دانشگاه صنعتی شریف

دانشکده مهندسی برق

درس آنالیز داده های حجیم – پاییز 1401

تمرین سری دوم

استاد: دکتر ایمان غلامپور

قوانین تحویل:

- پاسخ به تمرینات این درس می بایست حتماً تایپ شده باشند، لذا گزارش های دست نویس تصحیح نخواهند شد.
- بخش زیادی از نمره تمرینات به گزارش و نتیجه گیری های شما اختصاص دارد، لذا در نوشتن گزارش بخش های مختلف سوالات دقت کافی را داشته باشید و تمامی نتایج را تحلیل کرده و با حوصله آن ها را ذکر کنید، سعی کنید در تحلیل های خود از نمودارها و هر visualization ابتکاری دیگر استفاده کنید، گزارش هایی که صرفاً شامل کد باشند تنها نمره programming assignmentرا خواهند گرفت.
- پاسخ های قسمت های عملی می بایست حتماً در فرمت ipynb. ب اشند، بنابراین میبایست تمامی بخش های عملی به صورت یک jupyter notebook تحویل داده شوند.
- تمام فایل های خود را در قالب یک فایل زیپ به فرمت HWn_studentNumber_Family تحویل دهید، n شماره تمرین می باشد.

قوانین تاخیر:

در کل میتوانید برای تمامی تمرینات **حداکثر 12** روز تاخیر داشته باشید و به ازای هر تمرین بیشتر از 4 روز تاخیر، مشمول کسری نمره می باشد، بطوری که بعد از روز 4ام، به ازای هر روز اضافی، 20 درصد از نمره تمرین را از دست خواهد داد.

از آنجا که تمام سیاست به کار گرفته شده در این درس کار با دیتاهای واقعی و یادگیری عملی در دنیای واقعی در کنار مطالب تئوری ست، لذا وقت خود را با کپی کردن از یکدیگر هدر ندهید، در صورتی که در گزارش ها و کد ها، شباهت های غیرعادی دیده شود، بدون تذکر، 100 نمره منفی برای طرفین در نظر گرفته می شود، لذا می توانید صرفاً از یکدیگر مشورت بگیرید یا سوالات خود را به صورت ایمیل به آدرس زیر ارسال کنید.

golnooshabd@gmail.com

سوال اول)

Confidence احتمال رخداد B در سبد است اگر سید قبلا شامل A باشد:

$$\mathrm{conf}(A \to B) = \Pr(B|A),$$

Lift به معنی احتمال رخداد A ,B با یکدیگر است، با این پیش فرض که A ,B از یکدیگر مستقل هستند .مقدارS(B) برابر ساپورت Bتقسیم بر تعداد کل سبدهاست.

$$\mathrm{lift}(A \to B) = \frac{\mathrm{conf}(A \to B)}{S(B)},$$

الف) مقدار conviction به صورت زیر تعریف میشود.مفهوم آن را شرح دهی د(چه رخدادی را توصیف میکند)

$$\mathrm{conv}(A \to B) = \frac{1 - S(B)}{1 - \mathrm{conf}(A \to B)}.$$

ب) کدام یک از سه تعریف بالا نسبت به A و B متقارن اند .تقارن یا عدم تقارن را اثبات کنید.

ج) ضعف تعریف Confidence نسبت به دو تعریف دیگر در چیست .در واقع چه المانی در این تعریف در نظر گرفته نشده است .با ذکر مثال شرح دهید(.راهنمایی :احتمال رخداد B را یکبار زیاد و یک بار کم در نظر بگیرید.)

سوال دوم)

فرض کنید ۱۰۰ سبد و ۱۰۰ آیتم داریم. هر سبد شامل همه ی آیتمهایی است که بر آنها بخش پذیر است. مثلا سبد شمارهی شش شامل آیتمهای ۱ و ۲ و ۳ و ۶ است. با فرض آنکه ساپورت ترشولد برابر ۵ باشد. Maximal frequent itemset این دیتاست را بیابید.

سوال سوم) اختیاری

یکی از روش های فروش کالای بیشتر به مشتریان حاضر در سرویس های آنلاین cross-selling نام دارد. مثالی از این روش به این صورت است که بر اساس انتخاب هایی که مشتری در سبد خرید خود داشته، آیتم اضافه پیشنهاد داده میشود. یک راه ساده برای اینکار پیشنهاد آیتم های پرتکرار باهم میباشد.

از فایل records.txt استفاده کنید تا الگوریتم a-priori را پیاده کنید. هر خط نشان دهنده یک ریکورد ثبت شده از یک مشتری میباشد. در هر خط، هر رشته ی 8 کاراکتری نشان دهنده شناسه آیتمی است که در آن، مورد بازدید قرار گرفته شده است و آیتم ها با فاصله از هم جدا شده اند؛ برخی از خطوط حاوی موارد تکراری هستند. حذف یا نادیده گرفتن موارد تکراری نباید بر نتایج شما تأثیر بگذارد. itemset ها با سایز 2 و 3 را همراه با confidence score آن ها برای support s=100 بیابید.

(در این قسمت هدف آشنایی با الگوریتم می باشد و استفاده از کتابخانه pyspark توصیه میشود اما اجباری نیست)

سوال چهارم)

در این سوال قصد داریم تا به مقایسه الگوریتم های A-priori و SON بر روی داده های واقعی ترافیک شهر تهران بپردازیم. اطلاعات لازم و لینک دانلود فایل دیتاست در CW موجود است. توجه کنید که برای حل سوال نمی توانید از الگوریتم های frequent item اسپارک استفاده کنید و حتما دیتاها به صورت rdd پردازش شوند. همچنین می توانید از دیتاست کوچک شده Sample_Data.zip استفاده کنید. خروجی این دو الگوریتم گزارش رفتارهای تکرار شونده می باشد که به صورت

Frequent_itemset از داده خام استخراج می‌شوند. برای شروع یک rdd به فرمت زیر بسازید تا در نهایت **مسیر های پرتردد و تعداد شمارش آن ها** بر اساس دوربین ها مشخص شوند.
key= (plate, date), value= [list of device codes]

گزارش شما باید حاوی موارد زیر باشد:

- خروجی، نحوه پیاده سازی و نتایج بدست آمده از الگوریتم A-priori.
- خروجی، نحوه پیاده سازی و نتایج بدست آمده از الگوریتم SON.
- مقایسه بین دو روش.
- مقدار و نحوه انتخاب ساپورت و دیگر پارامترهای استفاده شده.
- در صورت نیاز، نحوه پاک سازی داده ها از دیتای پرت.
- توضیح دهید ترکیب دوربین هایی که کنار هم قرار ندارند چگونه در این فرایند حذف میشوند و به عنوان مسیر پرتردد در خروجی ظاهر نمی‌شوند.
- بررسی کنید آیا امکان بدست آوردن رفتار های تکرار شونده دیگری مانند زمان های پرتردد و یا خوردروهای پرتردد وجود دارد و توجیح پذیر است؟ به صورت تئوری چگونگی آن را توضیح دهید.

راهنمایی اول: پیاده سازی الگوریتم A-priori در چند مرحله انجام میشود؛ ابتدا از rdd گفته شده استفاده کنید تا تعداد تردد های ثبت شده برای هر دوربین بدست آید سپس با استفاده از ساپورت مناسب، مرحله اول که یافتن تک‌مسیر های پرتردد میباشد به پایان میرسد. برای مرحله بعد، از نتیجه ذخیره شده مرحله اول (مثلا به صورت لیست یا دیکشنری) استفاده کنید و ترکیب دوتایی از مسیرهای محتمل را بسازید (با استفاده از قوانین معرفی شده در درس برخی داده ها پیش از شمارش حذف می‌شوند)؛ ادامه کار همانند مرحله اول با شمارش داده ها ادامه پیدا می‌کند. همین مراحل را برای ترکیب های چندتایی و بزرگ تر از مسیرهای پرتردد تعمیم دهید.

راهنمایی دوم: برای پیاده سازی الگوریتم SON، rdd را به دو یا سه بخش تقسیم کنید.