به نام خدا

تکلیف دوم درس داده کاوی

ترم دوم 1401-1400

- **معرفی دیتاست Churn**: ریزش مشتری یک اصطلاح مالی است که به از دست دادن مشتری برای یک شرکت یا کسب و کار اشاره دارد. نرخ ریزش بالاتر از یک آستانه خاص می تواند تأثیرات ملموس و نامحسوسی در موفقیت تجاری شرکت داشته باشد. در حالت ایده‌آل، شرکت‌ها دوست دارند تا آنجا که می‌توانند مشتری خود را حفظ کنند. هر نمونه در مجموعه‌ی داده شامل 20 ویژگی و یک متغیر بولی "Churn?" است.

قسمت اول: تمرینات عملی با پایتون

برای تمرینات زیر، از مجموعه داده‌های churn استفاده کنید.

1. خلاصه‌سازی داده‌ها

1.1. با استفاده از کتابخانه‌ی pandas دیتاست را به دیتافریم تبدیل کرده و سپس اطلاعات کلی در مورد این دیتاست و ویژگی‌های آن را نمایش دهید.

1.2. مقادیر یکتای ویژگی‌های دسته‌ای دیتاست را همراه با تعداد موجود در هر دسته نمایش دهید.

2. دسته‌بندی داده‌ها

2.1. داده‌های موجود در دیتاست را با استفاده از متد groupby و بر اساس مقدار Gender دسته‌بندی کنید و برای نمایش داده‌ها در هر دسته از میانگین استفاده کنید.

3. نرمال‌سازی داده‌ها

3.1. با استفاده از توابع Scale، normalize و minmax_scale مقادیر دیتاست را نرمال‌سازی کنید.

4. خلاصه‌سازی و بصری‌سازی

4.1. نمودار هیستوگرام هر یکی از ویژگی‌های دیتاست را نمایش دهید.

5. بررسی همبستگی بین متغیرها

5.1. در دیتاست، همبستگی بین متغیرها را با استفاده از نمودار pairplot بررسی کرده و این نمودار را تفسیر کنید.

5.2. در دیتاست، همبستگی بین متغیرهای Day Charge، Day Calls و Day Mins را با استفاده از تابع pearsonr بدست آورید.

5.3. با استفاده از متد corr از کتابخانه‌ی pandas مقدار همبستگی بین متغیرهای Day Charge، Day Calls و Day Mins را در دیتاست نشان دهید.

5.4. دیتافریم بدست‌آمده در مرحله‌ی قبل را با استفاده از نمودار heatmat از پکیج seaborn نشان دهید. این نمودار برای چه مواقعی مناسب است؟

6. Chi-Square

6.1. در خصوص رابطه‌ی بین evening minutes و churn relationship در دیتاست، فرض H0 و H1 را تعیین کنید. هدف بررسی وابستگی یا استقلال این دو پارامتر است.

قسمت دوم: سوالات تشریحی – بدون نیاز به برنامه نویسی

1. جدول زیر را که یک دیتاست کوچک با 10 رکورد است در نظر بگیرید:

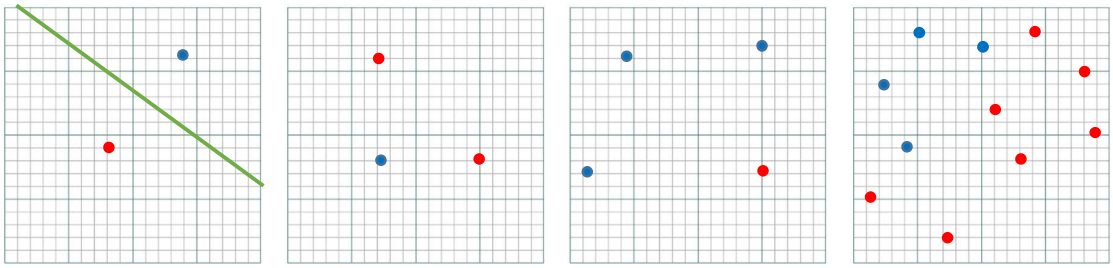
Record	Age	Marital	Income	Risk
1	22	Single	$46,156.98	Bad loss
2	33	Married	$24,188.10	Bad loss
3	28	Other	$28,787.34	Bad loss
4	51	Other	$23,886.72	Bad loss
5	25	Single	$47,281.44	Bad loss
6	39	Single	$33,994.90	Good risk
7	54	Single	$28,716.50	Good risk
8	55	Married	$49,186.75	Good risk
9	50	Married	$46,726.50	Good risk
10	66	Married	$36,120.34	Good risk

با در نظر گرفتن K=3 در روش k-nearest neighbor کلاس نمونه‌ی x=(30, Single, $30000) را بیابید. دقت شود که داده‌ها باید نرمال شوند.

2. در دیتاست داده‌شده در سوال 1 تشریحی، فرض کنید ویژگی‌ها فقط شامل دو ستون Age و Income هستند. با استفاده از بیز ساده‌ی گوسی، کلاس نمونه‌ی x=(24, $25000) را بیابید.

3. در دیتاست داده‌شده در سوال 1 تشریحی، با استفاده از گسسته‌سازی مبتنی بر آنتروپی، ویژگی عددی "Age" را به سه مقدار "a1"، "a2" و "a3" تبدیل کرده و دیتاست جدید را نشان دهید.
4. داده‌های آموزشی زیر را در فضای دوبعدی xy در نظر بگیرید.

(الف) مرز تصمیم طبقه‌بند 1-NN با فاصله‌ی اقلیدسی را رسم کنید. برای نمونه در شکل (1) مرز تصمیم را نشان داده‌ایم. (راهنمایی: عمودمنصف پاره‌خط واصل بین دو نقطه‌ی A و B، مکان هندسی نقاطی را نشان می‌دهد که فاصله‌ی اقلیدسی آن از دو نقطه‌ی A و B برابر است.)



(ب) می‌دانیم که طبقه‌بند نزدیکترین همسایگی یک طبقه‌بند Lazy به حساب می‌آید. بنابراین بایستی همه‌ی داده‌های آموزشی را به منظور استفاده در زمان تست، ذخیره کنیم. در قسمت قبل نشان داده‌ایم می‌توان برای طبقه‌بند 1-NN مرز تصمیم بدست آورد. در صورتی‌که به جای ذخیره‌ی همه‌ی داده‌های آموزشی، مرز تصمیم را ذخیره کنیم، آیا از نظر حافظه‌ی مورد نیاز برای ذخیره‌سازی همیشه بهبود خواهیم داشت؟ (یک جواب بله یا خیر مشخص کنید و در دو تا سه جمله دلیل خودتان را توضیح دهید.)

(پ) برای ساخت درخت تصمیم بایستی همه‌ی داده‌های آموزشی را در ابتدا در اختیار داشته باشیم. اگر داده‌ی آموزشی جدیدی وارد شود، باید به دقت مدیریت شود. آیا KNN نیز این مشکل را دارد. چرا؟

(ت) فرض کنید درخت تصمیم D از روی دیتاست T ایجاد شده است. بعد از ساخت درخت، تعدادی داده‌ی آموزشی دیگر (D') به ما داده می‌شود. چطور می‌توان درخت T را گسترش داد به طوری‌که درخت گسترش داده‌شده (T') از روی داده‌ی D+D' باشد. (احتمالا T' به خوبی درختی که از ابتدا با استفاده از داده‌ی D+D' ساخته شود، نخواهد بود با این‌حال در این سوال به دنبال ساخت درخت از ریشه نیستیم.) اگر نیازی به اطلاعات دیگری از درخت T دارید، عنوان کنید.

5. چرا هرس کردن درخت در الگوریتم‌های درخت تصمیم خوب است؟ پیش‌هرس و پس‌هرس کردن را مقایسه کنید.

نکات تکلیف:

1. الزامی به تایپ پاسخ‌های بخش تشریحی نیست. اما لازم است پاسخ‌ها مرتب و خوانا نوشته شوند و تصویر با کیفیتی از آن‌ها در قالب یک فایل pdf قرار داده شود. در صورت خوانا نبودن پاسخ‌ها نمره‌ای به آن تعلق نخواهد گرفت.
2. پاسخ‌ها به زبان فارسی نوشته شود.
3. راه حل سوالات محاسباتی به صورت کامل نوشته شود.
4. به تکالیف کپی‌شده بین گروه‌های مختلف، نمره‌ای تعلق نخواهد گرفت.
5. زبان برنامه‌نویسی سوالات پایتون است. ترجیحاً در محیط jupyter notebook کدنویسی شود.
6. تا پنج روز بعد از تاریخ تحویل، امکان آپلود پاسخ‌ها با کسر نمره وجود دارد. تاریخ آخرین ویرایش به عنوان تاریخ تحویل تکلیف در نظر گرفته می‌شود.
7. روش تحویل:

الف) فایل‌های مربوط به کدهای هر سوال بخش کدنویسی، در یک فایل با نام Bx.zip که x شماره سوال است زیپ شوند.

ب) برای هر کدام از سوالات کدنویسی، یک ویدئوی کوتاه با نام Bx که x شماره سوال است قرار داده و در آن می‌بایست کد سوال، اجرا و توضیح داده شود.

پ) کلیه‌ی فایل‌های زیپ به همراه فایل‌های ویدئویی و فایل pdf پاسخ‌های بخش تشریحی، در یک فایل واحد با نام HW2-Lastname.zip که Lastname نام خانوادگی شما و هم‌گروه(های) شماست، زیپ شده و **فقط در سامانه‌ی یکتا** تا ساعت 23:59 روز 19 فروردین آپلود شوند.

8. در صورت داشتن هرگونه سوال، می‌توانید با دستیاران آموزشی درس در تلگرام در ارتباط باشید. ترجیحاً سوالات درسی را در گروه تلگرامی درس مطرح کنید.

موفق باشید