«به نام خدا»

تمرین سوم درس داده‌کاوی

هدف: آشنایی با مفاهیم SVM، روش‌های ensemble و قواعد انجمنی

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_DM03.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 در دیتاست این مسئله تنها ۵ دیتا داریم و مجموعه زیر مجموعه آیتم ست های پرتکرار[۱] سه تایی هستند

{۱, ۲, ۳}, {۱, ۲, ۴}, {۱, ۲, ۵}, {۱, ۳, ۴}, {۱, ۳, ۵}, {۲, ۳, ۴}, {۲, ۳, ۵}, {۳, ۴, ۵}.

الف) لیست آیتم ست های پرتکرار ۴ تایی که با روش $Fk-1 \times F1$ بدست می آید را بنویسید.

ب) لیست ایتم ست های ۴تایی تولید شده توسط الگوریتم a-priori در مرحله نولید کاندیداها را بنویسید.

ج) لیست آیتم ست های پرتکرار ۴تایی تولید شده توسط الگوریتم a-priopri را بنویسید.

۲) معیارهای زیادی جهت استفاده برای آزمودن مدل های classification وجود دارد. از جمله معیارهایی که در کلاس با آنها آشنا شدیم Precision، Recall و F-Measure بوده است. معیار های زیر را تعریف نمایید.

الف) معیار های Sensitivity، Specificity و G-Mean.

ب) معیار های True Positive Rate و False Positive Rate.

ج) معیار های Brier Score و Log Loss.

۳) در یک شهر، ۷۰٪ مردم ماسک میزنند و فقط ۲۰٪ مردم از وسایل ضدعفونی کننده استفاده میکنند. در این شهر طبق آمار انجام شده، کسانی که از ماسک و وسایل ضدعفونی کننده استفاده میکنند ۲۵٪ گلودرد و تب دارند، کسانی که ماسک میزنند و وسایل ضدعفونی کننده استفاده نمیکنند ۴۰٪ گلودرد و تب دارند، کسانی که ماسک نمیزنند و از وسایل ضدعفونی کننده دیگر استفاده میکنند حدود ۵۰٪ گلودرد و تب دارند و کسانی که از ماسک و وسایل ضدعفونی کننده استفاده نمیکنند حدود ۷۰٪ گلودرد و تب دارند. حال با مراجعه به پزشک و انجام آزمایش و تست کرونا مشخص شده است که ۴۵ درصد افراد مبتلا به گلودرد و تب، مبتلای به کرونا نیز هستند. از میان افرادی که گلودرد ندارند نیز حدود ۵ درصد مبتلا به کرونا هستند که علائم آن هنوز بروز نکرده بود.

الف) مدل گرافی جهتدار آماری سلامت این شهر را رسم نمایید.

[۱] frequent itemset

ب) با استفاده از فرض Markov، احتمال گلودرد افراد به شرط داشتن کرونا و استفاده کردن از ماسک و استفاده نکردن از وسایل ضدعفونی کننده را محاسبه نمایید.

۴) دو دسته اصلی روش های یادگیری گروهی ensemble learning را نام برده، نحوه عملکرد هر کدام و تفاوت آن ها به طور کامل توضیح دهید.

۵) در ارتباط با ماشین های بردار پشتیبان به سوالات زیر پاسخ دهید.

الف) توجه به SVM دو کلاسه شرح داده شده در کلاس، یک سناریو برای SVM چند کلاسه (مثلا با m کلاس) ارائه دهید.

ب) منظور از مدل با حاشیه سخت چیست؟

ج) فرض کنید یک ماشین بردار پشتیبان با مرزبندی خطی آموزش داده اید و متوجه می شوید دچار کم برازش شده است. برای حل این مشکل پارامترهای مدل خود را چگونه تغییر می دهید؟

د) function kernel چیست؟ ۲ نمونه رایج ترین آن ها را نام برده و transformation مربوط به هر یک را به طور مختصر توضیح دهید

۶) تراکنش های زیر را طبق مراحل خواسته شده ارزیابی کنید.

الف) الگوریتم apriori را برروی تراکنش های زیر اجرا کنید. فرض کنید آستانه پشتیبانی(support) برابر ۳۳ ٪و آستانه اطمینان(confidence) ٪۶۰ می باشد. تمامی مراحل تولید مجموعه آیتم های کاندید را نشان دهید و درنهایت مجموعه آیتم های پرتکرار را بدست آوردید. همچنین تمامی قواعد انجمنی قابل تولید از مجموعه آیتم ها را نوشته، آنهایی که مطمئن هستند را مشخص کرده و براساس میزان اطمینان مرتب کنید

ب) با استفاده از داده های بخش قبل و با همان آستانه پشتیبانی، یک درخت الگو پرتکرار(FP-tree) بسازید، مراحل گسترش درخت با هر تراکنش را نشان دهید

ج) با استفاده از الگوریتم FP-Growth ،مجموعه آیتم های پرتکرار را بدست آورید.

	Items
T1	پیتزا، نوشابه ، برگر
T2	پیتزا، نوشابه
T3	برگر، سیب‌زمینی
T4	پیتزا، سیب‌زمینی، دلستر
T5	برگر، دلستر
T6	پیتزا، سیب‌زمینی، دلستر

۷) مجموعه داده‌های زیر، مربوط به بیماران قلبی می‌باشد. بر اساس داده‌های زیر، مدل adaboost را تا ۲ مرحله تشکیل دهید(تمامی فرضیات مورد نیاز در نظر گرفته شده و جزییات به طور کامل نوشته شود) و در نهایت confusion matrix مدل را برای همین داده ها بدست آورید.

بیماری قلبی	وزن	گرفتگی رگ	درد قفسه سینه
دارد	205	دارد	دارد
دارد	180	دارد	ندارد
دارد	210	ندارد	دارد
دارد	167	دارد	دارد
ندارد	156	دارد	ندارد
ندارد	125	دارد	ندارد
ندارد	168	ندارد	دارد
ندارد	172	دارد	دارد

بخش پیاده سازی:

۱) ابتدا مجموعه دادهی ۲ بعدی txt.data را خوانده و آن را به نسبت ۹۰ به ۱۰ جدا کنید. دقت کنید که ۲ دو عدد اول هر سطر مختصات داده و عدد سوم برچسب آن است. حال میخواهیم این مجموعه داده را با svm linear طبقه بندی کنیم.

الف) با توجه به مطالب گفته شده در کلاس، خط مرزی را بدست آورده و نمودار خط مرزی و داده های آموزش را گزارش کنید. هم چنین میزان دقت مدل را محاسبه کرده وگزارش کنید.

ب) این کار را برای مجموعه دادهی txt.data2 هم انجام دهید و نمودار و دقت مدل را گزارش کنید.

ج) همانطور که مشاهده میکنید،svm linear برای این مجموعه داده مناسب نیست و باید از svm nonlinear استفاده کنیم. برای این کار میتوانید از کتابخانه ی sklearn استفاده کنید و نیازی به پیاده سازی خود الگوریتم نیست. مدل را آموزش داده و خروجی را برای دادههای تست محاسبه کرده و دقت را گزارش کنید. نقاط آموزش و مرز به دست آمده را رسم کنید.

۲) در این تمرین میخواهیم به پیاده سازی مدل forest random بپردازیم. مجموعه دادهی txt.data3 را خوانده و آن را به نسبت ۸۰ به۲۰ جدا کنید. سپس مدل forest random را پیاده سازی کرده و آن را بر اساس داده های آموزش، آموزش دهید(کلاس هدف، صفت Species است). دقت مدل، confusionmatrix مدل برای داده های تست را بدست آورده و آن را گزارش کنید. همچنین نشان دهید کم یا زیاد شدن تعداد random treeها، چه تأثیری در دقت مدل دارد و آن را گزارش کنید.

۳) در این تمرین میخواهیم به پیاده سازی مدل adaboost بپردازیم. مجموعه دادهی txt.data4 که اطلاعات مربوط به مسافران تایتانیک می باشد را خوانده و آن را به نسبت ۸۰ به ۲۰ جدا کنید (پیش‌پردازش فراموش نشود). سپس. سپس مدل adaboost را پیاده سازی کرده و آن را بر اساس داده های آموزش، آموزش دهید.دقت مدل، matrix confusion مدل برای داده های تست را بدست آورده و آن را گزارش کنید.

** دقت شود که در تمامی سوالات باید الگوریتم های خواسته شده به طور کامل توسط خود دانشجو پیاده سازی شود و خلاصه ای از عملکرد کلی و همچنین جزییات مدل ها نیز در کنار موارد خواسته شده، گزارش شود.