نمایش داده می‌شوند.

تراکنش:

تراکنش یک پردازش یکپارچه و هم روند روی پایگاه داده است. هدف این بود که اگر می خواهیم کاری انجام دهیم، این کار در قالبی انجام شود که یقینا یکپارچه انجام بشود و اثر جانبی روی بقیه نگذارد.

تعمیم:

به طور ساده یعنی با دیدن چندین داده که یک حقیقت را نشان می دهند، آن حقیقت را استنتاج کردن تعمیم گفته می شود. یعنی اینکه مثلا ما امروز از خانه بیرون می رویم می بینیم که خورشید هست؛ فردا هم همین طور؛ روز بعد هم همین طور؛ پس نتیجه می گیریم که هر وقت روز بود خورشید هم هست. به این فرایند از جز به کل رسیدن تعمیم می گوییم.

رده بند:

ساختن مدلی که بتواند یک الگو را در یکی از اعضای یک مجموعه از مفاهیم پیش تعیین شده به نام رده قرار دهد را رده بند گویند. این انتساب باید بگونه ای انجام پذیرد که الگوهای موجود در یک رده، بیشترین و الگوهای موجود در رده های متفاوت کمترین شباهت را به هم داشته باشند.

قانون انجمن:

یعنی ارتباط بین دو پدیده که با هم مکرر رخ می‌دهند؛ مثلا هر وقت فرد خواسته پول بردارد، قبل آن یک رسید دریافت کرده است.

خوشه بندی:

به نوعی از رده بندی گوییم که از قبل دسته یا خوشه ها مشخص نیستند.

۲- داده کاوی تعریف کنید؟ مراحل آن را نام ببرید و توضیح دهید.

به بیان ساده می‌توان گفت داده کاوی استخراج دانش از حجم زیادی از داده ها و یا اطلاعات است. به عبارت دیگر، عموما داده کاوی را فقط یکی از گام های اساسی در فرآیند کشف دانش می دانند. کشف دانش شامل مراحل زیر است.

i. پاکسازی داده (حذف نویز و داده های متناقض)
ii. یکپارچه سازی داده (که در آن ممکن است منابع چند گانه ی داده ها با هم ترکیب شود)
iii. انتخاب داده ها (داده ها برای تجزیه و تحلیل از پایگاه داده ها بازیابی می شود)
iv. تبدیل داده (در آن داده ها به شکل های مناسب تبدیل و ثبت می شوند تا عملیات کاوش آسان تر انجام شود)
v. داده کاوی (فرآیندی ضروری که روش هایی هوشمند برای الگوی استخراج داده معرفی می شود)
vi. ارزیابی الگو (شناسایی الگوهای جالب برای کسب دانش بر پایه تعریف الگوی جالب؛)
vii. نمایش دانش (که از تکنیک هایی برای تجسم و ارائه دانش استخراج شده به کاربر استفاده می شود)

۳- انواع داده های که روی آنها داده کاوی قابل انجام است، را نام ببرید و توضیح دهید؟

۱ داده های پایگاه داده رابطه ای

یک سیستم پایگاه داده، یا سیستم مدیریت پایگاه داده، از یک مجموعه ای از داده هایی که به هم مرتبط هستند، پایگاه داده، و یک سری نرم افزارهایی برای مدیریت و دسترسی به داده ها تشکیل می شود.

۲ انبارهای داده

فرض کنید که شرکت AllElectronics یک شرکت بین المللی موفق می باشد که شعبه هایی را در سرتاسر دنیا دارد. هر شعبه دارای پایگاه داده های خودش می باشد. رئیس شرکت خواسته است که یک تحلیلی را در مورد فروش های هر کالا در هر شعبه برای فصل سوم سال بداند. این امر یک کار سخت برای پرس و جوهای رابطه ای می باشد؛ چرا که داده ها روی چندین پایگاه داده توزیع شده در سرتاسر دنیا قرار دارد. اگر این شرکت یک انبار داده داشت این کار آسان می بود. یک انبار داده، یک انباری از اطلاعات می باشد که از چندین منبع جمع شده اند و تحت یک شمای یکپارچه ذخیره شده اند و اغلب در یک مکان نگهداری می شوند.

۳ پایگاه داده های تراکنشی

به طور کلی یک پایگاه تراکنشی از یک فایل که هر کدام از رکوردهایش یک تراکنش را نشان می دهد گفته شده است. یک تراکنش معمولا شامل یک شناسه تراکنش و لیستی از اقلامی که آن تراکنش را می سازند، می باشد؛ برای نمونه تعداد مورد های خرید شده را شامل می شود.

۴ سیستم های اطلاعاتی و داده ای پیشرفته و کاربردهای پیشرفته

۴-۱ پایگاه داده های زمانی

اغلب پایگاه داده های رابطه ای هستند که ویژگی های مرتبط به زمان را نگهداری می کنند.

۴-۲ پایگاه داده های دنباله ای

رشته ای از حوادث مرتب شده و متوالی را نگهداری می کنند که هر کدام به ترتیب در بستر زمان رخ داده اند بدون معنای مستقیم زمان. برای مثال، دنباله ی کلیک کردن در یک وب سایت را می توان پایگاه داده های دنباله ای بنامیم؛ در حالی که پایگاه داده های سری زمانی پایگاه داده هایی می باشند که مقادیر یک سری حوادث یا ویژگی ها را بر روی زمان نشان دهند. مثلا داده هایی که هر چند دقیقه مقدار بازار بورس را نمایش می دهد یا نگهداری دما بر حسب زمان.

۴-۳ پایگاه داده های زمانی مکانی

پایگاه داده های مکانی شامل اطلاعات مرتبط به مکان می باشند. مثال های از این نوع پایگاه داده ها، پایگاه داده های جغرافیایی و مجتمع سازی در سطح بسیار بالا یا پایگاه داده های طراحی به کمک کامپیوتر و پایگاه داده های تصاویر ماهواره ای و پزشکی است.

۴-۴ پایگاه داده های متنی و چندرسانه ای

پایگاه داده هایی که از کلمات به منظور توصیف اشیا استفاده می کنند پایگاه داده های متنی گفته می شوند. ویژگی اصلی این پایگاه داده ها این است که به شدت بی ساختار می باشند.

۴-۵ پایگاه داده های متنی نیمه ساختارمند

بعضی از پایگاه داده های متنی که تا حدودی ساختارمند هستند نیمه ساختارمند به آن ها گفته می شود. برای مثال ایمیل ها و بعضی وب پیج های HTML از این دسته هستند.

۴-۶ تار نمای جهانی وب

تار نمای جهانی وب و سرویس های اطلاعاتی توزیع شده مخصوص به خودش شبیه yahoo، google، America online، Alta vista و سایر موارد، سرویس های اطلاعاتی برخط بسیار غنی و جهانی را فراهم می کند که اشیا داده های ما از طریق لینک

های به همدیگر متصل اند که دسترسی های درون اینترنتی را برای کاربران تسهیل کند. کاربران از یک وب پیج به وب پیج دیگر و مورد علاقه خود به وسیله لینک ها نقل مکان می کنند.

۴- وظایف اصلی داده کاوی را نام ببرید و توضیح دهید؟

۱ توصیف کلاس یا مفهوم

مشخص سازی خواص و تفکیک سازی داده ها می توانند به رده ها و یا مفاهیم منتسب بشوند. برای مثال در شرکت *AllElectronics* کلاس های اقلام فروخته شده، اقلامی که برای فروش هستند، می توانند شامل کامپیوترها، پرینترها باشند در حالی که مفاهیم و یا کلاس ها در مشتری ها می توانند خرج کننده های بزرگ یا خرج کننده های کوچک تقسیم بشوند.

۲ کاوش الگوهای پرتکرار، ارتباطات و همرخداد

الگوهای پرتکرار همانگونه که از اسمشان بر می آید الگوهایی هستند که به صورت فراوان در داده ها رخ می دهند. اگرچه انواع گوناگونی از این الگوها وجود دارند اما به طور معمول به مجموعه ای از اقلامی که به طور همزمان در یک مجموعه داده تراکنشی رخ می دهد، ما اصطلاحا مجموعه اقلام فراوان یا پرتکرار می گوییم.

۳ رده بندی و پیش بینی

رده بندی به فرآیند یافتن مدل یا تابع توصیف کننده و تمایز دهنده ای که رده های داده ای و مفاهیم داده ای را به منظور تواناسازی ما به تعیین کلاس یا رده اشیا جدید (با کمک آن مدل) گفته می شود. آن مدل بر اساس یک تحلیل بر روی داده های آموزشی به وجود آمده است که برای آن مجموعه آموزشی، برچسب کلاس آن ها را به عنوان ورودی به آن مدل می دهیم.

۴- تحلیل خوشه

تحلیل خوشه برعکس پیش بینی و رده بندی که ما تحلیلمان را بر روی یک سری اشیایی که بر چسب کلاس آن ها را می دانیم می باشد، است. خوشه بندی تحلیل اشیایی است که هیچ گونه مجموعه آموزشی برای آن ها وجود ندارد. هدف در خوشه بندی این می باشد که ما داده ها را در خوشه هایی قرار بدهیم که مشابهت بین داده های درون خوشه ای به حداکثر برسد، در حالی که مشابهت بین داده های بیرون خوشه ای به حداقل برسد.

۵- تحلیل داده دور افتاده

یک پایگاه داده ممکن است شامل اشیائی یا داده هایی باشد که با رفتار عمومی و مدل داده ها همخوانی چندانی ندارد. این داده ها را به اصطلاح داده های دور افتاده می گوییم. تحلیل داده دور افتاده را اصطلاحا کاوش داده پرت یا دور افتاده می گوییم.

۶ تحلیل توسعه

تحلیل توسعه داده اغلب به یافتن مدل هایی برای اشیایی که در طول زمان رفتارشان را عوض می کنند گفته می شود.

۵- مراحل پیش پردازش را نوشته و توضیح دهید؟

(الف) پاکسازی داده: می تواند برای حذف و یا تصحیح خطا و سازگار سازی داده ها به کار گرفته شود،

(ب) ادغام و یکپارچه سازی داده: داده هایی را که از چندین منبع می باشند، در یک مجموعه داده واحد منسجم می کند.

(ج) تبدیل داده ها: همانند تکنیک نرمال سازی (نرمال سازی باعث بهبود و صحت کارایی الگوریتم های داده کاوی می شود) می تواند به کار گرفته شود.

(د) کاهش داده می تواند حجم داده را با استفاده از اجتماع، حذف صفات تکراری و یا خوشه بندی داده ها کاهش دهد.

۶- انواع روش های اندازه گیری پراکندگی چند دسته اند؟ تعریف کنید.

الف) معیار توزیعی، معیاری است که برای یک مجموعه داده ای با تقسیم کردن آن به زیر مجموعه های کوچکتر محاسبه می شود؛ با محاسبه معیار برای هر زیر مجموعه و سپس ادغام نتایج برای رسیدن به مقدار کلی برای مجموعه داده اصلی انجام می شود. هر دو تابع *Sum()* و *Count()* معیار های توزیعی هستند

ب) معیار جبری: یک معیار جبری معیاری است که با به کارگیری تابع جبری روی یک یا چند معیار توزیعی محاسبه می شود. از این رو میانگین (یا *Mean()*) یک معیار جبری است.

ج) معیار کلی: یک معیار کلی، معیاری است که روی کل مجموعه داده ای محاسبه می شود. این مقدار با تقسیم بندی داده به زیر مجموعه ها و ادغام مقادیر بدست آمده، حاصل نمی شود. میانه نمونه ای از یک معیار کلی است. معیارهای کلی خیلی پرهزینه تر از معیارهای توزیعی است.

۷- انواع تکنیک های کاهش داده را نام ببرید؟

تکنیک های کاهش داده برای به دست آوردن نمایش مختصر مجموعه داده ای که از لحاظ حجم خیلی کوچکتر و در عین حال صحت و جامعیت داده اصلی را داراست به کار می روند. با این روش کاوش داده کاهش یافته موثر و کارآمدتر و منجر به تولید همان نتایج اصلی می شود.

استراتژی های کاهش داده در زیر ذکر شده اند:

۱-اجتماع مکعب داده ای، که عملیات اجتماع روی داده ها به منظور ساخت مکعب داده ای به کار می روند.

۲-انتخاب زیر مجموعه صفات: که صفات غیر مرتبط، کم مرتبط و زائد کشف و حذف شوند.

۳-کاهش ابعاد، که مکانیسم های کد گذاری برای کاهش اندازه مجموعه داده ای مورد استفاده قرار می گیرند.

۴-کاهش چندی، نمایش های داده ای کوچکتر همچون مدل های پارامتری (که نیازمند ذخیره سازی فقط پارامترهای مدل و نه خود داده واقعی می باشند) و یا روش های غیر پارامتری همچون خوشه بندی، نمونه گذاری و استفاده از هیستوگرام جایگزین داده اصلی می شوند.

۵-گسسته سازی و تولید سلسله مراتب مفهومی، مقادیر داده ای خام با محدوده یا سطوح مفهومی بالاتر جایگزین می شوند. گسسته سازی داده شکل دیگری از کاهش چندی است که برای تولید خودکار سلسله مراتب مفهومی مفید است. گسسته سازی و تولید سلسله مراتب مفهومی ابزارهای قدرتمندی برای کاوش داده می باشند. آنها امکان کاوش داده را در چند سطح انتزاع فراهم می کنند.

۸- تکنیک های کاهش چندی را نام ببرید؟

۱- مدل های رگرسیون و *Log-linear*

۲- هیستوگرام ها

۳- خوشه بندی

۴- نمونه گیری

۹- انبار داده را تعریف کنید و کلمات کلیدی آن را توضیح دهید؟

انبار داده یک مجموعه‌ی موضوع‌گرا، ادغام شده، متغیر با زمان و غیر فرار از داده‌ها است که برای پشتیبانی از فرایند اتخاذ تصمیم استفاده می‌شود.

۱ - موضوع‌گرا: یک انبار داده بر اساس موضوع‌های اساسی از قبیل مشتری، فروشنده، محصول و خرید سازماندهی می‌شود. علاوه بر تمرکز بر روی عملیات روز به روز و پردازش تراکنش یک سازمان، یک انبار داده بر روی مدلسازی و تحلیل (آنالیز) داده برای تصمیم‌گیرندگان نیز تمرکز می‌کند. از این رو، انبار داده معمولا به موضوعاتی که در فرایند تصمیم‌گیری مفید نیستند، توجه کمی دارد.

۲ - ادغام شده: یک انبار داده معمولا به وسیله مجتمع کردن چندین منبع غیر متجانس از قبیل پایگاه داده رابطه‌ای، فایل‌های بدون قالب و رکوردهای تراکنش بر خط ساخته می‌شود. تکنیک‌های پاکسازی و ادغام داده برای اطمینان از ثبات قراردادها، ساختارهای رمزگشایی، میزان صفات و غیره به کار برده می‌شوند.

۳ - متغیر با زمان: داده‌ها به منظور داشتن اطلاعات از زمان‌های گذشته (۵-۱۰ سال گذشته) ذخیره می‌شوند. هر ساختار اصلی در انبار داده شامل یک عنصر زمان به صورت آشکار یا ناآشکار است.

۴ - غیر فرار: یک انبار داده، معمولا یک مخزن جداگانه‌ی فیزیکی از داده‌هایی است که از داده‌های کاربردی موجود در محیط عملیاتی انتقال داده شده‌اند. به خاطر این جداسازی، یک انبار داده به مکانیزم‌های پردازش تراکنش، بازیافت و کنترل همزمانی نیازی ندارد. انبار داده معمولا به دو عملیات در دستیابی داده نیاز دارد: *بارگیری اولیه داده و دستیابی داده*

۱۰- مکعب داده چیست؟

یک مکعب داده به داده اجازه می‌دهد تا به صورت چندبعدی مدل شده و نشان داده شوند. مکعب داده به وسیله‌ی ابعاد و حقایق تعریف می‌شود.

۱۱- عملیات *OLAP* در مدل داده‌ی چند بعدی را نام ببرید و توضیح دهید.

Roll-up: این عملیات، با بالا رفتن از نمودار سلسله مراتبی مفهومی در یک بعد یا با کاهش بعد، متراکم سازی مکعب داده را اجرا می‌کند.

Drill-down: عکس عملیات *Roll-up* است. این عملیات از داده‌های با جزییات بیشتر به سمت داده‌های با جزییات کمتر می‌رود. عملیات *drill-down*، با پایین آمدن از نمودار سلسله مراتبی مفهومی در یک بعد یا معرفی کردن بعد جدید انجام می‌گیرد.

Slice and dice: عملیات *slice*، کار انتخاب یک بعد از مکعب مشخص را اجرا می‌کند که منجر به یک زیر مکعب می‌شود.

pivot: عملیات مجسم سازی است که بردارهای داده را به منظور ایجاد یک نمایش متناوب از داده‌ها می‌چرخاند.

۱۲- الف) کل قوانین انجمنی قابل تعریف بر روی جدول زیر چند قلم می باشد؟

ب) حمایت و پوشش قانون $A = \{F\} \rightarrow \mathrm{B} = \{\mathrm{T}\}$ را به دست آورید؟

ج) کدام قانون انجمنی در بین قوانین انجمنی یک صفت به یک صفت بیشترین اعتماد را دارد؟

د) کدام قانون انجمنی در بین قوانین انجمنی یک صفت به یک صفت بیشترین پوشش را دارد؟

A	B	C	D
T	T	T	T
T	F	T	F
F	T	T	F
F	F	T	T
F	F	F	T

الف)

$$\text{صفت} \rightarrow \text{صفت} \Longrightarrow 4 \times 3 \times 3^2$$

$$\text{دوصفت} \rightarrow \text{صفت} \Longrightarrow 4 \times 3 \times 2 \times 3^3$$

$$\text{دوصفت} \rightarrow \text{دوصفت} \Longrightarrow \frac{4 \times 3}{2} \times 3^4$$

۱۲۴۲ قلم قانون انجمنی گوناگون خواهیم داشت.

ب) پوشش برابر با ۶۰ درصد و حمایت برابر با ۶۷ درصد می باشد.

ج)

$$A = \{T, F\} \rightarrow \mathrm{B} = \{\mathrm{T}, \mathrm{F}\}$$

د)

$$A = \{T, F\} \rightarrow \mathrm{B} = \{\mathrm{T}, \mathrm{F}\}$$

۱۳- الف) نمودار جعبه ای را برای داده های زیر ترسیم کنید؟

-7, 14, 7, 15, 8, 9, 12, 11, 9, 10, 100, 21, 13, 11, 10, 15, 9, 10, 7, 13

ب) نمودار جعبه ای را برای داده های زیر ترسیم کنید؟

-5, 44, 1, 12, 3, 1, 12, 14, 9, 10, 40, 1, 1, 17, 12, 15, 9, 10, 7, 13, 3, 4, 5, 11, 6

جواب الف)

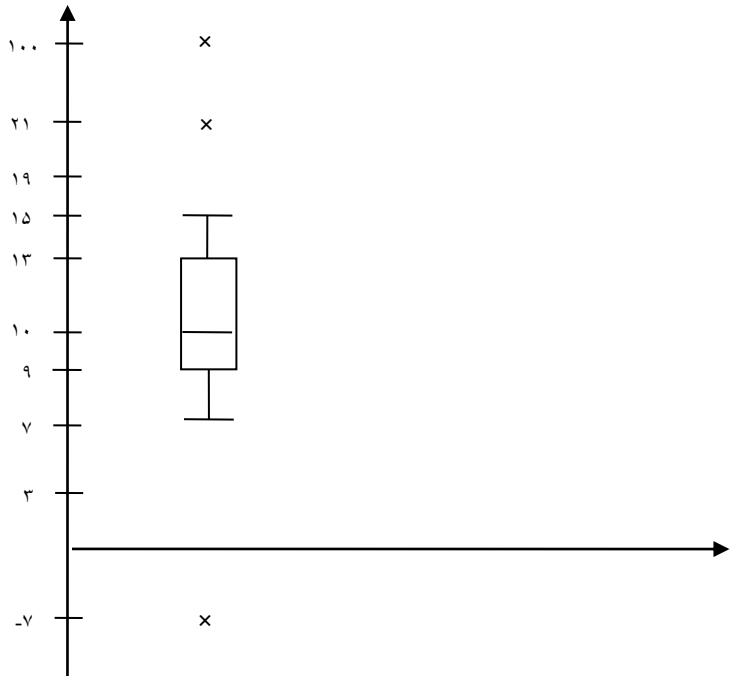
ابتدا داده ها را مرتب می کنیم.

-7, 7, 7, 8, 9, 9, 9, 10, 10, 10, 11, 11, 12, 13, 13, 14, 15, 15, 21, 100

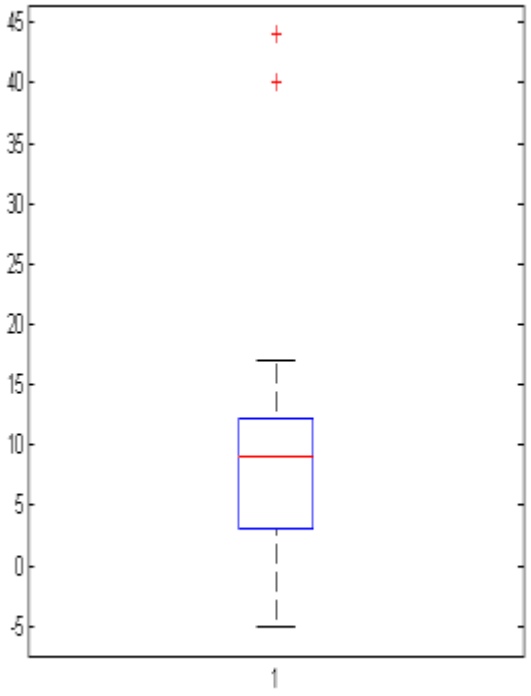
اندیس چارک اول مطابق با رابطه زیر محاسبه می شود:

$$Index_{Q_i} = \left\lceil \frac{DataNumber \times i \times 25}{100} \right\rceil$$

پس اندیس چارک اول برابر است با ۵ که مقدار چارک اول عدد ۹ می باشد. پس اندیس چارک دوم یا همان میانه برابر است با ۱۰ که مقدار چارک دوم عدد ۱۰ می باشد. همچنین اندیس چارک سوم برابر است با ۱۵ که مقدار چارک سوم عدد ۱۳ می باشد. بنابراین IQR برابر است با ۴. در نتیجه بازه اطمینان که به شکل $[Q_1 - 1.5 \times IQR, Q_3 + 1.5 \times IQR]$ تعریف می شود برابر است با $[3,19]$.



جواب ب)



۱۴-داده های زیر را در نظر بگیرید. این داده ها را با انواع روش های نرمال سازی زیر نرمال کنید.

الف) نرمال سازی ۱ و ۵

ب) نرمال سازی میانگین ۰ و انحراف معیار ۱

-7, 14, 7, 15, 8, 9, 12, 11, 9, 10, 100, 21, 13, 11, 10, 15, 9, 10, 7, 13

پاسخ الف)

$$y = \frac{max_2 - min_2}{max_1 - min_1}(x - min_1) + min_2$$

در این معادله max_2 برابر است با ۵، min_2 برابر است با ۱، max_1 برابر است با ۱۰۰ و min_1 برابر است با ۷-. بنابراین معادله بال به شکل زیر نوشته می شود.

$$y = \frac{4}{107}(x + 7) + 1$$

حال هر عدد را در معادله فوق گذاشته و حاصل را می نویسیم.

1, 1.79, 1.52, 1.82, 1.56, 1.6, 1.71, 1.67, 1.6, 1.64, 5, 2.05, 1.75, 1.67, 1.64, 1.82, 1.6, 1.64, 1.52, 1.75

پاسخ ب)

$$\mu = \frac{-7 + 14 + 7 + 15 + 8 + 9 + 12 + 11 + 9 + 10 + 100 + 21 + 13 + 11 + 10 + 15 + 9 + 10 + 7 + 13}{20}$$

$$\mu = \frac{297}{20} = 14.85$$

$$\sigma = 20.19$$

-1.08, -.04, -.39, .01, -.34, -.29, -.14, -.19, -.29, -.24, 4.22, .3, -.09, -.19, -.24, .01, -.29, -.24, -.39, -.09

۱۵-الف) فرض کنید A و B دو صفت عددی در یک پایگاه داده هستند. همچنین فرض کنید α و β دو ثابت عددی هستند. فرض کنید $r_{A,B}$ همبستگی صفات A و B است. نشان دهید $r_{\beta A+\alpha,B} = r_{A,B}$.

ب) داده های زیر را در نظر بگیرید. ویژگی A بیشترین همبستگی را با کدام ویژگی دارا می‌باشد.

A	B	C	D	E	F
۲	۳	۲	۴	۱	-۳
۳	۴	۱	۵	۲	-۵
۳	۴	۱	۵	۲	-۵
۴	۴	۱	۵	۳	-۵
۴	۵	۰	۶	۳	-۷
۵	۵	۰	۶	۳	-۷

جواب)

ابتدا همبستگی بین ویژگی A و B را بر طبق رابطه زیر محاسبه می‌کنیم.

$$r_{A,B} = \frac{\sum_{i=1}^{N}(a_i - \bar{A})(b_i - \bar{B})}{N\sigma_A\sigma_B} = \frac{\sum_{i=1}^{N}(a_i b_i)\ - N\bar{A}\bar{B}}{N\sigma_A\sigma_B}$$

$$r_{A,B} = \frac{2\times 3 + 3\times 4 + 3\times 4 + 4\times 4 + 4\times 5 + 5\times 5 - 6\times 3.5\times 4.17}{6\times 1.05\times 0.75} = 0.74$$

چنان که می دانیم:

$$r_{\beta A+\alpha,B} = \frac{\sum_{i=1}^{N}((\beta a_i + \alpha)b_i)\ - N(\beta\bar{A} + \alpha)\bar{B}}{N\beta\sigma_A\sigma_B} = \frac{\sum_{i=1}^{N}(\beta a_i b_i)\ + \sum_{i=1}^{N}(\alpha b_i)\ - N(\overline{\beta A} + \alpha)\bar{B}}{N\beta\sigma_A\sigma_B}$$
$$= \frac{\sum_{i=1}^{N}(\beta a_i b_i)\ + \alpha\sum_{i=1}^{N} b_i\ - N\beta\bar{A}\bar{B} - N\alpha\bar{B}}{N\beta\sigma_A\sigma_B} = \frac{\beta\sum_{i=1}^{N}(a_i b_i)\ + N\alpha\bar{B} - N\beta\bar{A}\bar{B} - N\alpha\bar{B}}{N\beta\sigma_A\sigma_B}$$
$$= \frac{\beta\sum_{i=1}^{N}(a_i b_i)\ - N\beta\bar{A}\bar{B}}{N\beta\sigma_A\sigma_B} = \frac{\beta(\sum_{i=1}^{N}(a_i b_i)\ - N\bar{A}\bar{B})}{N\beta\sigma_A\sigma_B} = \frac{\sum_{i=1}^{N}(a_i b_i)\ - N\bar{A}\bar{B}}{N\sigma_A\sigma_B} = r_{A,B}$$

از آنجایی که صفت $c_i = 5 - b_i$ است، $r_{A,C}$ نیز برابر $r_{A,B}$ یعنی ۰,۷۴ است. به دلیل مشابه، آنجایی که صفت $d_i = b_i + 1$ است، $r_{A,D}$ نیز برابر $r_{A,B}$ یعنی ۰,۷۴ است. همبستگی بین ویژگی A و E را بر طبق رابطه زیر محاسبه می‌کنیم.

$$r_{A,E} = \frac{\sum_{i=1}^{N}(a_i - \bar{A})(e_i - \bar{E})}{N\sigma_A\sigma_E} = \frac{\sum_{i=1}^{N}(a_i e_i)\ - N\bar{A}\bar{E}}{N\sigma_A\sigma_E}$$

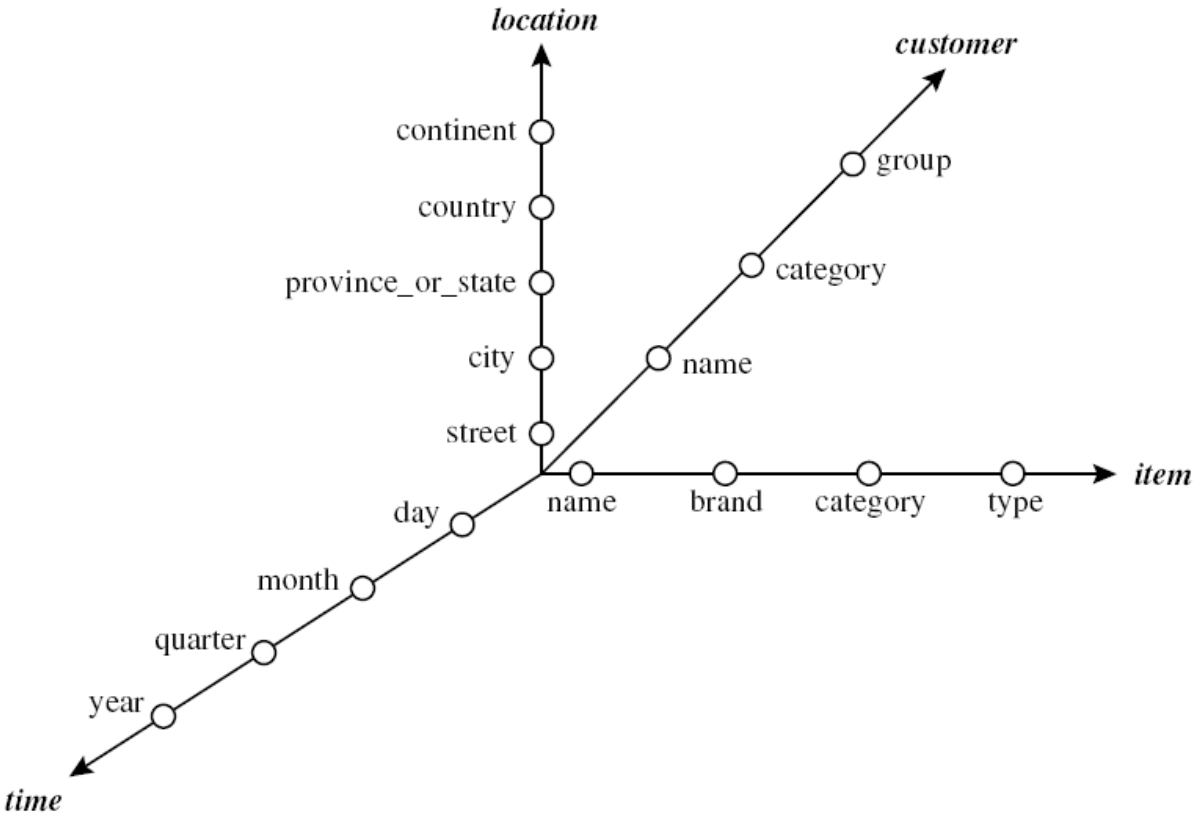
$$r_{A,E} = \frac{2\times 1 + 3\times 2 + 3\times 2 + 4\times 3 + 4\times 3 + 5\times 3 - 6\times 3.5\times 2.33}{6\times 1.05\times 0.82} = 0.78$$

آنجایی که صفت $f_i = 3 - 2\times b_i$ است، $r_{A,D}$ نیز برابر $r_{A,F}$ یعنی ۰,۷۴ است. پس همبستگی صفات A و E از همه بیشتر است.

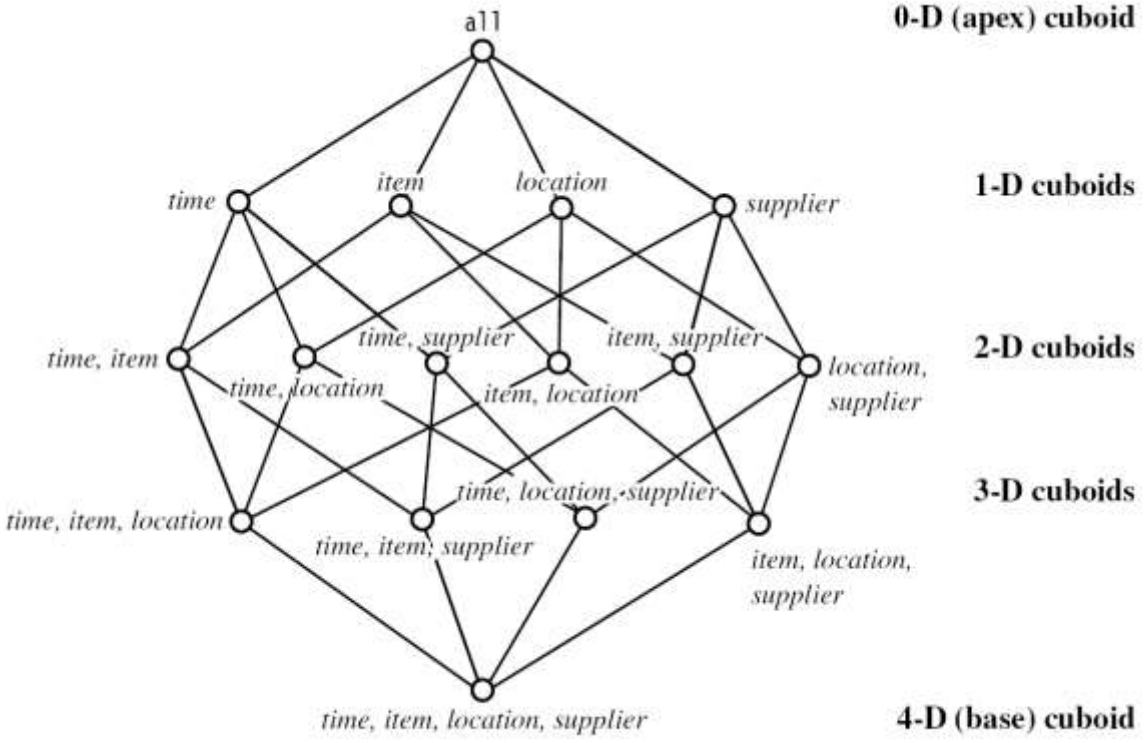
۱۶-چهار ویژگی location، customer، item و time را در نظر بگیرید.

الف) شبکه کعب‌ها را برای آنها ترسیم کنید.

ب) اگر سلسله مراتب مفهومی روی هر ویژگی شبیه به زیر باشد، چند مکعب گوناگون روی این ویژگی‌ها تعریف می-شود؟



جواب الف)



جواب ب)

$$(5+1)\times(3+1)\times(4+1)\times(4+1)=600$$

۱۷- فرض کنید یک گروه ۱۵۰۰ نفری مورد تحقیق قرار بگیرند. جنسیت هر شخص ذکر شده است. از هر شخص در مورد نوع کتاب مورد مطالعه که داستانی یا غیرداستانی است پرسیده شده است. بنابراین ما دو صفت داریم، جنس شخص یا ***gender*** و نوع کتاب یا ***preferred_Reading***. فراوانی (تعداد) مشاهده شده از رخداد توام در جدول پیشامد زیر خلاصه شده است.

	Male	***Female***	***Total***
Fiction	*250*	*200*	*450*
Non-fiction	*50*	*1000*	*1050*
Total	*300*	*1200*	*1500*

مطلوب است تحلیل همبستگی صفات رده بندی (گسسته) با استفاده از χ^2. با چه اطمینانی این دو وابسته هستند؟ فرض کنید جدول زیر را داریم.

درجه	سطح اطمینان	سطح قبول برای وابستگی

۱	۰,۱	۲,۷۱
۲	۰,۱	۴,۶۱
۳	۰,۱	۶,۲۵
۴	۰,۱	۷,۷۸
۱	۰,۰۱	۶/۶۳
۲	۰,۰۱	۹/۲۱
۳	۰,۰۱	۱۱/۳۴
۴	۰,۰۱	۱۳/۲۸
۱	۰,۰۰۱	۱۰/۸۳
۲	۰,۰۰۱	۱۳/۸۲
۳	۰,۰۰۱	۱۶/۲۷
۴	۰,۰۰۱	۱۸/۴۷
۱	۰,۰۰۰۱	۱۵/۱۴
۲	۰,۰۰۰۱	۱۸/۴۲
۳	۰,۰۰۰۱	۲۱/۱۱
۴	۰,۰۰۰۱	۲۳/۵۱
۱	۰,۰۰۰۰۱	۱۹/۵۱
۲	۰,۰۰۰۰۱	۲۳/۰۳
۱	۰,۰۰۰۰۰۱	۲۳/۹۳
۲	۰,۰۰۰۰۰۱	۲۷/۶۳
۱	۰,۰۰۰۰۰۰۱	۲۸/۳۷
۲	۰,۰۰۰۰۰۰۱	۳۲/۲۴

جواب)

ابتدا فراوانی مورد انتظار برای هر یک از درایه‌های جدول را با استفاده از معادله زیر به دست می‌آوریم.

$$e_{ij} = \frac{count(A = a_i) \times count(B = b_j)}{N}$$

با استفاده از فرمول بالا می توان فراوانی مورد انتظار برای هر خانه جدول را محاسبه کرد. برای مثال فراوانی مورد انتظار برای خانه (*male, fiction*) در زیر محاسبه شده است.

$$e_{11} = \frac{count(male) \times count(fiction)}{N} = \frac{300 \times 450}{1500} = 90$$

توجه کنید که مجموع فراوانی های مورد انتظار در هر سطر مساوی با تعداد کل فراوانی مشاهده شده برای آن سطر و مجموع فراوانی های مورد انتظار در هر ستون مساوی با تعداد کل فراوانی های مشاهده شده برای آن ستون است.

جدول زیر فراوانی مورد انتظار درایه ها است.

	Male	Female	Total
Fiction	90	360	450
Non-fiction	210	840	1050
Total	300	1200	1500

حال با استفاده از معادله زیر χ^2 را محاسبه می‌نماییم.

$$\chi^2 = \sum_{i=1}^{c}\sum_{j=1}^{r}\frac{(o_{ij}-e_{ij})^2}{e_{ij}} = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840}$$
$$= 284/44 + 121/90 + 71/11 + 30/48 = 507/93$$

برای این جدول 2×2، درجه آزادی برابر با ۱= (۱-۲)(۱-۲) است. برای ۱ درجه آزادی، مقدار χ^2 مورد نیاز برای رد این فرضیه در سطح ۰/۰۰۱ برابر با ۱۰,۸۳ (از نقاط بالای توزیع χ^2 از جدول گرفته شده است) است. از آن جایی که مقدار محاسبه شده از این مقدار بالاتر است، ما این فرضیه که دو صفت *gender* و *preferred_Reading* مستقل هستند را رد و نتیجه می گیریم که دو صفت ذکر شده برای این گروه افراد کاملاً وابسته است. با همین تحلیل برای ۱ درجه آزادی، مقدار χ^2 مورد نیاز برای رد این فرضیه در سطح ۰/۰۰۰۱ برابر با ۱۵,۱۴ است. از آن جایی که مقدار محاسبه شده از این مقدار بالاتر است، ما این فرضیه که دو صفت *gender* و *preferred_Reading* مستقل هستند را رد و نتیجه می گیریم که دو صفت ذکر شده برای این گروه افراد کاملاً وابسته است. با همین تحلیل برای ۱ درجه آزادی، مقدار χ^2 مورد نیاز برای رد این فرضیه در سطح ۰/۰۰۰۰۱ برابر با ۱۹,۵۱ است. از آن جایی که مقدار محاسبه شده از این مقدار بالاتر است، ما این فرضیه که دو صفت *gender* و *preferred_Reading* مستقل هستند را رد و نتیجه می گیریم که دو صفت ذکر شده برای این گروه افراد کاملاً وابسته است. با همین تحلیل برای ۱ درجه آزادی، مقدار χ^2 مورد نیاز برای رد این فرضیه در سطح ۰/۰۰۰۰۰۱ برابر با ۲۳,۹۳ است. از آن جایی که مقدار محاسبه شده از این مقدار بالاتر است، ما این فرضیه که دو صفت *gender* و *preferred_Reading* مستقل هستند را رد و نتیجه می گیریم که دو صفت ذکر شده برای این گروه افراد کاملاً وابسته است. ولی برای ۱ درجه آزادی، مقدار χ^2 مورد نیاز برای رد این فرضیه در سطح ۰/۰۰۰۰۰۰۱ برابر با ۲۸,۳۷ است. از آن جایی که مقدار محاسبه شده از این مقدار پایین‌تر است، ما این فرضیه که دو صفت *gender* و *preferred_Reading* مستقل هستند را رد و نتیجه می گیریم که دو صفت ذکر شده برای این گروه افراد کاملاً وابسته است. پس سطح اطمینان وابستگی این دو برابر زیر است:

$$1 - 0.0000001 = 0.9999999$$

۱۸- بهترین نقطه برش برای **برای صفت Taxable Income را محاسبه** کنید؟

Taxable Income	Cheat
125K	No
100K	No
70K	No
120K	No
95K	Yes
60K	No
220K	No
85K	Yes
75K	No
90K	Yes

جواب

در گام اول، ویژگی Taxable Income را تبدیل به یک ویژگی باینری می‌کنیم.

ابتدا برش بر روی 60K را انجام می‌دهیم. در نتیجه مجموعه داده به شکل زیر، در خواهد آمد.

Taxable Income	Cheat
1	No
1	No
1	No
1	No
1	Yes
0	No
1	No
1	Yes
1	No
1	Yes

ارزش این برش را محاسبه می‌کنیم.

$$Info_{TI}^{60K} = \frac{1}{10} * 0 + \frac{9}{10} * \left(-\left(\frac{3}{9} * \log_2^{\frac{3}{9}} + \frac{6}{9} * \log_2^{\frac{6}{9}} \right) \right) = 0.83$$

$$Info_{TI}^{70K} = \frac{2}{10} * 0 + \frac{8}{10} * \left(-\left(\frac{3}{8} * \log_2^{\frac{3}{8}} + \frac{5}{8} * \log_2^{\frac{5}{8}} \right) \right) = 0.76$$

$$Info_{TI}^{75K} = \frac{3}{10} * 0 + \frac{7}{10} * \left(-\left(\frac{3}{7} * \log_2^{\frac{3}{7}} + \frac{4}{7} * \log_2^{\frac{4}{7}} \right) \right) = 0.69$$

$$Info_{TI}^{85K} = \frac{4}{10} * \left(-\left(\frac{1}{4} * \log_2^{\frac{1}{4}} + \frac{3}{4} * \log_2^{\frac{3}{4}} \right) \right) + \frac{6}{10} * \left(-\left(\frac{2}{6} * \log_2^{\frac{2}{6}} + \frac{4}{6} * \log_2^{\frac{4}{6}} \right) \right) = 0.88$$

$$Info_{TI}^{90K} = \frac{5}{10} * \left(-\left(\frac{2}{5} * \log_2^{\frac{2}{5}} + \frac{3}{5} * \log_2^{\frac{3}{5}} \right) \right) + \frac{5}{10} * \left(-\left(\frac{1}{5} * \log_2^{\frac{1}{5}} + \frac{4}{5} * \log_2^{\frac{4}{5}} \right) \right) = 0.85$$

$$Info_{TI}^{95K} = \frac{6}{10} * \left(-\left(\frac{3}{6} * \log_2^{\frac{3}{6}} + \frac{3}{6} * \log_2^{\frac{3}{6}} \right) \right) + \frac{4}{10} * \left(-\left(\frac{3}{6} * \log_2^{\frac{3}{6}} + \frac{3}{6} * \log_2^{\frac{3}{6}} \right) \right) = 0.6$$

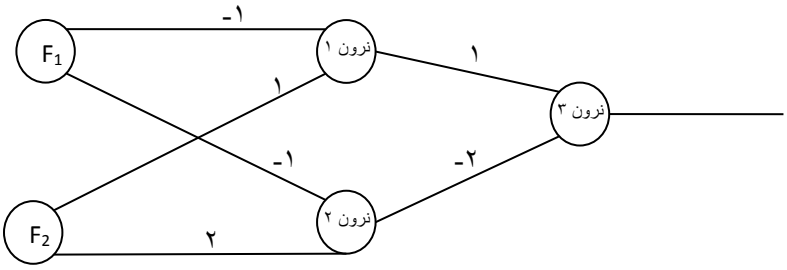
$$Info_{TI}^{100K} = \frac{7}{10} * \left(-\left(\frac{4}{7} * \log_2^{\frac{4}{7}} + \frac{3}{7} * \log_2^{\frac{3}{7}} \right) \right) + \frac{3}{10} * 0 = 0.69$$

$$Info_{TI}^{120K} = \frac{8}{10} * \left(-\left(\frac{5}{8} * \log_2^{\frac{5}{8}} + \frac{3}{8} * \log_2^{\frac{3}{8}} \right) \right) + \frac{2}{10} * 0 = 0.76$$

$$Info_{TI}^{125K} = \frac{9}{10} * \left(-\left(\frac{6}{9} * \log_2^{\frac{6}{9}} + \frac{3}{9} * \log_2^{\frac{3}{9}} \right) \right) + \frac{1}{10} * 0 = 0.83$$

پس بهترین برش "<=95K" است.

۱۹- یک رده‌بند شبکه عصبی MLP آموزش دیده بر روی داده‌های افراد سرطانی و غیر سرطانی، در زیر نمایش داده شده است. نرون‌های یک و دو از یک تابع فعالیت خطی (PureLine) استفاده می‌کنند و نرون ۳ از تابع فعالیت logsig استفاده می‌کند. مقدار بایاس (b) نرون‌های ۱، ۲ و ۳ به ترتیب ۱، ۱، ۰ است. این رده‌بند را بر روی بیست داده زیر تست می‌کنیم. دقت (Accuracy)، صحت (Precision)، فراخوان (Recall)، معیار فیشر (F-Measure)، و نمودار ROC این رده‌بند را به دست آورید. نکته: حالتی را در نظر بگیرید که بیشترین کارایی را دارد. در شرایط مساوی (Tie)، به نفع رده غیر سرطانی قضاوت شود.



نکته: $logsig(x) = \frac{1}{1+e^{-x}}$

مجموعه داده تست

برچسب	ویژگی ۲ (F_2)	ویژگی ۱ (F_1)
+	۱	۱
+	۲	۱

+	۳	۱
+	۱	۲
+	۲	۲
+	۳	۲
+	۱	۳
+	۲	۳
+	۳	۳
+	۴	۴
−	−۴	−۴
−	−۱	−۱
−	−۲	−۱
−	−۳	−۱
−	−۱	−۲
−	−۲	−۲
−	−۳	−۲
−	−۱	−۳
−	−۲	−۳
−	−۳	−۳

جواب

خروجی کل شبکه از تابع زیر به دست می‌آید.

$$\frac{1}{1+e^{-((1)(-F_1+F_2+1)+(-2)(-F_1+2F_2+1))}}=\frac{1}{1+e^{-(-F_1+F_2+1+2F_1-4F_2-2)}}=\frac{1}{1+e^{3F_2-F_1+1}}$$

خروجی شبکه بر مجموعه داده آزمایشی به شکل زیر است.

مقادیر پیش‌بینی شده	برچسب واقعی
۰/۰۴۷۴	+
۰/۰۰۲۵	+
۰/۰۰۰۱	+
۰/۱۱۹۲	+
۰/۰۰۶۷	+
۰/۰۰۰۳	+
۰/۲۶۸۹	+
۰/۰۱۸۰	+
۰/۰۰۰۹	+
۰/۰۰۰۱	+

۰/۹۹۹۱	–
۰/۷۳۱۱	–
۰/۹۸۲۰	–
۰/۹۹۹۱	–
۰/۵۰۰۰	–
۰/۹۵۲۶	–
۰/۹۹۷۶	–
۰/۲۶۸۹	–
۰/۸۸۰۸	–
۰/۹۹۳۳	–

برای بخش اول، ابتدا سطح رده مثبت را (0.5 < مقدار) در نظر می‌گیریم.

برچسب پیش‌بینی شده	مقادیر پیش‌بینی شده	برچسب واقعی
+	۰/۰۴۷۴	+
+	۰/۰۰۲۵	+
+	۰/۰۰۰۱	+
+	۰/۱۱۹۲	+
+	۰/۰۰۶۷	+
+	۰/۰۰۰۳	+
+	۰/۲۶۸۹	+
+	۰/۰۱۸۰	+
+	۰/۰۰۰۹	+
+	۰/۰۰۰۱	+
–	۰/۹۹۹۱	–
–	۰/۷۳۱۱	–
–	۰/۹۸۲۰	–
–	۰/۹۹۹۱	–
–	۰/۵۰۰۰	–
–	۰/۹۵۲۶	–
–	۰/۹۹۷۶	–
+	۰/۲۶۸۹	–
–	۰/۸۸۰۸	–
–	۰/۹۹۳۳	–

$$Accuracy = \frac{19}{20} = 95\%$$

$$Precision = \frac{10}{11} = 91\%$$

$$Recall = \frac{10}{10} = 100\%$$

$$F\_Measure = \frac{2 \times Recall \times Precision}{Recall + Precision} = 95\%$$

برای محاسبه ROC، ابتدا سطح رده مثبت را (1 > مقدار) در نظر می‌گیریم. پس برچسب‌های پیش‌بینی شده به شکل زیر خواهند بود.

برچسب پیش‌بینی شده	مقادیر پیش‌بینی شده	برچسب واقعی
+	۰/۰۴۷۴	+
+	۰/۰۰۲۵	+
+	۰/۰۰۰۱	+
+	۰/۱۱۹۲	+
+	۰/۰۰۶۷	+
+	۰/۰۰۰۳	+
+	۰/۲۶۸۹	+
+	۰/۰۱۸۰	+
+	۰/۰۰۰۹	+
+	۰/۰۰۰۱	+
+	۰/۹۹۹۱	−
+	۰/۷۳۱۱	−
+	۰/۹۸۲۰	−
+	۰/۹۹۹۱	−
+	۰/۵۰۰۰	−
+	۰/۹۵۲۶	−
+	۰/۹۹۷۶	−
+	۰/۲۶۸۹	−
+	۰/۸۸۰۸	−
+	۰/۹۹۳۳	−

یعنی TP و FP از روابط زیر برابر ۰ و ۰ خواهند بود.

$$FP = \frac{\text{تعداد برچسب‌های مثبت پیش‌بینی شده غلط}}{\text{تعداد برچسبهای منفی}} = \frac{10}{10} = 100\%$$

$$TP = \frac{\text{تعداد برچسب‌های مثبت پیش‌بینی شده درست}}{\text{تعداد برچسبهای مثبت}} = \frac{10}{10} = 100\%$$

سپس سطح رده مثبت را (0.9991 > مقدار) در نظر می‌گیریم. پس برچسب‌های پیش‌بینی شده به شکل زیر خواهند بود.

برچسب پیش‌بینی شده	مقادیر پیش‌بینی شده	برچسب واقعی
+	۰/۰۴۷۴	+
+	۰/۰۰۲۵	+
+	۰/۰۰۰۱	+
+	۰/۱۱۹۲	+
+	۰/۰۰۶۷	+
+	۰/۰۰۰۳	+
+	۰/۲۶۸۹	+
+	۰/۰۱۸۰	+
+	۰/۰۰۰۹	+
+	۰/۰۰۰۱	+
−	۰/۹۹۹۱	−
+	۰/۷۳۱۱	−
+	۰/۹۸۲۰	−
−	۰/۹۹۹۱	−
+	۰/۵۰۰۰	−
+	۰/۹۵۲۶	−
+	۰/۹۹۷۶	−
+	۰/۲۶۸۹	−
+	۰/۸۸۰۸	−
+	۰/۹۹۳۳	−

یعنی:

$$FP = \frac{\text{تعداد برچسب‌های مثبت پیش‌بینی شده غلط}}{\text{تعداد برچسبهای منفی}} = \frac{8}{10} = 80\%$$

$$TP = \frac{\text{تعداد برچسب‌های مثبت پیش‌بینی شده درست}}{\text{تعداد برچسبهای مثبت}} = \frac{10}{10} = 100\%$$

سپس سطح رده مثبت را (مقدار $0.9976 <$) در نظر می‌گیریم. پس برچسب‌های پیش‌بینی شده به شکل زیر خواهند بود.

برچسب پیش‌بینی شده	مقادیر پیش‌بینی شده	برچسب واقعی
+	۰/۰۴۷۴	+
+	۰/۰۰۲۵	+
+	۰/۰۰۰۱	+
+	۰/۱۱۹۲	+
+	۰/۰۰۶۷	+

+	۰/۰۰۰۳	+
+	۰/۲۶۸۹	+
+	۰/۰۱۸۰	+
+	۰/۰۰۰۹	+
+	۰/۰۰۰۱	+
−	۰/۹۹۹۱	−
+	۰/۷۳۱۱	−
+	۰/۹۸۲۰	−
−	۰/۹۹۹۱	−
+	۰/۵۰۰۰	−
+	۰/۹۵۲۶	−
−	۰/۹۹۷۶	−
+	۰/۲۶۸۹	−
+	۰/۸۸۰۸	−
+	۰/۹۹۳۳	−

یعنی:

$$FP = \frac{\text{تعداد برچسب‌های مثبت پیش‌بینی شده غلط}}{\text{تعداد برچسبهای منفی}} = \frac{7}{10} = 70\%$$

$$TP = \frac{\text{تعداد برچسب‌های مثبت پیش‌بینی شده درست}}{\text{تعداد برچسبهای مثبت}} = \frac{10}{10} = 100\%$$

سپس سطح رده مثبت را (مقدار $0.9933 <$) در نظر می‌گیریم. پس برچسب‌های پیش‌بینی شده به شکل زیر خواهند بود.

برچسب پیش‌بینی شده	مقادیر پیش‌بینی شده	برچسب واقعی
+	۰/۰۴۷۴	+
+	۰/۰۰۲۵	+
+	۰/۰۰۰۱	+
+	۰/۱۱۹۲	+
+	۰/۰۰۶۷	+
+	۰/۰۰۰۳	+
+	۰/۲۶۸۹	+
+	۰/۰۱۸۰	+
+	۰/۰۰۰۹	+
+	۰/۰۰۰۱	+
−	۰/۹۹۹۱	−

+	۰/۷۳۱۱	–
+	۰/۹۸۲۰	–
–	۰/۹۹۹۱	–
+	۰/۵۰۰۰	–
+	۰/۹۵۲۶	–
–	۰/۹۹۷۶	–
+	۰/۲۶۸۹	–
+	۰/۸۸۰۸	–
–	۰/۹۹۳۳	–

یعنی:

$$FP = \frac{\text{تعداد برچسب‌های مثبت پیش‌بینی شده غلط}}{\text{تعداد برچسبهای منفی}} = \frac{6}{10} = 60\%$$

$$TP = \frac{\text{تعداد برچسب‌های مثبت پیش‌بینی شده درست}}{\text{تعداد برچسبهای مثبت}} = \frac{10}{10} = 100\%$$

سپس سطح رده مثبت را (0.9820 > مقدار) در نظر می‌گیریم. پس برچسب‌های پیش‌بینی شده به شکل زیر خواهند بود.

برچسب پیش‌بینی شده	مقادیر پیش‌بینی شده	برچسب واقعی
+	۰/۰۴۷۴	+
+	۰/۰۰۲۵	+
+	۰/۰۰۰۱	+
+	۰/۱۱۹۲	+
+	۰/۰۰۶۷	+
+	۰/۰۰۰۳	+
+	۰/۲۶۸۹	+
+	۰/۰۱۸۰	+
+	۰/۰۰۰۹	+
+	۰/۰۰۰۱	+
–	۰/۹۹۹۱	–
+	۰/۷۳۱۱	–
–	۰/۹۸۲۰	–
–	۰/۹۹۹۱	–
+	۰/۵۰۰۰	–
+	۰/۹۵۲۶	–
–	۰/۹۹۷۶	–

+	۰/۲۶۸۹	–
+	۰/۸۸۰۸	–
–	۰/۹۹۳۳	–

یعنی:

$$FP = \frac{\text{تعداد برچسب‌های مثبت پیش‌بینی شده غلط}}{\text{تعداد برچسبهای منفی}} = \frac{5}{10} = 50\%$$

$$TP = \frac{\text{تعداد برچسب‌های مثبت پیش‌بینی شده درست}}{\text{تعداد برچسبهای مثبت}} = \frac{10}{10} = 100\%$$

سپس سطح رده مثبت را ($0.9526 >$ مقدار) در نظر می‌گیریم. پس برچسب‌های پیش‌بینی شده به شکل زیر خواهند بود.

برچسب پیش‌بینی شده	مقادیر پیش‌بینی شده	برچسب واقعی
+	۰/۰۴۷۴	+
+	۰/۰۰۲۵	+
+	۰/۰۰۰۱	+
+	۰/۱۱۹۲	+
+	۰/۰۰۶۷	+
+	۰/۰۰۰۳	+
+	۰/۲۶۸۹	+
+	۰/۰۱۸۰	+
+	۰/۰۰۰۹	+
+	۰/۰۰۰۱	+
–	۰/۹۹۹۱	–
+	۰/۷۳۱۱	–
–	۰/۹۸۲۰	–
–	۰/۹۹۹۱	–
+	۰/۵۰۰۰	–
–	۰/۹۵۲۶	–
–	۰/۹۹۷۶	–
+	۰/۲۶۸۹	–
+	۰/۸۸۰۸	–
–	۰/۹۹۳۳	–

یعنی:

$$FP = \frac{\text{تعداد برچسب‌های مثبت پیش‌بینی شده غلط}}{\text{تعداد برچسبهای منفی}} = \frac{4}{10} = 40\%$$

$$TP = \frac{\text{تعداد برچسب‌های مثبت پیش‌بینی شده درست}}{\text{تعداد برچسبهای مثبت}} = \frac{10}{10} = 100\%$$

سپس سطح رده مثبت را (0.8808 > مقدار) در نظر می‌گیریم. پس برچسب‌های پیش‌بینی شده به شکل زیر خواهند بود.

برچسب پیش‌بینی شده	مقادیر پیش‌بینی شده	برچسب واقعی
+	۰/۰۴۷۴	+
+	۰/۰۰۲۵	+
+	۰/۰۰۰۱	+
+	۰/۱۱۹۲	+
+	۰/۰۰۶۷	+
+	۰/۰۰۰۳	+
+	۰/۲۶۸۹	+
+	۰/۰۱۸۰	+
+	۰/۰۰۰۹	+
+	۰/۰۰۰۱	+
−	۰/۹۹۹۱	−
+	۰/۷۳۱۱	−
−	۰/۹۸۲۰	−
−	۰/۹۹۹۱	−
+	۰/۵۰۰۰	−
−	۰/۹۵۲۶	−
−	۰/۹۹۷۶	−
+	۰/۲۶۸۹	−
−	۰/۸۸۰۸	−
−	۰/۹۹۳۳	−

یعنی:

$$FP = \frac{\text{تعداد برچسب‌های مثبت پیش‌بینی شده غلط}}{\text{تعداد برچسبهای منفی}} = \frac{3}{10} = 30\%$$

$$TP = \frac{\text{تعداد برچسب‌های مثبت پیش‌بینی شده درست}}{\text{تعداد برچسبهای مثبت}} = \frac{10}{10} = 100\%$$

سپس سطح رده مثبت را (0.7311 > مقدار) در نظر می‌گیریم. پس برچسب‌های پیش‌بینی شده به شکل زیر خواهند بود.

برچسب پیش‌بینی شده	مقادیر پیش‌بینی شده	برچسب واقعی
+	۰/۰۴۷۴	+

+	۰/۰۰۲۵	+
+	۰/۰۰۰۱	+
+	۰/۱۱۹۲	+
+	۰/۰۰۶۷	+
+	۰/۰۰۰۳	+
+	۰/۲۶۸۹	+
+	۰/۰۱۸۰	+
+	۰/۰۰۰۹	+
+	۰/۰۰۰۱	+
−	۰/۹۹۹۱	−
−	۰/۷۳۱۱	−
−	۰/۹۸۲۰	−
−	۰/۹۹۹۱	−
+	۰/۵۰۰۰	−
−	۰/۹۵۲۶	−
−	۰/۹۹۷۶	−
+	۰/۲۶۸۹	−
−	۰/۸۸۰۸	−
−	۰/۹۹۳۳	−

یعنی:

$$FP = \frac{\text{تعداد برچسب‌های مثبت پیش‌بینی شده غلط}}{\text{تعداد برچسبهای منفی}} = \frac{2}{10} = 20\%$$

$$TP = \frac{\text{تعداد برچسب‌های مثبت پیش‌بینی شده درست}}{\text{تعداد برچسبهای مثبت}} = \frac{10}{10} = 100\%$$

سپس سطح رده مثبت را ($0.5000 <$ مقدار) در نظر می‌گیریم. پس برچسب‌های پیش‌بینی شده به شکل زیر خواهند بود.

برچسب پیش‌بینی شده	مقادیر پیش‌بینی شده	برچسب واقعی
+	۰/۰۴۷۴	+
+	۰/۰۰۲۵	+
+	۰/۰۰۰۱	+
+	۰/۱۱۹۲	+
+	۰/۰۰۶۷	+
+	۰/۰۰۰۳	+
+	۰/۲۶۸۹	+

+	۰/۰۱۸۰	+
+	۰/۰۰۰۹	+
+	۰/۰۰۰۱	+
–	۰/۹۹۹۱	–
–	۰/۷۳۱۱	–
–	۰/۹۸۲۰	–
–	۰/۹۹۹۱	–
–	۰/۵۰۰۰	–
–	۰/۹۵۲۶	–
–	۰/۹۹۷۶	–
+	۰/۲۶۸۹	–
–	۰/۸۸۰۸	–
–	۰/۹۹۳۳	–

یعنی:

$$FP = \frac{\text{تعداد برچسب‌های مثبت پیش‌بینی شده غلط}}{\text{تعداد برچسبهای منفی}} = \frac{1}{10} = 10\%$$

$$TP = \frac{\text{تعداد برچسب‌های مثبت پیش‌بینی شده درست}}{\text{تعداد برچسبهای مثبت}} = \frac{10}{10} = 100\%$$

سپس سطح رده مثبت را ($0.2689 <$ مقدار) در نظر می‌گیریم. پس برچسب‌های پیش‌بینی شده به شکل زیر خواهند بود.

برچسب پیش‌بینی شده	مقادیر پیش‌بینی شده	برچسب واقعی
+	۰/۰۴۷۴	+
+	۰/۰۰۲۵	+
+	۰/۰۰۰۱	+
+	۰/۱۱۹۲	+
+	۰/۰۰۶۷	+
+	۰/۰۰۰۳	+
–	۰/۲۶۸۹	+
+	۰/۰۱۸۰	+
+	۰/۰۰۰۹	+
+	۰/۰۰۰۱	+
–	۰/۹۹۹۱	–
–	۰/۷۳۱۱	–
–	۰/۹۸۲۰	–

–	۰/۹۹۹۱	–
–	۰/۵۰۰۰	–
–	۰/۹۵۲۶	–
–	۰/۹۹۷۶	–
–	۰/۲۶۸۹	–
–	۰/۸۸۰۸	–
–	۰/۹۹۳۳	–

یعنی:

$$FP = \frac{\text{تعداد برچسب‌های مثبت پیش‌بینی شده غلط}}{\text{تعداد برچسبهای منفی}} = \frac{0}{10} = 0\%$$

$$TP = \frac{\text{تعداد برچسب‌های مثبت پیش‌بینی شده درست}}{\text{تعداد برچسبهای مثبت}} = \frac{9}{10} = 90\%$$

سپس سطح رده مثبت را ($0.1192 <$ مقدار) در نظر می‌گیریم:

$$FP = \frac{0}{10} = 0\%$$

$$TP = \frac{8}{10} = 80\%$$

سپس سطح رده مثبت را ($0.0474 <$ مقدار) در نظر می‌گیریم:

$$FP = \frac{0}{10} = 0\%$$

$$TP = \frac{7}{10} = 70\%$$

سپس سطح رده مثبت را ($0.0180 <$ مقدار) در نظر می‌گیریم:

$$FP = \frac{0}{10} = 0\%$$

$$TP = \frac{6}{10} = 60\%$$

سپس سطح رده مثبت را ($0.0067 <$ مقدار) در نظر می‌گیریم:

$$FP = \frac{0}{10} = 0\%$$

$$TP = \frac{5}{10} = 50\%$$

سپس سطح رده مثبت را ($0.0025 <$ مقدار) در نظر می‌گیریم:

$$FP = \frac{0}{10} = 0\%$$

$$TP = \frac{4}{10} = 40\%$$

سپس سطح رده مثبت را ($0.0009 <$ مقدار) در نظر می‌گیریم:

$$FP = \frac{0}{10} = 0\%$$

$$TP = \frac{3}{10} = 30\%$$

سپس سطح رده مثبت را ($0.0003 <$ مقدار) در نظر می‌گیریم:

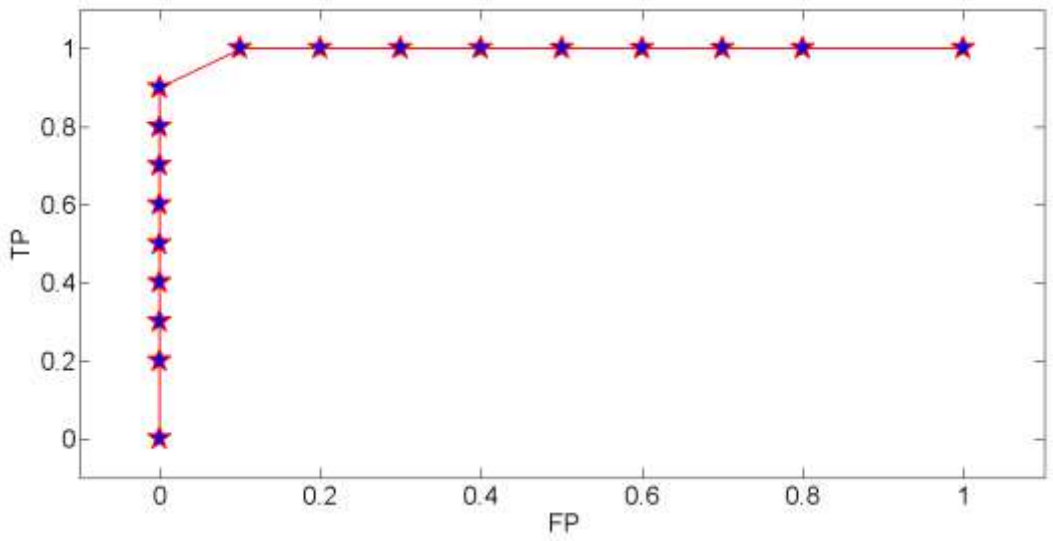
$$FP = \frac{0}{10} = 0\%$$

$$TP = \frac{2}{10} = 20\%$$

سپس سطح رده مثبت را ($0.0001 <$ مقدار) در نظر می‌گیریم:

$$FP = \frac{0}{10} = 0\%$$

$$TP = \frac{0}{10} = 0\%$$



۲۰- نمونه‌های آموزشی زیر را در نظر بگیرید.

X	y	z	تعداد نمونه‌های کلاس C1	تعداد نمونه‌های کلاس C2
0	0	0	5	40

0	0	1	0	15
0	1	0	10	5
0	1	1	45	0
1	0	0	10	5
1	0	1	25	0
1	1	0	5	20
1	1	1	0	15

در درخت تصمیم به دست آمده، آنتروپی هر یک از گره‌های برگ را محاسبه نمایید. (۳۰ نمره)

۲۱- مجموعه داده‌های زیر را در نظر بگیرید.

Tid	Refund	Marital Status	Taxable Income	Evad
1	Yes	Single	125k	No
2	No	Married	100k	No
3	No	Single	70k	No
4	Yes	Married	120k	No
5	No	Divorced	95k	Yes
6	No	Married	60k	No
7	Yes	Divorced	220k	No
8	No	Single	85k	Yes
9	No	Married	75k	No
10	No	Single	90k	Yes

فرض کنید احتمال تعلق هر صفت پیوسته Taxable Income به هر یک از کلاس‌های Yes و No دارای توزیع نرمال باشد. همچنین فرض کنید هر یک از صفت‌ها برای هر رده مفروض مستقل از هم باشند. همچنین رکورد تست زیر را در نظر بگیرید.

X=(Refund=Yes, Marital Status=Married, Income=80)

با استفاده از رده‌بند Naïve Bayes رده این نمونه تست را تعیین نمایید. (۲۰ نمره)

۲۲- Feature Subset Selection چیست؟ چه رهیافت‌هایی برای انجام آن روی داده‌ها وجود دارد؟ تفاوت آن با Dimension Reduction چیست؟ (۵ نمره)

۲۳- وظایف اصلی داده‌کاوی را نام برده و به صورت مختصر و مفید شرح دهید. (۱۰ نمره)

۲۴- مسئله "نفرین ابعاد" یا "مشکل ابعاد" یا "Curse of Dimensionality" چه مسئله‌ای است؟ (۵ نمره)

۲۵- شباهت و تفاوت شبکه عصبی مصنوعی (ANN) با ماشین بردار پشتیبان (SVM) در چیست؟ توضیح دهید. (۵ نمره)

۲۶- داده‌های مقابل را در نظر بگیرید. بهترین حد آستانه را برای gain حداکثر برای ویژگی X (بر اساس آنتروپی) به دست آورید. (۱۶ نمره)

Number	X	Y	Class
1	15	4	C1
2	20	2	C2
3	25	1	C1
4	30	3	C1
5	35	2	C2
6	25	2	C1
7	15	9	C2
8	20	6	C2

۲۷- داده‌های زیر را در نظر بگیرید. الگوریتم k-means را روی این داده‌ها تا حداکثر ۳ تکرار اعمال کنید. نمونه‌های ۱، ۲ و ۴ را به عنوان مراکز اولیه در نظر بگیرید. روی نتیجه خوشه‌بندی به دست آمده از الگوریتم k-means، خطای کل (یا SSE) را محاسبه نمایید. (۲۰ نمره)

Number	X	Y	Z	Class
1	-2	4	0	1
2	-2	2	0	1
3	-3	3	1	1
4	-10	3	-10	2
5	-8	2	-8	2
6	-8	4	-9	2
7	-10	9	10	3
8	-8	9	8	3
9	-9	10	8	3

۲۸- با استفاده از ماتریس مجاورت زیر نقاط داده شده را طبق الگوریتم Average_Linkage، به طور کامل به صورت سلسله مراتبی خوشه‌بندی کرده و درخت dendrogram آن را رسم نمایید. در هر مرحله از الگوریتم ماتریس مجاورت به روز رسانی شده را به دست آورید. (۳۰ نمره)

Points	P1	P2	P3	P4	P5	P6
P1	1	0.7	0.35	0.9	0.2	0.5
P2	0.7	1	0.55	0.45	0.1	0.8
P3	0.35	0.55	1	0.6	0.3	0.2
P4	0.9	0.45	0.6	1	0.75	0.4
P5	0.2	0.1	0.3	0.75	1	0.85
P6	0.5	0.8	0.2	0.4	o.85	1

۲۹- OverTraining و UnderTraining را توضیح دهید. (۶ نمره)

۳۰- یک مجموعه داده با دو ویژگی X_1 و X_2 و برچسب y را در نظر بگیرید. ماتریس کواریانس زیر بین ویژگی‌ها و برچسب را در نظر بگیرید.

	x_1	x_2	y
x_1	a	d	e
x_2	d	b	f
y	e	f	c

الف) فرض کنید d عددی بسیار بزرگ در مقایسه با سایر اعداد این ماتریس است. آیا می‌توان نتیجه گرفت که با حذف یکی از این ویژگی‌ها کارایی رده‌بندی افت نخواهد کرد؟ توضیح دهید.

ب) فرض کنید d عددی بسیار کوچک در مقایسه با سایر اعداد این ماتریس است. آیا می‌توان نتیجه گرفت که با حذف یکی از این ویژگی‌ها کارایی رده‌بندی افت نخواهد کرد؟ توضیح دهید.

ج) فرض کنید d صفر است. آیا می‌توان نتیجه گرفت که با حذف یکی از این ویژگی‌ها کارایی رده‌بندی افت نخواهد کرد؟ توضیح دهید.

د) فرض کنید e صفر است. آیا می‌توان نتیجه گرفت که با حذف یکی از این ویژگی‌ها کارایی رده‌بندی افت نخواهد کرد؟ توضیح دهید.

ذ) فرض کنید f صفر است. آیا می‌توان نتیجه گرفت که با حذف یکی از این ویژگی‌ها کارایی رده‌بندی افت نخواهد کرد؟ توضیح دهید.

ه) فرض کنید c صفر است. آیا می‌توان نتیجه گرفت که با حذف یکی از این ویژگی‌ها کارایی رده‌بندی افت نخواهد کرد؟ توضیح دهید.

جواب

الف) درست- کوواریانس بالا یعنی همبستگی زیاد. همبستگی زیاد بین دو صفت یعنی یکی از صفات اضافه است.

ب) درست (چرا که ممکن است منفی باشد و اندازه آن بسیار بزرگ) اندازه همبستگی زیاد بین دو صفت یعنی یکی از صفات اضافه است.

ج) غلط- اندازه کوواریانس کم یعنی همبستگی کم و عدم ارتباط بین آن دو صفت. همبستگی کم بین دو صفت یعنی هیچ یک از صفات اضافه نیست.

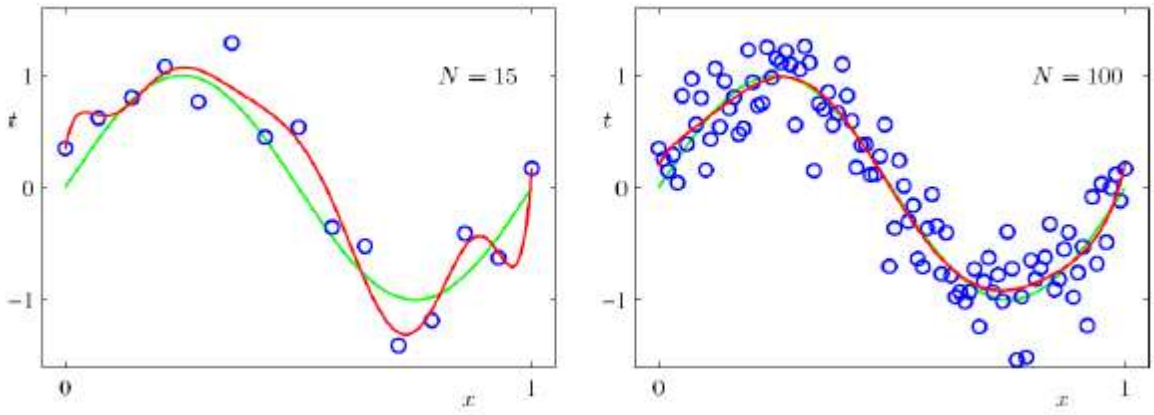
د) درست- اندازه کوواریانس کم یعنی همبستگی کم و عدم ارتباط بین صفت x_1 و برچسب. پس حذف x_1 صفت بی تاثیر است.

ذ) درست- به دلیل قسمت د، این بار صفت x_2 بی تاثیر است.

ه) درست- اگر c صفر است، پس برچسب فقط یک حالت دارد. یعنی مساله تک رده‌ای است. پس حذف هر دو صفت نیز در کارایی رده‌بندی بی تاثیر است.

۳۱- آیا ارتباطی بین اندازه داده‌ها و پدیده یادگیری بیش از حد (Overfitting) وجود دارد؟ با مثال توضیح دهید.

جواب) بله- نمونه‌های کم باعث کم شدن تعمیم و کم عمق شدن اطلاعات لازم برای یادگیری می‌شود. پس با یادگیری زیاد داده-های کم، پدیده یادگیری بیش از حد رخ می‌دهد.



۳۲- فرض کنید از مجموعه داده زیر یک ویژگی را می‌خواهیم حذف کنیم. کدام یک از ۳ ویژگی زیر را حذف کنیم. چرا؟

X1	X2	X3	Y
0	0	0	0
0	0	1	1
0	1	0	0
0	1	1	1
1	0	0	1
1	0	1	0
1	1	0	1
1	1	1	0

جواب) با به دست آوردن ماتریس کوواریانس خواهیم فهمید که هیچ صفتی بر دیگری ارجحیت ندارد. پس هر صفتی را می‌توان حذف کرد.

$$\text{Cov} = \begin{bmatrix} 0.25 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.25 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.25 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.25 \end{bmatrix}$$

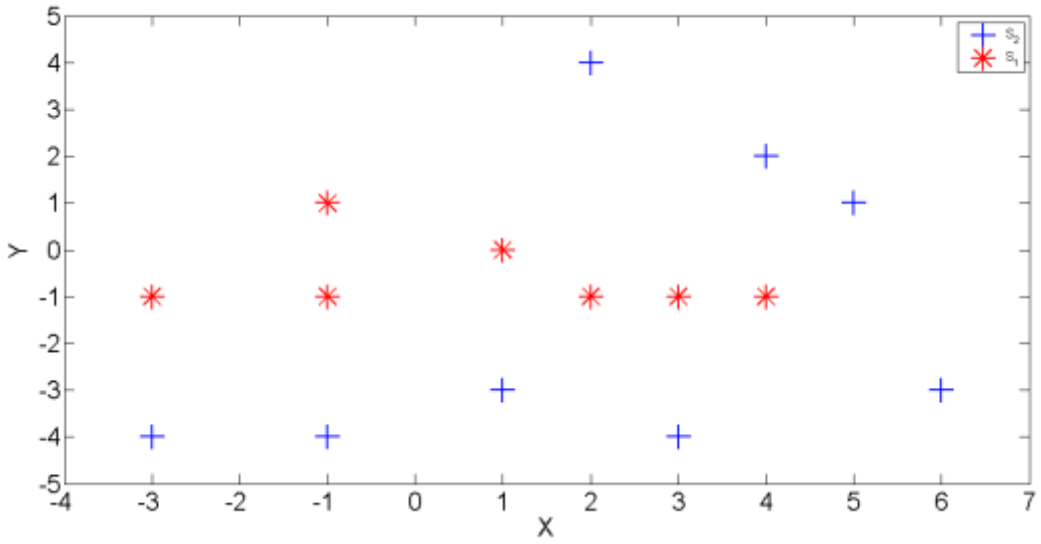
۳۳- یک شبکه عصبی MLP برای داده‌های زیر ترسیم کنید به گونه‌ای که روی این داده‌ها خطا نداشته باشد. وزن‌ها را به صورت دستی تنظیم کنید.

S_1: (-3, -1), (-1, -1), (-1, 1), (1, 0), (2, -1), (3, -1), (4, -1)

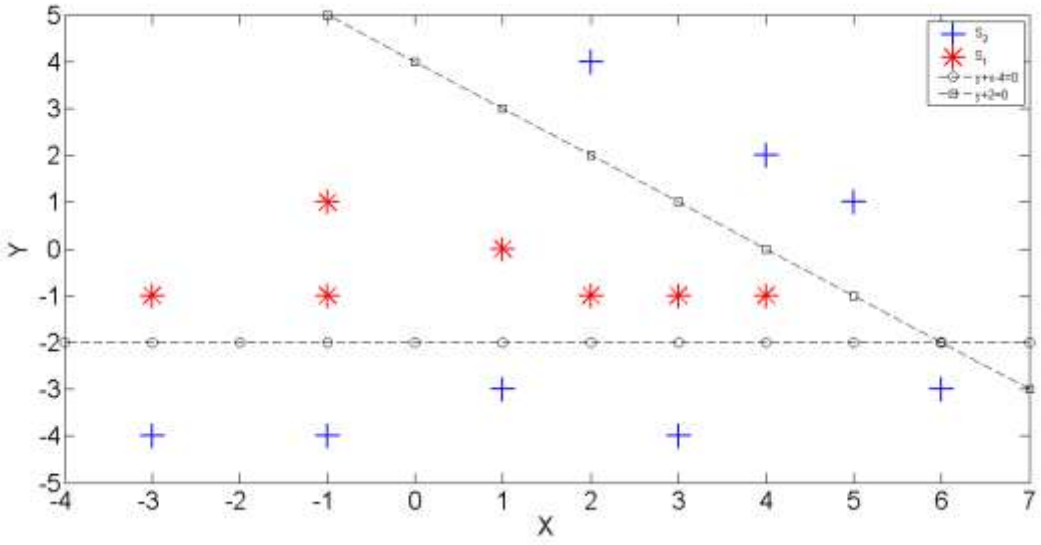
S_2: (-3, -4), (-1, -4), (1, -3), (1, -3), (2, 4), (3, -4), (4, 2), (5, 1), (6, -3)

جواب

فضای داده‌های دو رده در شکل زیر نشان داده شده است.



دو خط $y + x - 4 = 0$ و $y + 2 = 0$ لازم است تا رده * را تشخیص داد.

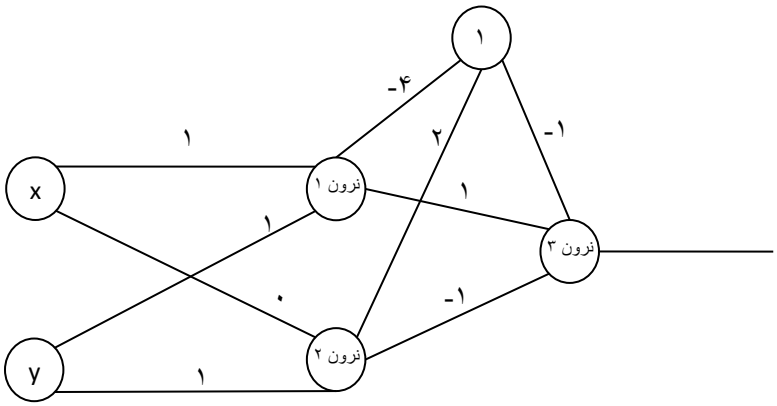


رده * به شکل زیر تشخیص داده می‌شود:

$y + x - 4 < 0$

$y + 2 \geq 0$

یعنی اگر هر دو شرط بالا بر قرار باشد، رده * است و در غیر این صورت، رده + است. پس از یک شبکه به شکل زیر استفاده می-کنیم.



هر سه نرون از تابع فعالیت sgn(x) که به شکل زیر تعریف می‌شوند، استفاده می‌کنند.

$$\text{sgn}(x) = \begin{cases} 1 & x \geq 0 \\ -1 & x < 0 \end{cases}$$

نرون ۱ تفکیک خطی $y + x - 4 \geq 0$ را انجام می‌دهد. نرون ۲ تفکیک خطی $y + 2 \geq 0$ را انجام می‌دهد. نرون ۳ نیز در صورتی ۱ را در خروجی می‌برد که خروجی نرون ۱، ۱- باشد و خروجی نرون ۲، ۱ باشد.

در نهایت این شبکه اگر خروجی ۱ تولید کرد، به معنی رده * است و اگر خروجی ۱- تولید کرد، به معنی رده + است.

۳۴- دو متغیر تصادفی X و Y را در نظر بگیرید. فرض کنید که μ و σ به ترتیب میانگین و انحراف معیار را نشان دهند. فرض کنید که μ_X و σ_X به ترتیب میانگین و انحراف معیار X را نشان دهند. رابطه متقابل بین X و Y را با یکی از روش‌های زیر نمایش می-دهند.

I) همبستگی:

$$Cov(X,Y) = E\left((X - \mu_X)(Y - \mu_Y)\right) = E(XY) - \mu_X\mu_Y$$

II) ضریب همبستگی:

$$\rho_{XY} = \frac{Cov(X,Y)}{\sigma_X\sigma_Y}$$

III) اطلاعات متقابل:

$$MI(X,Y) = H(X) - H(X|Y) = KL\big(P(X,Y)||P(X)P(Y)\big)$$

IV) اطلاعات متقابل نرمال شده:

$$NMI(X,Y) = \frac{MI(X,Y)}{H(Y)}$$

الف) ثابت کنید اندازه ضریب همبستگی حداکثر یک است. نکته: $E(XY)^2 \leq E(X^2)E(Y^2)$.

ب) با مثال بگویید چه موقع همبستگی یک می‌شود. با مثال بگویید چه موقع همبستگی منفی یک می‌شود.

ت) ثابت کنید اطلاعات متقابل نرمال شده حداکثر یک است. نکته: $H(X) = \int p(X)\log\big(p(X)\big)\,dX$ آنتروپی متغیر تصادفی X است.

پ) با مثال بگویید اطلاعات متقابل نرمال شده چه موقع یک می‌شود.

ج) اگر ضریب هبستگی صفر باشد، آیا اطلاعات متقابل نرمال شده صفر می‌شود. اثبات کنید یا مثال نقض بیاورید.

چ) اگر اطلاعات متقابل نرمال شده صفر باشد، آیا ضریب هبستگی صفر می‌شود. اثبات کنید یا مثال نقض بیاورید.

د) اگر اطلاعات متقابل نرمال شده یک باشد، آیا ضریب هبستگی یک می‌شود. اثبات کنید یا مثال نقض بیاورید.

جواب

الف)

$$\rho_{XY} = \frac{Cov(X,Y)}{\sigma_X\sigma_Y} = \frac{E\left((X-\mu_X)(Y-\mu_Y)\right)}{\sigma_X\sigma_Y} = \frac{E(XY)-\mu_X\mu_Y}{\sigma_X\sigma_Y}$$

$$|\rho_{XY}|^2 = \left|\frac{E(XY)-\mu_X\mu_Y}{\sigma_X\sigma_Y}\right|^2 = \frac{E(XY)^2 - 2E(XY)\mu_X\mu_Y + \mu_X^2\mu_Y^2}{\sigma_X^2\sigma_Y^2}$$
$$\leq \frac{E(X^2)E(Y^2) - 2E(XY)E(X)E(Y) + E(X)^2E(Y)^2}{(E(X^2)-E(X)^2)(E(Y^2)-E(Y)^2)}$$
$$= \frac{E(X^2)E(Y^2) + E(X)^2E(Y)^2 - 2E(XY)E(X)E(Y)}{E(X^2)E(Y^2) + E(X)^2E(Y)^2 - E(X^2)E(Y)^2 - E(Y^2)E(X)^2}$$
$$= \frac{E(X^2)E(Y^2) + E(X)^2E(Y)^2 - E(X^2)E(Y)^2 - E(Y^2)E(X)^2}{E(X^2)E(Y^2) + E(X)^2E(Y)^2 - E(X^2)E(Y)^2 - E(Y^2)E(X)^2}$$
$$+ \frac{E(X^2)E(Y)^2 + E(Y^2)E(X)^2 - 2E(XY)E(X)E(Y)}{E(X^2)E(Y^2) + E(X)^2E(Y)^2 - E(X^2)E(Y)^2 - E(Y^2)E(X)^2}$$
$$= 1 - \frac{2E(XY)\mu_X\mu_Y - E(X^2)\mu_Y^2 - E(Y^2)\mu_X^2}{E(X^2)E(Y^2) + E(X)^2E(Y)^2 - E(X^2)E(Y)^2 - E(Y^2)E(X)^2}$$
$$= 1 - \frac{2E\left((\mu_Y X)(\mu_X Y)\right) - E\left((\mu_X Y)^2\right) - E\left((\mu_Y X)^2\right)}{E(X^2)E(Y^2) + E(X)^2E(Y)^2 - E(X^2)E(Y)^2 - E(Y^2)E(X)^2}$$
$$= 1 - \frac{\left(E\left((\mu_Y X)(\mu_X Y)\right) - E\left((\mu_X Y)^2\right)\right) + \left(E\left((\mu_Y X)(\mu_X Y)\right) - E\left((\mu_Y X)^2\right)\right)}{E(X^2)E(Y^2) + E(X)^2E(Y)^2 - E(X^2)E(Y)^2 - E(Y^2)E(X)^2}$$

فرض کنید تغییر متغیر زیر را انجام دهیم:

$$x = \mu_Y X$$

$$y = \mu_X Y$$

با جایگذاری متغیرهای بالا، معادله بالا به شکل زیر خواهد بود:

$$|\rho_{XY}|^2 \leq 1 - \left(E(xy) - E(y^2) + E(xy) - E(x^2)\right)$$

از طرفی داریم:

$$E(xy) - E(x^2) = \int x \int yp(x,y)dy\,dx - \int xxp(x)dx$$
$$= \int x \int yp(x,y)dy\,dx - \int x \int xp(x,y)dy\,dx = \int x \int (y-x)p(x,y)dy\,dx$$
$$= E\big(x(y-x)\big)$$

$$E(xy) - E(y^2) = E\big(y(x-y)\big) = E\big((-y)(y-x)\big)$$

با جایگذاری روابط بالا، معادله بالا به شکل زیر خواهد بود:

$$|\rho_{XY}|^2 \leq 1 - E\big(x(x-y) - y(x-y)\big) = 1 - E((x-y)^2)$$

از آنجایی که:

$$E((x-y)^2) \geq 0 \rightarrow -E((x-y)^2) \leq 0 \rightarrow 1 - E((x-y)^2) \leq 1$$

پس:

$$|\rho_{XY}|^2 \leq 1$$

ب)

همبستگی وقتی یک می‌شود که یکی از ویژگی‌ها با یک رابطه خطی با شیب مثبت نسبت به دیگری به دست آید. مثلا همبستگی دو بردار زیر یک است.

1	3
2	5
6	13
0	1
-1	-1

همبستگی وقتی منفی یک می‌شود که یکی از ویژگی‌ها با یک رابطه خطی با شیب منفی نسبت به دیگری به دست آید. مثلا همبستگی دو بردار زیر منفی یک است.

1	-3.1
2	-3.2
6	-3.6
0	-3
-1	-2.9

ج)

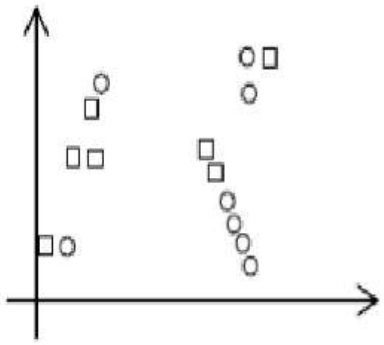
$$MI(X,Y) = H(X) - H(X|Y) = \int \big(p(X)\log\big(p(X)\big) - p(X|Y)\log\big(p(X|Y)\big)\big)dX$$
$$= \int \left(p(X)\log\big(p(X)\big) - \frac{p(X \cap Y)}{P(Y)}\log\left(\frac{p(X \cap Y)}{P(Y)}\right)\right)dX =$$

۳۵- نمونه‌های آموزشی زیر را در نظر بگیرید.

X	y	z	تعداد نمونه‌های کلاس C1	تعداد نمونه‌های کلاس C2
0	0	0	5	40
0	0	1	0	15
0	1	0	10	5
0	1	1	45	0
1	0	0	10	5
1	0	1	25	0
1	1	0	5	20
1	1	1	0	15

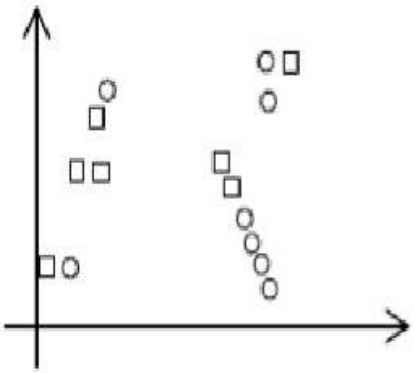
از آستانه هرس با خلوص ۰,۸ استفاده کنید.

۳۶- دقت k-NN را بر روی مجموعه داده زیر محاسبه کنید و جدول زیر را پر کنید. با فرض اینکه از تکنیک one-leave-out استفاده می‌شود. در صورت رخداد حالت برابر (Tie) بین دو رده، خطا را ۰,۵ (نیم) واحد در نظر بگیرید.



k	k-nearest neighbor accuracy
1	
2	
3	

جواب



k	k-nearest neighbor accuracy
1	9/15
2	8.5/15
3	8/15

۳۷- مجموعه داده‌هایی را در نظر بگیرید که خلاصه آن در جدول زیر را آورده شده است.

X	y	z	تعداد نمونه‌های کلاس C1	تعداد نمونه‌های کلاس C2
a	0	0	5	40
a	0	1	0	15
b	1	0	10	5
b	1	1	45	0
c	0	0	10	5
c	0	1	25	0
b	1	0	5	20
b	1	1	0	15

بین صفات x و y با چه اطمینانی همبستگی وجود ندارد؟ فرض کنید جدول زیر را داریم.

درجه	سطح اطمینان	سطح قبول برای وابستگی
۱	۰,۱	۲,۷۱
۲	۰,۱	۴,۶۱
۳	۰,۱	۶,۲۵
۴	۰,۱	۷,۷۸
۱	۰,۰۱	۶/۶۳
۲	۰,۰۱	۹/۲۱
۳	۰,۰۱	۱۱/۳۴
۴	۰,۰۱	۱۳/۲۸
۱	۰,۰۰۱	۱۰/۸۳
۲	۰,۰۰۱	۱۳/۸۲
۳	۰,۰۰۱	۱۶/۲۷
۴	۰,۰۰۱	۱۸/۴۷
۱	۰,۰۰۰۱	۱۵/۱۴
۲	۰,۰۰۰۱	۱۸/۴۲
۳	۰,۰۰۰۱	۲۱/۱۱
۴	۰,۰۰۰۱	۲۳/۵۱
۱	۰,۰۰۰۰۱	۱۹/۵۱
۲	۰,۰۰۰۰۱	۲۳/۰۳
۱	۰,۰۰۰۰۰۱	۲۳/۹۳
۲	۰,۰۰۰۰۰۱	۲۷/۶۳
۱	۰,۰۰۰۰۰۰۱	۲۸/۳۷
۲	۰,۰۰۰۰۰۰۱	۳۲/۲۴

جواب)

در اولین گام، جدول زیر را به دست می‌آوریم.

	$x = a$	$x = b$	$x = c$	*Total*
$y = 0$	*60*	*0*	*40*	*100*
$y = 1$	*0*	*100*	*0*	*100*
Total	*60*	*100*	*40*	*200*

سپس فراوانی مورد انتظار برای هر یک از درایه‌های جدول را با استفاده از معادله زیر به دست می‌آوریم.

$$e_{ij} = \frac{count(A = a_i) \times count(B = b_j)}{N}$$

با استفاده از فرمول بالا می توان فراوانی مورد انتظار برای هر خانه جدول را محاسبه کرد. برای مثال فراوانی مورد انتظار برای خانه $(0,a)$ در زیر محاسبه شده است.

$$e_{11} = \frac{count(0) \times count(a)}{N} = \frac{100 \times 60}{200} = 30$$

توجه کنید که مجموع فراوانی های مورد انتظار در هر سطر مساوی با تعداد کل فراوانی مشاهده شده برای آن سطر و مجموع فراوانی های مورد انتظار در هر ستون مساوی با تعداد کل فراوانی های مشاهده شده برای آن ستون است.

جدول زیر فراوانی مورد انتظار درایه ها است.

	$x = a$	$x = b$	$x = c$	*Total*
$y = 0$	*30*	*50*	*20*	*100*
$y = 1$	*30*	*50*	*20*	*100*
Total	*60*	*100*	*40*	*200*

حال با استفاده از معادله زیر χ^2 را محاسبه می‌نماییم.

$$\chi^2 = \sum_{i=1}^{c}\sum_{j=1}^{r}\frac{(o_{ij}-e_{ij})^2}{e_{ij}} = 2*\frac{(60-30)^2}{30} + 2*\frac{(100-50)^2}{50} + 2*\frac{(40-20)^2}{20}$$
$$= 60 + 100 + 40 = 200$$

برای این جدول 2×3، درجه آزادی برابر با ۲= (۱-۳)(۱-۲) است. برای ۲ درجه آزادی، مقدار χ^2 مورد نیاز برای رد این فرضیه در هر سطحی، مقدار محاسبه شده از این مقدار بالاتر است. پس ما این فرضیه که دو صفت *gender* و *preferred_Reading* مستقل هستند را رد و نتیجه می گیریم که دو صفت ذکر شده برای این گروه افراد کاملاً وابسته است. پس سطح اطمینان وابستگی این دو برابر زیر است:

$$1 - 0.0000001 = 0.9999999$$

دوره جامع آموزش برنامه نویسی متلب

MATLAB