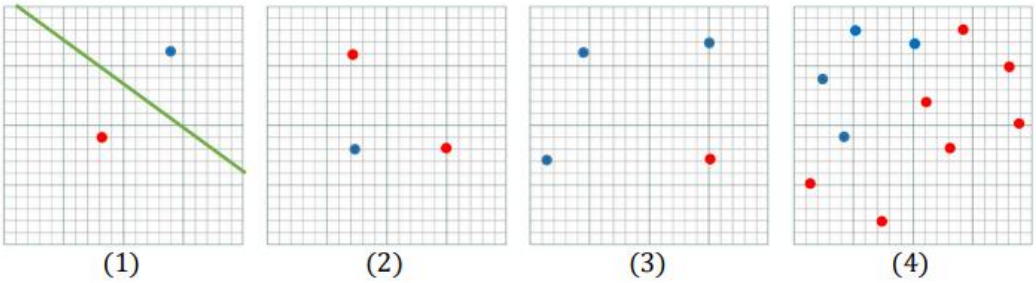
تمرین دوم درس مبانی داده کاوی

(زمستان ۴۰۱)

مهلت تحویل تمرین: ۱۵ فروردین ماه

سوالات تئوری

سوال ۱- داده‌های آموزشی زیر را در فضای دو بعدی xy در نظر بگیرید.



- مرز تصمیم طبقه‌بند NN-1 با فاصله اقلیدسی را رسم کنید. برای نمونه در شکل (۱) مرز تصمیم را نشان داده ایم.
- 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 داده ایم می‌توان برای طبقه‌بند NN-1 مرز تصمیم بدست آورد. در صورتیکه به جای ذخیره‌ی همه‌ی داده‌های آموزشی، مرز تصمیم را ذخیره کنیم، آیا از نظر حافظه‌ی مورد نیاز برای ذخیره سازی همیشه بهبود خواهیم داشت؟ (یک جواب بله یا خیر مشخص کنید و در دو تا سه جمله دلیل خودتان را توضیح دهید.)
- 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؟

سوال ۲- چرا هرس کردن درخت در الگوریتم‌های درخت تصمیم خوب است؟ پیش هرس و پس هرس کردن را مقایسه کنید.

- فرض کنید درخت تصمیم T از روی دیتاست D 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 داده شده (T′) از روی داده‌های D+D′ باشد. (احتمالا T′ به خوبی درختی که از ابتدا با استفاده از داده‌های D+D′ ساخته شود، نیست با این حال در این سوال به دنبال ساخت درخت از ریشه نیستیم.) اگر نیازی به اطلاعات دیگری از درخت T دارید، عنوان کنید.

سوال ۳- با استفاده از رکوردهای جدول زیر و قانون بیز محاسبه کنید درصورتی که کسی دارای تب، عدم سرفه و دارای سردرد باشد، آیا آن فرد سرماخوردگی دارد یا خیر.

شماره رکورد	سردرد	سرفه	تب	سرماخوردگی؟
۱	دارد	دارد	دارد	آری
۲	دارد	دارد	دارد	خیر
۳	دارد	دارد	دارد	آری
۴	دارد	دارد	ندارد	خیر
۵	ندارد	دارد	ندارد	آری
۶	ندارد	ندارد	ندارد	خیر
۷	ندارد	ندارد	ندارد	آری
۸	دارد	ندارد	ندارد	خیر
۹	ندارد	ندارد	دارد	خیر
۱۰	ندارد	دارد	دارد	آری

سوال ۴- (اختیاری) جدول زیر را که یک دیتاست کوچک با ۱۰ 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 نمونه X=(30, Single, $30000) را بیابید. دقت شود که داده‌ها باید نرمال شوند.
- در دیتاست داده شده، فرض کنید ویژگی‌ها فقط شامل دو ستون Age و Incom هستند. با استفاده از قانون بیز، کلاس نمونه X=(۲۴, $۲۵۰۰۰) را بیابید

سوالات عملی

سوال ۵ - با استفاده از مجموعه داده Kaggle titanic به سوالات زیر پاسخ دهید:

۱- نخست دیتاست را با استفاده از کتابخانه pandas خوانده و تبدیل به دیتافریم نمایید.
۲- ابتدا ستون Cabin را حذف کنید و سپس سطرهای شامل Null را در این دیتاست حذف کنید.
۳- ستون‌های Passengerld, Name, Ticket, SibSp, Parch از مجموعه داده حذف کنید.
۴- خصوصیاتی با نوع categorical را به روش oneHot Encoding به عدد تبدیل کنید.
۵- همه ستون‌ها بجز Survived را به x و Survived را به y تعریف کنید.
۶- داده‌ها را به دو بخش train و test تفکیک کنید. (۷۰ درصد داده train و ۳۰ درصد داده test)
۷- مدل درخت تصمیم را با random_state = 0 روی داده اعمال کنید و دقت بدست آمده برای train و test را به تفکیک مشخص کنید.
۸- درخت به دست آمده را با graphViz و pydotplus نمایش دهید.
۹- آیا overfitting رخ داده است؟ چرا؟ اگر جواب شما مثبت است راه‌حل این مشکل چیست؟
۱۰- Confusion Matrix را برای test رسم کنید و آنرا تفسیر کنید.
۱۱- با استفاده از Grid Search بهترین مقدار برای پارامترهای max_depth و min_samples_leaf را بدست آورید. آیا دقت مدل بهتر خواهد شد؟ توضیح دهید.
۱۲- بهترین درخت بدست آمده توسط Grid Search را با استفاده از GraphViz نمایش دهید.
۱۳- مدل Random Forest را بر روی داده‌های train اجرا کنید.
۱۴- دقت بدست آمده بر روی train و test نسبت به درخت تصمیم چقدر تغییر کرده است؟
۱۵- با استفاده از Random Search بهترین حالت برای پارامترهای max_depth و min_samples_split و bootstrap و min_samples_leaf و criterion و n_estimators را بدست آورید.
۱۶- آیا معیار gini موثر است یا entropy ؟

سوال ۶ - روی مجموعه داده iris، برای جداسازی داده آموزشی از داده برای تست از روش Leave One Out که هربار یکی از رکوردهای داده را برای تست و مابقی را برای آموزش استفاده می‌کند، بهره بگیرید (برای اینکار می‌توانید از تابع LeaveOneOut در کتابخانه sklearn استفاده کنید)، الگوریتم SVM خطی را اجرا کنید. پس از آموزش مدل روی مجموعه آموزشی به کمک مدل آموزش دیده روی مجموعه داده تست پیش‌بینی انجام دهید. با تحلیل ماتریس Confusion نتیجه را بررسی کنید.

۱- این‌بار می‌خواهیم kernel SVM را روی مجموعه داده iris اجرا کنیم. برای اینکار از متد polynomial استفاده کنید. مقادیر پارامتر چندجمله‌ای را از ۱ تا ۱۰ تغییر دهید و نمودار نرخ خطای کلاس‌بندی را برحسب درجه چندجمله‌ای را رسم و نتایج نمودار را تحلیل کنید.
۲- باتوجه به نتایج قسمت قبلی kernel SVM با متد polynomial و درجه بهینه را پیاده سازی کنید و نتایج را بر اساس ماتریس Confusion تحلیل کرده و با قسمت قبل مقایسه کنید.

سوال ۷ (اختیاری) - با استفاده از مجموعه داده Kaggle titanic به سوالات زیر پاسخ دهید:

۱- موارد یک تا پنج سوال اول عملی را انجام دهید.
۲- داده‌ها را به ۷۰ درصد آموزشی و ۳۰ درصد تست تقسیم کنید. (train_test_split)
۳- با استفاده از MLPClassifer و پارامترهای پیش‌فرض آن و بدون استانداردسازی داده، مدل‌سازی انجام دهید و دقت مدل را روی داده تست گزارش دهید.
۴- در صورت استانداردسازی داده مدل MLPClassifer چقدر باعث افزایش دقت تست می‌شود؟
۵- با استفاده از GridSearchCV و انجام HyperParameters Tuning بهترین مقدار پارامترهای زیر را از میان مقادیر زیر بدست آورید و گزارش دهید چقدر به بهبود دقت مدل کمک شده است.
   a. الگوریتم بهینه‌سازی: SGD و Adam
   b. نرخ یادگیری: $1e^{-1}$ و $1e^{-2}$ و $1e^{-3}$ و $1e^{-4}$ و $1e^{-5}$
   c. تعداد لایه و نورون‌ها: بین یک تا سه لایه پنهان و هر لایه بین ۱۰۰ تا ۱۰۰۰ نورون
   d. تابع فعال‌سازی: خطی، tanh و relu
۶- برای مدل بدست آمده بخش قبل ماتریس Confusion رسم کنید و مقادیر Precision Recall F1-Score را برای هر کلاس جداگانه بدست آورید.

سوال ۸ (اختیاری) - در این سوال روی مجموعه داده iris کار کنید. برای جداسازی داده آموزشی از تست از تابع train_test_split استفاده کرده و ۲۰ درصد از داده‌ها را برای تست و مابقی را برای آموزش استفاده کنید. دقت کنید که قبل از اجرای الگوریتم داده‌ها را نرمال کنید.

۱- الگوریتم KNN را با استفاده از KDTree برای مقادیر k از ۱ تا ۳۰ اجرا کنید و نمودار نرخ خطای کلاس‌بندی را براساس مقدار k رسم کنید و آن را تحلیل کنید. برای k بهینه ماتریس Confusion را محاسبه و تحلیل کنید.
۲- قسمت قبل را اینبار به کمک BallTree تکرار کنید. پس از تحلیل نتایج این دو روش را مقایسه کنید.

نحوه تحویل: سوالات تئوری را به صورت تایپ شده و در قالب یک فایل PDF تحویل دهید. به علاوه هر یک از سوالات عملی را در قالب یک فایل ipynb به همراه نتایج قرار داده و فایل را به صورت Qn نام گذاری نمایید که n شماره سوال مربوطه می باشد. در انتها فایل های پایتون را به همراه فایل PDF تماما در قالب یک فایل zip نامگذاری شده به صورت NAME_STUDENTID در سامانه درس بارگذاری کنید. برای سوالات عملی توضیحات خود را به صورت Markdown در فایل پایتون بنویسید.

"It is often in the darkest skies that we see the brightest stars." – Richard Evans