«به نام خدا»

تمرین دوم درس داده‌کاوی

هدف: آشنایی با مفاهیم پایه درخت تصمیم‌گیری، دسته‌بندهای بیز و KNN

تذکر۱ : ملاک اصلی انجام تمارین بخش پیاده سازی، گزارش است و ارسال کد بدون گزارش فاقد ارزش است. لذا برای این بخش یک فایل گزارش مستقل در قالب pdf تهیه کنید و در آن برای هر سوال، تصاویر ورودی، تصاویر خروجی و توضیحات مربوط به آن را ذکر کنید. سعی کنید
توضیحات کامل و جامعی تهیه کنید. همچنین قابل توجه است که زبان مورد قبول برای بخش پیاده‌سازی، تنها **پایتون** می‌باشد.

تذکر۲ : مطابق قوانین دانشگاه هر نوع کپی برداری و اشتراک کار دانشجویان غیر مجاز بوده و شدیدا برخورد خواهد شد. استفاده از
کدها و توضیحات اینترنت به منظور یادگیری **الزاما با ذکر منبع** بلامانع است.

راهنمایی: در صورت نیاز میتوانید سواالت خود را در خصوص پروژه از تدریسیار درس، از طریق ایمیل زیر بپرسید .

E-mail: zahra.dehghanian97@gmail.com

ارسال: فایل های کد و گزارش خود را در قالب یک فایل فشرده با فرمت StudentID_DM02.zip تا تاریخ ۱۴/۰۸/۱۴۰۰ ارسال نمایید.

* شایان ذکراست در مجموع برای تمارین ۷ روز تاخیر مجاز در نظر گرفته شده است و افزون بر آن هر روز تاخیر باعث کسر **۲۰٪ نمره** کل تمرین خواهد شد.

• بخش نوشتاری

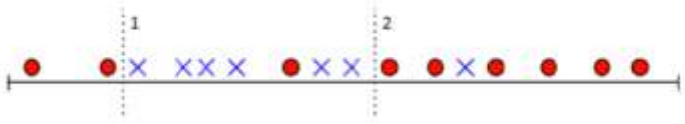
۱) فرض کنید یک نرم افزار فیلتر برای تشخیص اتوماتیک spam وجود دارد. این سیستم براساس حضور و یا عدم حضور بعضی کلمات در متن ایمیل، نوع آن را تشخیص می‌دهد. در زیر داده‌های آموزشی برای این نرم‌افزار را مشاهده می‌کنید:

'study'	'free'	'money'	Category
1	0	0	Regular
0	0	1	Regular
1	0	0	Regular
1	1	0	Regular
0	1	0	Spam
0	1	0	Spam
1	1	0	Spam
0	1	0	Spam
0	1	1	Spam
0	1	1	Spam
0	1	1	Spam
0	0	1	Spam

الف) اگر این نرم افزار مقدار$p(spam) = 0.1$ را در نظر بگیرد، توضیح دهید که آیا این کار منطقی است؟ چرا؟

ب) بر اساس قاعده بیز(naïve bayes) و $p(spam) = 0.1$ مشخص کنید که با جمله money for psychology study چگونه برخورد می‌شود و در چه دسته‌ای قرار می‌گیرد؟

۲) در شکل زیر داده‌های دو کلاس بر روی یکی از متغیرهای پیوسته خود رسم شده‌اند .در صورتی که بخواهیم برای درخت تصمیم این ویژگی پیوسته را به یک ویژگی گسسته باینری تبدیل نماییم، کدام یک از نقاط ۱و یا ۲ برای این کار مناسب‌تر هستند؟



۳) شما سعی دارید دسته‌بند مناسبی را برای تعیین اینکه کدام رستوران برای شام با دوستانتان مناسب‌تر است، طراحی کنید. برای این کار شما اطلاعات مربوط به ۱۱ رستوران مختلف و به طور خاص اطلاعاتی در مورد نوع رستوران‌ها، محل آن‌ها، محدوده قیمت و اینکه آیا آن‌ها می‌توانند محدودیت‌های غذایی را پوشش دهند و آیا از آن‌ها لذت می‌برید یا خیر را جمع‌آوری کرده‌اید.

داده‌ها در جدول زیر گزارش شده‌است:

Restaurant	Type	Price	Neighborhood	Restriction	OK
R_1	Fast Food	$	Oakland	Vegetarian	0
R_2	Ethnic	$$	Squirrel Hill	Gluten Free	0
R_3	Casual Dining	$$	Squirrel Hill	None	0
R_4	Casual Dining	$$$	Shadyside	Vegetarian	0
R_5	Casual Dining	$	Oakland	Vegetarian	1
R_6	Fast Food	$$	Squirrel Hill	None	1
R_7	Ethnic	$	Squirrel Hill	None	1
R_8	Casual Dining	$	Shadyside	Gluten Free	0
R_9	Fast Food	$$$	Oakland	None	0
R_{10}	Ethnic	$$	Shadyside	Vegetarian	1
R_{11}	Casual Dining	$$	Shadyside	Gluten Free	1

الف) با استفاده از این داده‌ها یک درخت تصمیم‌گیری برای تصمیم‌گیری درباره اینکه آیا می‌توانید از یک رستوران خاص لذت ببرید یا نه طراحی کنید. از معیار ناخالصی انتروپی به این منظور استفاده کنید و درخت تصمیم را تا عمق ۴ ادامه دهید. در هر سطح نحوه تصمیم‌گیری درباره اینکه کدام ویژگی را گسترش دهید، نشان دهید.

ب) خطای درخت تصمیم طراحی شده در مرحله قبل را محاسبه کنید. (تعداد نقاط در مجموعه آموزشی که به اشتباه دسته‌بندی شده است.)

۴) فرض کنید داده‌های پنج رستوران دیگرعلاوه بر جدول سوال ۳ به شرح زیر به شما داده می‌شود:

Restaurant	Type	Price	Neighborhood	Restriction
R_{12}	Fast Food	$	Squirrel Hill	None
R_{13}	Ethnic	$$	Shadyside	None
R_{14}	Ethnic	$	Oakland	Gluten Free
R_{15}	Casual Dining	$	Shadyside	Vegetarian
R_{16}	Ethnic	$	Squirrel Hill	Gluten Free

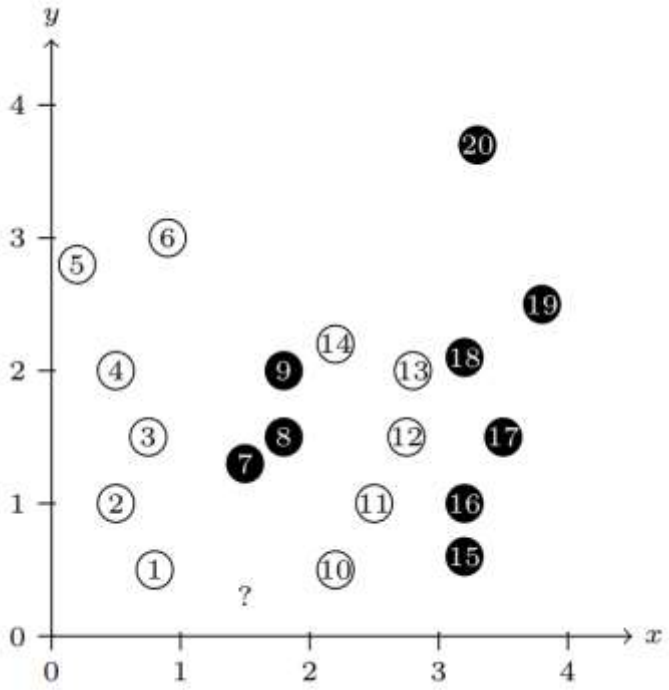
الف) به کدام یک از این رستوران‌ها می‌روید؟

ب) از روی کنجکاوی و برای بررسی صحت درخت تصمیم، تصمیم می‌گیرید همه آن‌ها را امتحان کنید. نتایج عبارتند از:

Restaurant	OK
R_{12}	0
R_{13}	1
R_{14}	0
R_{15}	1
R_{16}	0

درمورد عملکرد درخت تصمیم خود توضیح دهید.

۵) در شکل زیر مجموعه‌ای از نقاط آموزشی را نشان می‌دهیم که به صورت سیاه یا سفید طبقه‌بندی می‌شوند. قصد داریم از الگوریتم KNN برای دسته‌بندی نقاط جدید استفاده کنیم. نقطه مشخص‌شده توسط ? با استفاده از فاصله اقلیدسی به عنوان معیار فاصله به ازای k=1,2,3 در چه دسته‌ای قرار می‌گیرد؟



تمرینات سری دوم درس داده کاوی دکتر مزلقانی

۶) آیا در شکل سوال ۵ نقطه‌ای در مجموعه آموزشی وجود دارد که با استفاده از k=1 به اشتباه دسته‌بندی شود؟ اگر اینطور است، آن‌ها را شناسایی کنید.

۷) در شکل سوال ۵، یک معیار فاصله ساده که تمام نقاط در مجموعه آموزشی را به ازای k=1 درست طبقه‌بندی می‌کند بیایید.

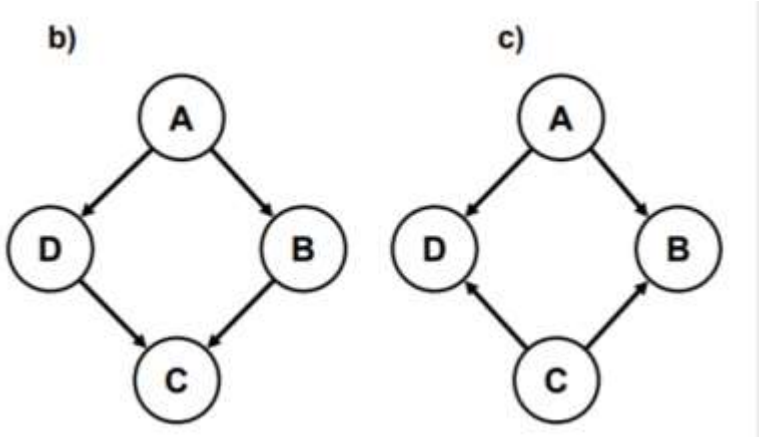
ب) زمانی که به ازای k=5 از معیار فاصله شما استفاده کنیم، چه اتفاقی می‌افتد ؟

ج) نقطه مشخص‌شده توسط ? با استفاده از معیار فاصله شما به ازای k=1,2,3 در چه دسته‌ای قرار می‌گیرد؟

۸) دو روش هرس کردن درخت (post-pruning و pre-pruning) را با هم مقایسه کرده و مزایا و معایب هرکدام را ذکر کنید.

۹) مشخص کنید هرکدام از گراف‌های زیر نشان دهنده کدام عبارت است؟

- $A \perp C|B,D$
- $B \perp D|A,C$



تمرینات سری دوم درس داده کاوی دکتر مزلقانی

۱۰) به عنوان بخشی از مطالعه جامع شادی مردم، ما داده‌های مهمی از فارغ التحصیلان را جمع‌آوری کرده‌ایم . در یک بررسی کاملاً اختیاری که همه دانشجویان ملزم به تکمیل آن بودند، از سوالات زیر استفاده کردیم:

*آیا شما مرتباً به مهمانی می‌روید؟[Party: Yes/No]

*آیا شما باهوش هستید؟ [Smart : Yes/No]

* آیا شما خلاق هستید؟ [Creative: Yes/No]

* آیا در تمام تکالیف خود خوب عمل کرده‌اید؟ [HW: Yes/No]

* آیا از یک مک استفاده می‌کنید؟ [Mac: Yes/No]

* آیا پروژه شما موفق شد؟ [Project: Yes/No]

* آیا شما در مهم‌ترین کلاس خود موفق بودید؟ [Success: Yes/No]

* آیا در حال حاضر خوشحال هستید؟ [Happy: Yes/No]

بعد از مشورت با یک روان‌شناس، مجموعه کاملی از روابط مشروط را به دست آوردیم:

HW فقط به Party و Smart بستگی دارد.

Mac تنها به Smart و Creative وابسته است.

Project تنها به Smart و Creative بستگی دارد.

Success فقط به HW و Project بستگی دارد.

Happy تنها به Party، Mac و Success بستگی دارد.

الف) شبکه بیزین این تحقیق را بکشید.

ب) توزیع مشترک را به عنوان محصول احتمالات شرطی بنویسید.

ج) تعداد پارامترهای مستقل مورد نیاز برای هر جدول احتمال شرطی چیست؟

د) چند پارامترهای مستقل داریم؟

- **بخش پیاده سازی:**

در این بخش شما باید با استفاده از سه الگوریتم درخت تصمیم‌گیری، knn و یکی از دسته بندهای بیز، داده‌ها را پیش‌بینی کنید.

- **قسمت اول : آماده‌سازی داده‌ها**

پیش از آنکه الگوریتم را روی داده‌ها پیاده‌سازی کنید، نیاز است که داده‌های عددی را به داده-های فهرستی(categorical data) تبدیل کنید. (برای مثال، داده هایی که درباره سن و سال (age) افراد است را به این صورت دسته بندی کنید: سن ۰ تا ۱۰ را با لغت "teenager" ، سن ۱۱ تا ۲۰ را با لغت "adult" نشان دهید و برای بقیه اعداد به همین ترتیب ادامه دهید.)
مجموعه داده‌های ارائه شده، شامل داده‌های عددی و فهرستی است.
در نهایت می‌توانید با استفاده از تکنیک "onehotencoding" در پکیج sklearn ، داده‌های فهرستی را encode کنید.

- **قسمت دوم: کلاس‌بندی داده‌ها**

پس از آماده‌سازی داده‌ها، میتوانیم آنها را کلاس‌بندی و پیش‌بینی کنیم.
در ابتدا داده‌ها را به دوقسمت مجزا، با نسبت ۸۰ / ۲۰ تقسیم کنید.(splitting)
در واقع مجموعه آموزش (train set) ۸۰ درصد مجموعه داده را شامل شود و ۲۰ درصد بقیه برای مجموعه تست (test set) استفاده شود.
مجموعه‌ی آموزش برای آموزش الگوریتم و مجموعه‌ی تست برای مشاهده‌ی الگوریتم آموزش دیده بر روی داده‌های مجموعه ی آموزش استفاده می‌شود.
کلاس‌بندهای درخت تصمیم‌گیری و دسته‌بندهای بیز و knn ، از پارامترهای مختلفی برای کلاس‌بندی کردن استفاده می‌کنند.

تمرینات سری دوم درس داده کاوی دکتر مزلقانی

۱) Dataset1.csv را با الگوریتم درخت تصمیم‌گیری آموزش دهید و سپس ستون "income" را برای مجموعه داده Dataset1_Unknown.csv پیش‌بینی کنید. برای آموزش الگوریتم درخت تصمیم‌گیری، یک بار پارامتر "criterion" را "gini" قرار دهید و سپس آن را "entropy" قرار دهید و نمودار درخت تصمیم‌گیری را رسم کنید.(visualize کنید). و پس از کلاسبندی مجموعه داده‌های ناشناخته اول)Dataset1_Unknown.csv(، یک بردار یا یک ستون از نتایج را به همراه دقت الگوریتم، با فایل گزارش ارسال کنید.

۲) Dataset2.csv را با الگوریتم knn آموزش دهید و سپس ستون "poisonous" را برای مجموعه داده Dataset2_Unknown.csv پیش‌بینی کنید. و پس از کلاسبندی مجموعه دادههای ناشناخته دوم Dataset2_Unknown.csv(، یک بردار یا ستون از نتایج را به همراهaccuracy الگوریتم، به ازای سه پارامتر k متفاوت، با فایل گزارش ارسال کنید.

۳) Dataset3.csv را با الگوریتم نایوبیز آموزش دهید و سپس ستون "disease" را برای مجموعه داده Dataset3_Unknown.csv پیش‌بینی کنید. و پس از کلاسبندی مجموعه دادههای ناشناخته سوم)Dataset3_Unknown.csv(، یک بردار یا ستون از نتایج را به همراهaccuracy الگوریتم، با فایل گزارش ارسال کنید.