درس کلان داده‌ها

تمرین دوم

نیم سال دوم تحصیلی ۱۳۹۹-۱۴۰۰
دانشگاه صنعتی امیرکبیر

به نام خدا

سوال ۱) پیاده سازی k-means با استفاده از spark

در این سوال می‌خواهیم با استفاده از داده‌های موجود برای این بخش و با استفاده از اسپارک، الگوریتم تکرار شونده‌ی k-means را پیاده سازی نماییم. فایل داده مورد نیاز برای این بخش حاوی ۴۶۰۱ سطر است که هر سطر آن بیانگر سندی است که با استفاده از یک بردار ویژگی ۵۸ بُعدی بازنمایی شده است. فایل‌های c1 و c2 نیز به ترتیب حاوی سنترویدهای اولیه k خوشه هستند که در c1 به صورت رندوم تعیین شده و در c2 این سنترویدها تا حد ممکن، با در نظر گرفتن معیار فاصله ی اقلیدسی، از یکدیگر فاصله دارند.

برای تمام قسمت‌های این بخش، حداکثر تعداد تکرارها را برابر با ۲۰ و تعداد خوشه‌ها را برابر با ۱۰ در نظر بگیرید.

الف) با در نظر گرفتن معیار فاصله‌ی اقلیدسی، برای هر تکرار تابع هزینه را محاسبه نمایید. این عمل بدین معنی است که میبایست برای تکرار نخست از مقادیر پیش فرض یکی از فایل‌های c1 و c2 استفاده کنید. الگوریتم k-means را روی داده‌های فایل data با استفاده از مقادیر فایل‌های c1 و c2 اجرا کرده سپس مقادیر بدست آمده تابع هزینه را به صورت نمودار بر حسب تکرار از ۱ تا ۲۰ برای هر یک از دو مورد c1 و c2 رسم نمایید.

ب) درصد تغییر هزینه الگوریتم k-means بین اجرای صفرم و اجرای دهم را طبق عبارت $\frac{cost[0]-cost[10]}{cost[0]}$، با استفاده از مقادیر اولیه سنترویدها در c1 و c2، با یکدیگر مقایسه نمایید (معیار فاصله را در این سوال اقلیدسی در نظر بگیرید). توضیح دهید که کدام یک مقداردهی اولیه بهتری داشته است.

ج) مورد الف را این بار با در نظر گرفتن فاصله‌ی منهتن به عنوان تابع هزینه تکرار نمایید.

د) درصد تغییرات بین اجرای صفرم و دهم را این بار با در نظر گرفتن فاصله‌ی منهتن بدست آورده و نتایج را برای دو فایل c 1 و c2 مقایسه نمایید.

الگوریتم k-means تکرار شونده برای این تمرین در زیر ارائه شده است.

Algorithm 1 Iterative k-Means Algorithm

```
procedure   Iterative k-Means
  Select k points as initial centroids of the k clusters.
  for iterations := 1 to MAX_ITER do
      for cach point p in the dataset do
          Assign point p to the cluster with closest centroid
      end for
      Calculate the cost for this iteration.
      for each cluster c do
          Recompute the centroid of c as the mean of all the data points assigned to c
      end for
  end for
end procedure
```

سوال ۲) Singular value decomposition

الف) اثبات کنید که مقادیر ویژه غیر صفر MM^T با مقادیر ویژه غیر صفر M^TM برابرند. همچنین بحث کنید که آیا بردارهای ویژه این دو نیز با یکدیگر برابرند یا خیر؟

ب) SVD ماتریس M را به حاصلضرب $U\Sigma V^T$ میشکند که در آن دو ماتریس U و V، به صورت ستونی orthonormal بوده و Σ نیز ماتریسی است قطری. با داشتن عبارت بالا برای ماتریس M، عبارتی ساده سازی شده و مشابه را به صورت V، V^T و Σ بدست آورید.

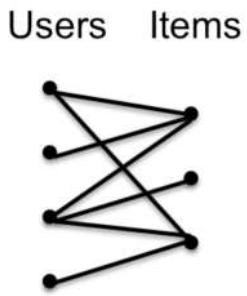
ج) ماتریس زیر را در نظر بگیرید.

$$M = \begin{bmatrix} 1 & 2 \\ 2 & 1 \\ 3 & 4 \\ 4 & 3 \end{bmatrix}$$

برنامه‌ای بنویسید که SVD ماتریس M را محاسبه نماید. مقادیر بازگشتی برای ماتریس‌های U، Σ و V^T را گزارش نمایید. سپس تجزیه‌ی مقدار ویژه حاصلضرب ماتریسی M^TM را محاسبه کرده، مقادیر ویژه بدست آمده را به صورت نزولی مرتب کرده، لیست بردارهای ویژه را نیز مرتب سازی نمایید به صورتی که بردار ویژه متناظر با بزرگترین مقدار ویژه در ابتدای لیست باشد، لیست هر دو را گزارش دهید. در نهایت ارتباط میان ماتریس V و U بدست آمده با استفاده از تجزیه SVD ماتریس M و ماتریس‌های بردارها و مقادیر ویژه گزارش شده در این بخش را بیان نمایید.

سوال ۳) سیستم‌های توصیه‌گر

یک گراف دو بخشی کاربر-اقلام را در نظر بگیرید، در صورتی که بین کاربر U و قلم جنس I یک یال وجود داشته باشد، می‌گوییم کاربر U قلم جنس I را می‌پسندد. ماتریس نمرات برای این مجموعه از کاربرها و اقلام را با R نمایش می‌دهیم؛ هر سطر از R مربوط به یک کاربر بوده و هر ستون متعلق به یکی از اقلام است. در صورتی که کاربر i قلم جنس j را بپسندد $R_{ij}=1$ و در غیر این صورت $R_{ij}=0$ خواهد بود. فرض کنید m کاربر و n قلم جنس وجود دارد. حال ماتریس P را یک ماتریس قطری m*m تعریف می‌کنیم، iامین عنصر قطری این ماتریس بیانگر تعداد عناصری است که کاربر i پسندیده است، به همین طریق ماتریس Q را یک ماتریس قطری n*n تعریف می‌کنیم که iامین عنصر قطری آن تعداد کاربرانی هستند که قلم جنس iام را پسندیده‌اند. شکل زیر را به عنوان مثال مشاهده کنید.



$$R=\begin{bmatrix} 1 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

$$P=\begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad Q=\begin{bmatrix} 3 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 3 \end{bmatrix}$$

الف) ماتریس شباهت کاربران را به صورت $T=R.R^T$ تعریف می‌کنیم. توضیح دهید T_{ii} و T_{ij} چه مفهومی دارد؟

ب) حال ماتریس شباهت اقلام را تعریف می‌کنیم، S_I یک ماتریس n*n خواهد بود، عنصر موجود در سطر i و ستون j بیانگر شباهت کسینوسی بین اقلام i و j است. نشان دهید S_I را می‌توان به فرم معادله (۴-۱) نوشت. توجه کنید $Q^{-1/2} = 1/\sqrt{Q}$.

$$S_I = Q^{-1/2} R^T R Q^{-1/2} \qquad (۴-۱)$$

در ادامه همین سوال را برای ماتریس شباهت کاربران تکرار کنید، در اینجا عنصر سطر iام و ستون jام ماتریس S_U بیانگر شباهت کسینوسی میان کاربر iام و jام است. توجه کنید جواب نهایی باید یک عبارت بر اساس ماتریس‌های R، P و Q باشد.

ج) روش توصیه بر اساس user-user collaborative filtering برای کاربر u می‌تواند به این صورت تعریف شود که برای همه اقلام مانند s معادله (۴-۲) حل شود و k قلم از اجناس که بیشترین $r_{u,s}$ را داشته باشند به عنوان جواب بازگردانده شود.

$$r_{u.s} = \sum_{x \in users} \cos_sim(x.u) \times R_{xs} \qquad (۴-۲)$$

به طور مشابه با استفاده از روش توصیه بر اساس item-item collaborative برای قلم جنس u با استفاده از معادله (۴-۳) k قلم جنس جهت توصیه انتخاب می‌شود.

$$r_{u.s} = \sum_{x \in items} R_{xs} \times \cos_sim(x.s) \qquad (۴-۳)$$

حال ماتریس توصیه به نام Γ را اینگونه تعریف می‌کنیم که $\Gamma_{(i.j)} = r_{i.j}$. برای هر دو روش توصیه (اقلام-اقلام و کاربر-کاربر) ماتریس Γ را بر اساس R، P و Q به دست آورید.

د) در این قسمت روش‌های گفته شده را روی دادگان واقعی پیاده‌سازی می‌کنیم. دادگان مربوط به این قسمت شامل اطلاعات برنامه‌های تلوزیونی است، شامل ۹۹۸۵ کاربر و ۵۶۳ برنامه معروف. فایل user-shows.txt ماتریس R است که هر سطر آن نشانگر یک کاربر و هر ستون آن مرتبط با یک برنامه است، در صورتی که یک کاربر یک برنامه را بیشتر از سه ماه تماشا کند، درایه متناظر با برنامه و کاربر یک می‌شود. ستون‌ها با space جدا شده‌اند. فایل shows.txt شامل اسامی برنامه‌ها به همان ترتیب ستون‌های ماتریس R.

روش‌های توصیه user-user collaborative filtering و item-item collaborative filtering برای کاربر ۵۰۰ام از دادگان مقایسه خواهند شد، این کاربر را Alex می‌نامیم، توجه کنید ایندکس این کاربر در دادگان ۴۹۹ است. برای انجام این کار ۱۰۰ ورودی را برای سطر الکس صفر می‌کنیم، بدین معنی که دیگر نمی‌دانیم الکس کدام برنامه‌ها را تماشا کرده است. بر اساس رفتار الکس با دیگر برنامه‌ها ۱۰۰ پیشنهاد برای الکس پیدا می‌کنیم و سپس با برنامه‌های واقعی(که قبلا حذفشان کرده بودیم) مقایسه می‌کنیم تا میزان تطابق را به دست آوریم.

۱- ماتریس‌های P و Q را محاسبه کنید.

۲- با استفاده از فرمول‌های قسمت (ج)، Γ را برای user-user collaborative filtering محاسبه کنید. فرض کنید S مجموعه صد برنامه اول است، از تمام برنامه‌های موجود در S، ۵ برنامه که بیشترین شباهت با الکس را دارد پیدا کنید(نام برنامه را ذکر کنید و در صورتی که دو برنامه امتیاز یکسان داشتند، برنامه با ایندکس کمتر را انتخاب کنید).

۳- Γ را برای movie-movie collaborative filtering محاسبه کنید، از میان همه برنامه‌های تلوزیونی در مجموعه S مانند حالت قبل ۵ برنامه با بیشترین امتیاز را برای الکس گزارش کنید. علاوه بر این امتیاز برنامه‌های انتخاب شده را نیز گزارش کنید.

توضیحات

- **مهلت تحویل تمرین: ۱۴۰۰/۳/۲۶**
- زمان ارائه: پس از تحویل مشخص خواهد شد.

نکاتی در مورد تحویل تمرین:

- ✓ فایل داده‌های مورد نیاز تمرین‌ها، به همراه فایل تمرین در سامانه بارگذاری شده است.
- ✓ 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 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 را داشته باشند. همچنین بر روی کد های خود کاملا مسلط باشید.
- ✓ توصیه میشود پیش از ارائه نیز مطالعه‌ای روی کد و گزارش خود داشته باشید تا سوالات 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