درس کلان داده‌ها

تمرین اول

به نام خدا

سوال ۱) Association rules

الف) با در نظر گرفتن سبدهای خرید زیر، به سوالات پاسخ دهید:

۱	{شیر, نوشابه، دستمال}
۲	{نان، کره، شیر}
۳	{شیر، دستمال، بیسکوییت}
۴	{نان، کره، بیسکوییت}
۵	{نوشابه، دستمال، بیسکوییت}
۶	{شیر، دستمال، نان، کره}
۷	{شیر، کره، دستمال}
۸	{نوشابه، دستمال}
۹	{شیر، دستمال، نان، کره}
۱۰	{نوشابه، بیسکوییت}

۱) چه تعداد association rule متفاوت (حتی با support صفر) می‌توان از این سبدهای خرید استخراج کرد؟
۲) بیشترین اندازه‌ی ItemSet ای که می‌توانیم استخراج کنیم چقدر است؟
۳) عبارتی برای بیشینه‌ی تعداد ItemSet هایی با اندازه‌ی ۳ که می‌توانیم از این اطلاعات استخراج کنیم بنویسید.
۴) کدام ItemSet با اندازه‌ی حداقل ۲، بیشترین support را دارد؟
۵) یک جفت association rule به فرم A→B و B→A بیابید که confidence یکسانی داشته باشند.

ب) کدام یک از معیارهای ارائه شده در زیر متقارن هستند؟ با ذکر مثال نقض و یا اثبات مشخص نمایید که آیا یک معیار متقارن است یا خیر. (یک معیار زمانی متقارن است که داشته باشیم $measure(A \longrightarrow B) = measure(B \longrightarrow A)$)

۱) ***confidence***: احتمال رخ دادن B در سبد خرید، به شرط که سبد خرید در حال حاضر دارای آیتم A باشد.

$$conf(A \longrightarrow B) = \Pr(B|A)$$

۲) ***Lift***: این معیار نسبت دو حالت زیر را با یکدیگر می‌سنجد

الف) مواردی که A و B بهمراه یک دیگر رخ داده اند.

ب) حالت مورد انتظار در صورتیکه A و B از نظر آماری از یکدیگر مستقل باشند.

$$lift(A \longrightarrow B) = \frac{conf(A \longrightarrow B)}{S(B)}$$

که در آن $S(B) = \frac{support(B)}{total\ number\ of\ transactions}$.

۳) ***conviction***: این معیار دو حالت زیر را با یکدیگر مقایسه می‌کند

الف) احتمال اینکه A بدون B رخ دهد با فرض اینکه این دو مستقل از یکدیگر هستند.

ب) فرکانس واقعی رخ دادن A بدون B.

$$conv(A \longrightarrow B) = \frac{1 - S(B)}{1 - conf(A \longrightarrow B)}$$

سوال ۲) کاربرد Association rules

یکی از راه‌های توصیه اقلام در وب سایت‌ها پیشنهاد اقلامی است که معمولا به همراه یکدیگر جستجو یا خریداری شده‌اند. با استفاده از الگوریتم A-priori روی دادگان browsing.txt، یک برنامه توصیه‌گر بنویسید که اقلامی که به طور توام جستجو شده‌اند را پیشنهاد بدهد. در هر خط browsing.txt هر ۸ کاراکتر بیانگر آیدی اقلام است که با space از هم جدا شده‌اند.

الف) در این بخش، جفت آیتم‌هایی را که دارای support حداقل ۱۰۰ هستند یافته و مقدار confidence را با توجه به قوانین زیر محاسبه نمایید. در انتها به ترتیب نزولی آن‌ها را مرتب سازی نموده و ۵ قانون ابتدایی را گزارش دهید (در صورت برابری، از ترتیب بندی lexicographical استفاده نمایید).

$$rule\ number\ one:\ \ X \Longrightarrow Y$$

$$rule\ number\ two:\ \ Y \Longrightarrow X$$

ب) جفت‌های سه تایی را که دارای support حداقل ۱۰۰ هستند یافته و مقدار confidence آن‌ها را طبق قوانین ارائه شده در زیر بیابید. در انتها آن‌ها را به صورت نزولی مرتب سازی کرده و لیستی از ۵ مورد ابتدایی گزارش دهید.

$$rule\ number\ one: (X.Y) \Longrightarrow Z$$

$$rule\ number\ two: (X.Z) \Longrightarrow Y$$

$$rule\ number\ three: (Y.Z) \Longrightarrow X$$

سوال ۳) LSH

در این سوال می‌خواهیم با یکی از کاربردهای LSH برای یافتن همسایگان نزدیک، آشنا شویم. فرض کنید که مجموعه ای از n داده در فضای داده‌ای مفروضی داریم که فاصله‌ی بین هر جفت داده از این مجموعه از طریق تابع d محاسبه می‌گردد. مسئله‌ی یافتن همسایه‌ی نزدیک با تقریب $(c.\lambda)$ بدین صورت تعریف می‌شود که با داشتن یک نقطه a و با فرض داشتن نقطه‌ی مفروضی به نام x که فاصله‌ی آن از داده‌ی a کوچکتر یا برابر مقدار λ است $(d(a.x) \leq \lambda)$، نقطه‌ی دیگری با نام x' در فضای داده باز می‌گرداند که فاصله‌ی آن با نقطه‌ی a برابر است با $d(x'.a) \leq c\lambda$. پارامتر c در این تعریف فاکتور تقریب مجاز بیشینه نامیده می‌شود.

فایل patches.csv برای حل این سوال فراهم شده است. هر ردیف در این مجموعه داده یک تصویر ۲۰*۲۰ است که توسط یک بردار ۴۰۰ بعدی بازنمایی شده است. با استفاده از معیار فاصله‌ی L_1 شباهت میان تصاویر را تعیین کرده، می‌خواهیم میزان کارایی تقریب با استفاده از LSH را با روش جستجوی خطی مقایسه نماییم. می‌توانید از کد فراهم آمده بهمراه دیتاست استفاده نمایید.

توضیحات کد: در کد اولیه ارائه شده در این تمرین، مواردی که میبایست توسط شما تکمیل گردند توسط **To do** مشخص شده‌اند. شما میبایست از توابع راه اندازی و جستجو استفاده کرده و تابع جستجوی خطی خود را پیاده سازی نمایید. می‌توانید از پارامترهای پیش فرض برای این تمرین که برابرند با L=10، k=24 استفاده نمایید؛ هر چند دست شما برای استفاده از هر مقدار دیگری برای این تمرین باز است مادامی که دلیل خود را برای انتخاب آن‌ها ذکر نمایید.

الف) برای هر یک از موارد ستون‌های ۱۰۰، ۲۰۰، ۳۰۰ تا هزار، ۳ مورد همسایه‌ی نزدیک را با استفاده از هر دو روش LSH و جستجوی خطی بدست آورید. میانگین زمان جستجوی خود را برای هر یک از این دو مورد ذکر کرده و با هم مقایسه نمایید.

ب) با فرض اینکه $\{z_j \mid 1 \leq j \leq 10\}$ مجموعه تصاویر مورد نظر ما که در آن z_j تصویری است از ستون $100j$ باشد و $\{x_{ij}^*\}_{i=1}^3$ سه همسایه ی نزدیک درست z_j باشند که از روش جستجوی خطی بدست آمده اند، میزان خطای زیر را گزارش دهید.

$$error = \frac{1}{10}\sum_{j=1}^{10}\frac{\sum_{i=1}^{3} d(x_{ij}.z_j)}{\sum_{i=1}^{3} d(x_{ij}^*.z_j)}$$

ج) نمودار مقدار خطا را یکبار به صورت تابعی از L (L=10, 12, …, 20) و با ثابت نگاه داشتن مقدار k برابر با k=24) و یکبار به صورت تابعی از k (k=16, 18, 20, 22, 24) و با ثابت نگاه داشتن مقدار L برابر با L=10) رسم نموده، مقادیر را گزارش نمایید. به طور خلاصه نمودارها را تحلیل نمایید.

د) با استفاده از هر یک از دو روش مورد مقایسه در این سوال، ۱۰ همسایه‌ی نزدیک برای تصویر موجود در صدمین ستون یافته و بهمراه خود تصویر رسم نمایید. در انتها این دو روش را از این منظر مقایسه نمایید.

سوال ۴) جریان داده

در این سوال تعداد تکرار عناصر مختلف در جریان داده بررسی خواهد شد. فرض کنید $S = <a_1,a_2,\ldots,a_t>$ جریان داده‌ای از مجموعه $\{1.2. \ldots . n\}$ بوده و برای هر $1 \le i \le n$، $F[i]$ تعداد دفعات ظاهر شدن i در S باشد. هدف در این مسئله تخمین مناسب $F[i]$ در تمامی زمان‌هاست. یک راه ساده برای این مسئله ذخیره کردن تمام عناصر داده است که نیازمند فضای ذخیره از مرتبه $O(n)$ است، اختصاص چنین مرتبه‌ای از حافظه برای بسیاری از کاربردها امکان‌پذیر نیست. در ادامه خواهیم دید با فضای ذخیره بسیار کمتر قادر خواهیم بود تا با استفاده از الگوریتمی که در زیر شرح داده شده است تخمین مناسبی از فرکانس عناصر جریان داده داشته باشیم. الگوریتم شامل دو پارامتر $\varepsilon,\delta > 0$ بوده و به تعداد $\left\lceil log\frac{1}{\delta}\right\rceil$ هش مستقل نیازمند است.

$$\forall j \in \left[\!\left[1; \left\lceil log\frac{1}{\delta}\right\rceil\right]\!\right].\ h_j: \{1.2. \ldots . n\} \to \left\{1.2. \ldots . \left\lceil\frac{e}{\epsilon}\right\rceil\right\}$$

لگاریتم در فرمول بالا لگاریتم طبیعی‌ست و همچنین به شمارنده‌های $c_{j.x}$ برای هر $1 \le j \le \left\lceil log\frac{1}{\delta}\right\rceil$ و $1 \le x \le \left\lceil\frac{e}{\epsilon}\right\rceil$ نیاز داریم. در ابتدای جریان داده همه این شمارنده‌ها با مقدار صفر مقداردهی می‌شوند، سپس با سر رسیدن اولین عنصر جریان داده مانند a_k که $1 \le k \le t$، متغیر $c_{j.h_j(a_k)}$ یک واحد اضافه می‌شود. برای هر $1 \le i \le n$ فرکانس تخمینی از فرمول زیر به دست می‌آید:

$$\tilde{F}[i] = \min_j\{c_{j.h_j(i)}\}$$

الگوریتم را پیاده‌سازی کرده، روی دادگان مرتبط با این سوال اجرا کنید. پارامترهای مرتبط با الگوریتم را با مقادیر $\delta = e^{-5}$ و $\epsilon = e \times 10^{-4}$ مقداردهی کنید. با توجه به مقدار انتخاب شده برای δ، به ۵ تابع هش نیاز خواهید داشت. داده‌ی کافی برای توابع هش مورد نیاز در فایل hash_params.txt موجودند. در مرحله بعد برای هر کلمه در دیتاست relative error را طبق زیر محاسبه کرده، نمودار خطا بر اساس فرکانس واقعی هر عنصر $(\frac{F[i]}{t})$ را رسم کنید، توجه داشته باشید که راستای x و y هر دو در مقیاس لگاریتمی در پایه ۱۰ باشد. مقادیر واقعی را می‌توانید از فایل counts بخوانید. با توجه نمودار بدست آمده، به نظر شما یک شرط تقریبی روی فرکانس واقعی عناصر که بتواند relative error همواره کمتر از یک نگه دارد، چیست؟

$$E_r[i] = \frac{\tilde{F}[i] - F[i]}{F[i]}$$

برای توابع هش مورد نیاز جهت حل این سوال، می‌توانید از سودوکد تابع هش نوشته شده در زیر استفاده کنید:

در این تابع مقدار p را برابر با ۱۲۳۴۵۷ قرار داده، برای پارامترهای a و b می‌توانید از داده های فایل hash_params استفاده نمایید.

Hash function template

```
def hash_function (a, b, p, n_buckets, x)
    y = x [modulo] p
    hash_val = (a*y + b) [modulo] p
    return hash_val [modulo] n_buckets
```

توضیحات

- **مهلت تحویل تمرین: ۱۴۰۰/۲/۲۰**
- زمان ارائه: پس از تحویل مشخص خواهد شد.

نکاتی در مورد تحویل تمرین:

- ✓ فایل داده‌های مورد نیاز تمرین‌ها را می‌توانید از این لینک دریافت نمایید.
- ✓ خروجی کد ها و نتایج سوالات را درون گزارش بنویسید و از توضیح اضافی کد و موارد دیگر خودداری فرمایید (کد بدون گزارش ارزشی ندارد).
- ✓ فرمت تحویل: برای هر سوال یک پوشه‌ی جداگانه در نظر گرفته، کد و مواردی از قبیل خروجی برنامه و نمودارها را در آن ذخیره نمایید. این پوشه‌ها به همراه یک فایل studentid_report.pdf برای گزارش و توضیح سوالات، درون یک فایل فشرده شده با فرمت zip. و یا rar. باشد. فرمت‌های دیگر پشتیبانی نمی‌شوند.
- ✓ در نظر داشته باشید کد های شما باید قابلیت اجرا در هنگام ارائه داشته باشند. همچنین بر روی کد های خود کاملا مسلط باشید.
- ✓ توصیه میشود پیش از ارایه نیز مطالعه‌ای روی کد خود داشته باشید تا سواالت تدریس یاران را به راحتی پاسخ دهید.
- ✓ می‌توانید از گوگل برای رفع سوالات و مشکلات خود استفاده نمایید. در صورت رفع نشدن مشکل، می‌توانید سوالات خود را با تدریس‌یاران درس از طریق ایمیل زیر در میان بگذارید.

- ایمیل تدریس‌یاران درس:

bdta00@gmail.com