دانشکده مهندسی برق

تحلیل داده‌های حجیم

تمرین سری سوم

استاد درس: دکتر ایمان غلامپور

طراح تمرین: حسین شریفی

بخش تئوری و تحقیق:

سوال اول)

تابع هزینه الگوریتم K-means به صورت زیر تعریف می‌شود:(منظور از S_j مجموعه‌ای از نمونه‌هاست که به مرکز μ_j نزدیکتر از هر خوشه دیگر هستند.)

$$L = \sum_{j=1}^{k} \sum_{x_i \in Sj} ||x_i - \mu_j||^2$$

الف) نشان دهید که برای کمینه کردن تابع هزینه کافی است میانگین هر خوشه به‌عنوان مرکز آن انتخاب گردد.

ب) اثبات کنید همواره این الگوریتم همگرا می‌شود و با رسم شکل نشان دهید، ممکن است این الگوریتم به نقطه local optimum همگرا شود.

ج) یکی از روش‌های پیدا کردن مقدار K بهینه، رسم تابع هزینه و پیدا کردن نقطه Knee می‌باشد. در مورد آن توضیح دهید و حداقل دومورد از روش‌های دیگر تخمین مقدار K شرح دهید.

د) می‌توان برای تولید ضرر بیشتر در تعداد کلاسترهای بالا تابع هزینه تغییر داد. یک تابع هزینه جدید پیشنهاد دهید. مزایا و معایب استفاده از این تابع هزینه جهت آموزش چیست؟

سوال دوم)

در الگوریتم خوشه‌بندی CURE، نقاط نماینده‌ی یک خوشه به میزان معینی از فاصله بین موقعیت اولیه‌شان و مرکز خوشه جابه‌جا می‌شوند. بهتر نیست که همه‌ی این نقاط به یک فاصله ثابت به سمت مراکز حرکت کنند؟ توضیح دهید.

سوال سوم)

الف) با توجه به تجزیه SVD، یک تجزیه مقدار منفرد برای A^T پیدا کنید. مقادیر منفرد A و A^T چه ارتباطی باهم دارند؟

ب) برای هر ماتریس $A_{n\times n}$، با استفاده از تجزیه SVD نشان دهید که یک ماتریس متعامد $Q_{n\times n}$ وجود دارد به طوری که: $A^TA = Q^TA^TAQ$

سوال چهارم)

الف) فرض کنید ماتریس $A_{100\times100}$ با $\| A \|_F = 50$ باشد. ۱۰ ستون به صورت تصادفی برای ساخت ماتریس C انتخاب می‌کنیم. اگر احتمال‌های نمونه‌گیری متناسب با مربع نرم ستون‌های ماتریس A باشد، خطای $\| A - CC^{\dagger}A \|_F$ را با فرض $\varepsilon = 0.1$ بدست آورید.(فرض کنید رنک ماتریس تخمین زده شده از تعداد ستون های A بسیار کمتر می‌باشد.($k \ll n$))

ب) دو ماتریس $A_{100\times100}$ و $B_{100\times100}$ درنظر بگیرید. قصد داریم حاصل AB با استفاده از نمونه‌گیری تصادفی با استفاده از ماتریس‌های C و R محاسبه کنیم. فرض کنید $s = 50$ (تعداد عبارت‌های نمونه‌برداری شده برای تقریب AB) و سطرها و ستون‌ها با احتمال‌هایی متناسب با نرم‌های مربع سطرها و ستون‌های A و B انتخاب می‌شوند. اگر $\| A \|_F = 50$ ، $\| B \|_F = 80$ و$0.01 \leq \delta$ خطای $\| AB - CR \|_F$ بدست آورید.(برای حل این سوال به SVDCUR.pdf در اسلایدهای بخش Dimensionality Reduction مراجعه کنید.)

سوال پنجم)

در این سوال قصد داریم آیتم‌هایِ ماتریسutility زیر را خوشه بندی کنیم.

	a	b	c	d	e	f	g	h
A	4	5		5	1		3	2
B		3	4	3	1	2	1	
C	2		1	3		4	5	3

الف) قصد داریم هشت تا از آیتم‌ها را به صورت hierarchically به چهار کلاستر تقسیم کنیم. برای خوشه‌بندی به صورت زیر عمل کنید:

- تمام اعداد 3،4و5 با 1 جایگزین کنید.
- تمام اعداد 1،2 و جاهای‌خالی را با 0 جایگزین کنید.
- از Jaccard distance برای اندازه‌گیری فاصله بین بردارهای ستونی حاصل استفاده کنید.
- برای خوشه‌هایی با بیش از یک عنصر، فاصله بین خوشه‌ها را حداقل فاصله بین جفت عناصر درنظر بگیرید.

ب) یک ماتریس جدید بسازید که سطرهای آن مربوط به کاربران باشد(مشابه قبل) و ستون‌های آن مربوط به خوشه‌ها باشد. مقدار هر درایه ماتریس جدید را با محاسبه میانگین مقادیر عددی(غیرخالی) برای آن کاربر و تمام آیتم‌های موجود در کلاستر به‌دست آورید.

ج) فاصله کسینوسی را برای هر جفت کاربر با استفاده از ماتریسی که در قسمت (ب) ساخته‌اید، محاسبه کنید.

بخش عملی:

سوال اول)

در این قسمت قصد داریم به بررسی Dimensionality reduction و Clustering بپردازیم. در ابتدا فایل MDA_P1.rar از حالت فشرده خارج کنید و فایل covtype.info جهت آشنایی با دیتاست مطالعه کنید.(توجه کنید که ستون اول صرفا شماره هر سمپل می‌باشد.)

الف) با استفاده از الگوریتم PCA، به رسم واریانس Principal components به استفاده از رابطه$\frac{\sum_{i=1}^{k}\lambda_i}{\sum_{i=1}^{d}\lambda_i}$ بپردازید. در ادامه مقدار k به نحوی انتخاب کنید که حداقل ۹۰ درصد از واریانس سمپل‌ها حفظ شود و در نهایت با استفاده از eigenvector های بدست آمده ابعاد نمونه‌هارا کاهش دهید.(می‌توانید از سایر الگوریتم‌های کاهش بعد جهت بهبود نتیجه قسمت (ب) استفاده کنید.)

ب) در این قسمت داده هارا به سه چانک تقسیم کنید و با استفاده از یکی از الگوریتم هایBRF یا Cure داده‌ها را به ۷ کلاستر گروه‌بندی کنید.(همه موارد طراحی از جمله نحوه نرمالایز کردن داده‌ها، نحوه initialize کردن اولیه کلاسترها و ... در گزارش خود بنویسید.)

ج) مقدار دو متریک زیر را برای ارزیابی کلاسترینگ صورت گرفته محاسبه کنید و بر اساس الگوریتمی که برای کلاسترینگ در قسمت (ب) انتخاب کردید، توضیح دهید کدام یک جهت ارزیابی کلاسترینگ مناسب تر می‌باشد. برای هر متریک توضیح دهید چه معیاری را ارزیابی می‌کند و رنج آن را مشخص کنید.

Silhouette Score:

$$S(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

$a(i)$: میانگین فاصله سمپل i ام تا سمپل‌های هم‌کلاستر

$b(i)$: میانگین فاصله سمپل i ام تا نقاط نزدیک‌ترین کلاستر

Davies-Bouldin Index:

$$DBI = \frac{1}{K}\sum_{i=1}^{K}\max_{i\neq j}\left(\frac{\sigma_i + \sigma_j}{dij}\right)$$

σ_i: میانگین فاصله سمپل‌ها در کلاستر i تا مرکز

dij: فاصله بین مرکز کلاستر i و j

K: تعداد کلاستر (۷)

د) در این قسمت، نتایج کلاسترینگ را در ستون‌های فایل Evaluation_P1 ذخیره کنید. (بخشی از نمره این تمرین به نتایج این بخش اختصاص دارد.)

سوال دوم)

در این بخش، سیستم‌های توصیه‌گر (recommender systems) را بر اساس امتیازات کاربران به کتاب‌های آمازون بررسی می‌کنیم. در ابتدا دیتاست از [اینجا] دانلود کنید. این دیتاست شامل ۴ ستون است:

- آیدی کاربر
- آیدی کتاب
- امتیاز
- زمان ثبتِ امتیاز

الف) در ابتدا بخشی از دیتاست به عنوان داده تست جدا کرده و در ادامه برای باقی دیتاست، بر اساس یکی از روش‌های Collaborative یا Latent factor based الگوریتم خود را طراحی و پیاده‌سازی کنید. تمام جزئیات درنظر گرفته‌شده و مراحل طراحی را گزارش دهید. برای نوشتن بخش LSH الگوریتم می‌توانید از توابع آماده استفاده کنید ولی باقی قسمت‌ها نیاز به کدنویسی از پایه دارند و درصورتی که نیاز به بهینه کردن تابع هزینه در الگوریتم خود دارید استفاده از کتابخانه‌ی MLlib مجاز نمی‌باشد.

ب) پس از طراحی سیستم توصیه‌گر، خطای RMSE برای بخش تست بدست آورید. می‌توانید جهت بهبود عملکرد قسمت (ج)، الگوریتم خود را به نحوی تغییر دهید که به کاهش این خطا منجر شود.(الگوریتم اولیه به همراه نتایج و مراحل تغییر منجر به کاهش این خطا گزارش کنید.)

ج) در این بخش باید فایل Evaluation_P2 که در پوشه MDA_P2 قرار دارد را تکمیل کنید. کافی است برای هر آیدی یوسر/کتاب داده شده، امتیاز تخمین زده شده بجای '?' وارد کنید.(بخشی از نمره این تمرین به نتایج این بخش اختصاص دارد.)