تحلیل کلان داده‌ها

تمرین اول

{MapReduce, Association Rule, LSH}

مهلت تحویل

۱۴۰۲/۰۱/۱۸

برای ارسال تمرین به نکات زیر توجه کنید.

۱. ملاک اصلی انجام تمرین گزارش آن است و ارسال کد بدون گزارش فاقد ارزش است. برای این تمرین یک فایل گزارش 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.
۲. زبان برنامه‌نویسی برای انجام تمرین‌ها، پایتون(Python) در نظر گرفته شده است.
۳. برای سهولت در انجام تمرینات، توصیه می‌شود که پلتفرم کولب گوگل استفاده نمایید.
۴. در نظر داشته باشید کدهای شما باید قابلیت اجرا در هنگام ارائه را داشته باشند. همچنین بر روی کدهای خود مسلط باشید
۵. کدهای ارسالی خود را برای افزایش خوانایی و درک بهتر به صورت مناسب کامنت‌گذاری کنید.
۶. پاسخ سوالات تشریحی، فایل‌های کد و گزارش خود را در یک فایل فشرده قرار داده و با نام با فرمت HW1_StudentID ارسال نمایید.
۷. درصورت وجود ابهام یا اشکال می‌توانید از طریق کانال با ما در ارتباط باشید(داخل کانال لینک گروه نیز موجود می‌باشد)

بخش اول MapReduce

۱- یکی از وظایفی که در آن از الگوی MapReduce استفاده می‌شود، ضرب ماتریس‌ها می‌باشد. بعضی مواقع حجم ماتریس به قدری بزرگ است که برای انجام محاسبات، لازم است از چند ماشین به صورت موازی استفاده شود. عمل ضرب ماتریس‌ها از دو مرحله MapReduce تشکیل شده است. به عبارتی می‌توان چند مرحله عمل MapReduce داشت به این صورت که ورودی به MapReduce اول وارد می‌شود و خروجی آن وارد عمل MapReduce بعدی شود.

با در نظر گرفتن دو ماتریس زیر به سوالاتی که در ادامه آمده است پاسخ دهید.

$$M1: \begin{bmatrix} 2 & 3 \\ 1 & 2 \end{bmatrix} \qquad M2: \begin{bmatrix} 1 & 4 \\ 2 & 3 \end{bmatrix}$$

الف) عملی که توابع Map و Reduce، در هرکدام از دو مرحله MapReduce انجام می‌دهند را به صورت مختصر شرح دهید.

ب) تمامی جفت کلید مقدار که توسط mapper اول ایجاد می‌شود را مشخص نمایید.

ج) تمامی جفت کلید مقدار، قبل از وارد شدن به reducer دوم را مشخص نمایید.

۲- در این قسمت قصد داریم با کاربرد الگوی MapReduce در دنیای واقعی آشنا شویم. معمولا کانال‌های شبکه مجازی برای بهتر دیده شدن با یکدیگر به تبادل لینک (تبلیغ) می‌پردازند. مجموعه دادگانی که در اختیار شما قرار گرفته است، هر سطرش شامل یک سری ID کانال است که ID اول مربوط به کانالی هست که تبلیغ شده است و سایر شماره‌ها ID کانال‌هایی است که کانال اول رو تبلیغ کرده‌اند. توجه کنید یک کانال در طی دوره‌های مختلف می‌تواند چندین بار با سایر کانال‌ها به تبادل لینک بپردازد. در پیاده سازی و گزارش موارد زیر را در نظر بگیرید:

- بایستی تمامی مراحل پیاده سازی با الگوی MapReduce نوشته شود.
- برای این بخش ترجیحا از PySpark که به راحتی در محیط گوگل کولب(google colab) قابل نصب است استفاده کنید.
- علاوه بر پیاده سازی روش مورد نظر نحوه عملکرد توابع Map و Reduce را در گزارش خود ذکر کنید.

الف) با استفاده از الگوی MapReduce برنامه ای بنویسید که پنج ID کانال با بیشترین تبادل لینک را در خروجی نشان دهد.

ب) تعداد تبادل لینک ID کانال با شماره‌های ۱۷۴۸، ۵۶۳۳ و ۳۴۶۹ را پیدا کنید

بخش دوم Association Rule

۱- فرض کنید آیتم‌های داریم که از ۱ تا ۱۰ شماره گذاری شده است. با استفاده از آیتم i ام هر سبد با احتمال 1/i به طور مستقل ساخته می‌شود. یعنی تمام سبدها شامل آیتم ۱، نیمی شامل آیتم ۲، یک سوم شامل آیتم ۳ و ... . فرض کنید که تعداد سبدها به اندازه کافی بزرگ است. اگر حدآستانه support را ۱٪ سبدها فرض کنیم، آیتم‌های پرتکرار را پیدا کنید. چه نتیجه‌ای از این سوال گرفته‌اید؟

۲- مجموعه داده کارت اعتباری به شما داده شده است. چیزی که در این سوال اهمیت دارد این است که شرکت‌های کارت اعتباری بتوانند تراکنش‌های جعلی کارت اعتباری را شناسایی کنند. این مجموعه داده شامل تراکنش‌های انجام شده توسط کارت‌های اعتباری در سپتامبر ۲۰۱۳ توسط دارندگان کارت اروپایی است. این مجموعه داده تراکنش‌هایی را ارائه می‌دهد که در دو روز اتفاق افتاده است که در آن ۴۹۲ تقلب از ۲۸۴۸۰۷ تراکنش اتفاق افتاده است. این مجموعه داده بسیار نامتعادل است، کلاس مثبت (تقلب ها) ۰/۱۷۲ درصد از کل تراکنش‌ها را تشکیل می‌دهد. متأسفانه، به دلیل مسائل محرمانه، ویژگی‌های اصلی و اطلاعات پس‌زمینه بیشتری درباره داده‌ها ارائه نشده است. ویژگی‌های V1، V2 ... V28 مولفه‌های اصلی به‌دست‌آمده با PCA[1] هستند، تنها ویژگی‌هایی که با PCA تبدیل نشده‌اند، «Time» و «Amount» است. ویژگی «Class» متغیر پاسخ است و در صورت تقلب مقدار ۱ و در غیر این صورت ۰ می‌گیرد. در این سوال از شما می‌خواهیم تا با استفاده از قوانین انجمنی، طبقه‌بندی دو کلاسه انجام دهید. یکی از چالش‌های اصلی حل این سوال، مقادیر عددی ویژگی‌ها است. در اینجا شما باید داده‌های عددی را گسسته کنید. اینکه با چه روشی اینکار را انجام می‌دهید و طول بازه‌ها چطور تعیین می‌کنید، در حل مسئله تاثیر گذار است. مواردی که از شما در این سوال می‌خواهیم به شرح زیر است:

- فرایند گسسته کردن داده‌ها و پیش‌پردازش‌های لازم را توضیح دهید.
- با استفاده از الگوریتم apriori این مسئله را حل کنید. طول قوانین انجمنی که باید گزارش کنید ۳ و ۴ است. به این صورت که سمت چپ قوانین ۲ یا ۳ ویژگی و سمت راست قوانین کلاس مورد نظر است.
- حد آستانه support برای فیلتر کردن الگوهای پرتکرار در هر مرحله مسئله مهمی است. این مقدار را در هر مرحله چگونه تعیین کردید؟ توضیح دهید.
- به ازای قوانین ۳ و ۴ تایی، ۵ قوانین برتر با معیار confidence و interest را برای هر کلاس گزارش کنید.
- در نهایت با استفاده از قوانینی که استخراج کردید صحت طبقه‌بندی برای هر کلاس را گزارش کنید. دقت کنید شما می‌توانید از ترکیب قوانینی که استخراج کردید این ارزیابی را انجام دهید (مثلا از همان ۵ قوانین برتر که گزارش کردید). هدف این است که شما با یکسری از قوانین اگر، آنگاه برای داده‌ها تصمیم بگیرید که به چه کلاسی تعلق دارد.

---

[1] **PCA، روشی برای کاهش ابعاد است.**

بخش سوم Locality Sensitive Hashing

۱- مهم ترین محدودیت LSH برای استفاده چیست ؟

۲- ثابت کنید Min-Hashing قابلیت شباهت نگه‌دار (Similarity preserving) دارد.

۳- در این قسمت هدف این است که با استفاده از LSH شباهت بین جملات را پیدا کنید. در این قسمت یک فایل lsh.ipynb در اختیار شما قرار داده شده است که شامل ۹ تابع است. ۳ تابع از این ۹ تابع از پیش نوشته شده است و وظیفه شما این است که ۶ تابع باقی مانده را تکمیل کنید. ورودی و خروجی هر تابع مشخص شده است. علاوه بر این در صورت نیاز نمونه ای از خروجی مورد انتظار نیز در تابع مربوطه آورده شده است. هر تابع مربوط به یک قسمت از LSH می‌شود به عنوان مثال تابع create_shingle برای ساختن k-shingle از روی داده‌های مورد نظر استفاده می‌شود.(در صورتی که Jupyter notebook را به صورت local بر روی سیستم خود نصب ندارید می‌توانید از محیط Google Colab استفاده کنید.)