به نام خدا

دانشگاه صنعتی امیرکبیر
(پلی تکنیک تهران)
دانشکده مهندسی کامپیوتر و فناوری اطلاعات

تمرین سری سوم داده کاوی

توضیحات:

- پاسخ به تمرین‌ها باید به صورت انفرادی صورت گیرد و در صورت مشاهده هرگونه تقلب نمره صفر برای کل تمرین منظور خواهد شد.
- تمیزی و خوانایی گزارش تمرین از اهمیت بالایی برخوردار است.
- گزارش تمرین خود را در قالب یک فایل PDF با نام **«HW3_StudentNumber.pdf»** به همراه کد های بخش پیاده سازی (فایل های ipynb. یا py.) در فایلی به نام **«HW3_StudentNumber.zip»** قرار داده و در سایت درس در مهلت معین بارگزاری نمایید.
- توجه داشته باشید که به سوالات پیاده سازی بدون گزارش نمره ای تعلق نمی گیرد.
- در صورت داشتن اشکال می‌توانید از طریق ایمیل **datamining.fall2020@gmail.com** با تدریس‌یاران درس در ارتباط باشید.
- همچنین لازم بذکر است که اگر مواردی در کلاس تدریس نشده انتظار می رود که خود دانشجویان جستجو کنند و انجام دهند.

نیم‌سال اول ۰۰-۹۹

سوال ۱- مزایا و معایب روش های دسته بندی مشتاق[1] (مثل درخت تصمیم، بیزین، شبکه عصبی) و روش ها دسته بندی تنبل[2](مثل KNN) را مقایسه کنید.

سوال ۲-

الف) تراکنش های زیر را در نظر بگیرید. با فرض min_sup = 60% , min_conf = 80% تمام frequent itemset ها را با استفاده از الگوریتم های Apriori و FP-growth پیدا کنید.

TID	Items_bought
T1	{M, O, N, K, E, Y}
T2	{D, O, N, K, E, Y}
T3	{M, A, K, E}
T4	{M, U, C, K, Y}
T5	{C, O, O, K, I, E}

ب) الگوریتم های Apriori و FP-growth را از لحاظ بهینگی عملکرد مقایسه کنید.

سوال ۳- در بسیاری از موارد در دنیای واقعی، برای انجام عمل دسته‌بندی، بخش زیادی از داده‌های موجود در پایگاه‌های داده برچسب مشخصی ندارند. از جمله روش‌های پیشنهادی برای حل این مشکل، روش‌های مبتنی بر یادگیری نیمه‌نظارتی[3]، یادگیری فعالانه[4] و یادگیری انتقالی[5] می‌باشند. درباره چگونگی این روش‌ها، موارد استفاده هر کدام و همچنین چالش‌های آن‌ها به اختصار توضیح دهید.

---

[1] eager
[2] lazy
[3] Semi-supervised learning
[4] Active learning
[5] Transfer learning

سوال ۴- دو دسته اصلی روش های یادگیری گروهی [6] را نام برده و تفاوت آن ها را توضیح دهید. مشخص نمایید هر یک از روش های پیشنهادی در بهبود چه مشکلی حین آموزش مدل یادگیری موثر هستند؟

سوال ۵- در ارتباط با ماشین های بردار پشتیبان[7] به سوالات زیر پاسخ دهید.

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

ب) منظور از مدل با حاشیه سخت[8] 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؟

سوال ۶- جدول زیر نوع کتاب‌های خریده شده یک فروشگاه را نشان می‌دهد:

نوع کتاب خریداری شده	شماره خرید
مذهبی، رمان، شعر	۱
مذهبی، رمان، تاریخی	۲
روانشناسی، رمان، تاریخی، شعر	۳
روانشناسی، تاریخی، شعر	۴
روانشناسی، مذهبی، رمان	۵
روانشناسی، مذهبی، رمان، تاریخی	۶

[6] Ensemble
[7] Support Vector Machines
[8] Hard Margin

روانشناسی	۷
روانشناسی، مذهبی، رمان	۸
روانشناسی، مذهبی، تاریخی	۹
روانشناسی، مذهبی	۱۰

الف) با استفاده از روش FP-Growth تمامی frequent itemset هایی که به کتاب نوع رمان ختم می‌شود را بیابید. (support = ۲۰٪).

ب) با درنظر گرفتن confidence = ۵۰٪ قواعد معتبر قابل استخراج از frequent itemset های به دست آمده از قسمت الف را بیابید.

پیاده سازی:

سوال ۸:

در این بخش هدف استفاده از جنگل تصادفی برای کلاس بندی می‌باشد. مسئله تایتانیک (سوال ۶ در تمرین ۲) را این بار با به جای درخت تصمیم با استفاده از جنگل تصادفی پیاده‌سازی کرده و دقت پیش‌بینی را اندازه‌گیری کنید (۲ حالت برای عمق درخت‌ها و ۲ حالت برای تابع تقسیم gini یا entropy در مجموع ۴ حالت مختلف). سپس مقایسه‌های زیر را انجام دهید:

۱- بالاترین دقت به دست آمده توسط مدل جنگل تصادفی را با مدل درخت تصمیم در تمرین قبل مقایسه کنید.

۲- سرعت یادگیری و تست جنگل تصادفی و درخت تصمیم را با هم مقایسه کنید.

توجه ۱- همانطور که در تمرین قبل اطلاع رسانی شد، برای محاسبه دقت پیش‌بینی روی مجموعه داده تایتانیک یکی از دو روش زیر را می‌توانید انتخاب کنید:

۱. تقسیم‌بندی داده های فایل train.csv به دو قسمت train set و test set (با نسبت ۸۰ به ۲۰) و محاسبه دقت مدل با استفاده از test set

۲. ثبت نام در سایت kaggle.com و بارگذاری نتایج پیش‌بینی شده برای فایل test.csv و دریافت امتیاز (که همان دقت پیش‌بینی شما می‌باشد) از سایت kaggle. نجوه بارگذاری و فرمت فایل ارسالی برای سایت در این لینک توضیح داده‌شده است.

روش انتخابی شما برای محاسبه دقت باید در تمرین دوم و سوم یکسان باشد تا مقایسه بین درخت تصمیم و جنگل تصادفی به درستی صورت بگیرد.

توجه ۲- نیازی به مصورسازی جنگل تصادفی نیست.

سوال ۹:

در این قسمت هدف آشنایی بیشتر و کلاس بندی با SVM (Support Vector Machine) می‌باشد. مجموعه داده مورد استفاده در اینجا همان مجموعه داده تایتانیک و هدف پیش‌بینی زنده ماندن یا نماندن مسافر خواهد بود. برای پیاده‌سازی این قسمت می‌توانید از کتاب‌خانه های موجود استفاده کنید. مراحل زیر را انجام داده و نتایج خواسته شده را در گزارش خود بیاورید:

۱- پیش‌پردازش مناسب روی مجموعه داده انجام دهید. دقت کنید که نحوه ارزیابی مدل به یکی از دو روش ذکر شده در قسمت قبل خواهد بود (تقسیم داده های train.csv به دو قسمت test set و train set و یا استفاده از سایت kaggle)، لذا در صورت استفاده از روش اول داده‌های train.csv را در این مرحله به‌هم زده و تقسیم کنید.

۲- مدل خود را با SVM با هسته خطی (Linear kernel function) آموزش داده و دقت آن را محاسبه کنید.

۳- دقت به دست آمده در بخش قبل را با دقت درخت تصمیم (در تمرین قبل) و جنگل تصادفی مقایسه کنید. علت پایین‌تر (یا بالاتر) بودن دقت را از نظر خود بیان کنید.

۴- مدل خود را با SVM با یک هسته غیر خطی (به انتخاب خودتان) آموزش داده و دقت آن را به دست آورید.

۵- دقت پیش‌بینی در بخش ۲ و ۴ را با هم مقایسه کرده و علت تغییر آن را ذکر کنید.