دانشگاه صنعتی شریف

دانشکده مهندسی برق

درس آنالیز داده ها ی حجیم

تمرین سری سوم

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 assignment را خواهند گرفت.
- پاسخ های قسمت های عملی می بایست حتماً در فرمت ipynb. باشند، 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 ۱۲** روز تاخیر داشته باشید و به ازای هر تمرین بیشتر از ۴ روز تاخیر، مشمول کسری نمره می باشد، بطوری که بعد از روز ۴ام، به ازای هر روز اضافی، ۲۰ درصد از نمره تمرین را از دست خواهید 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، ۱۰۰ نمره منفی برای طرفین در نظر گرفته می شود، لذا می توانید صرفاً از یکدیگر مشورت بگیرید یا به صورت ایمیل به آدرس زیر بپرسید.

alishojaei7697@gmail.com

سوال اول)

در این سوال قصد داریم جایگشتی تصادفی از سطرها که در بخش ۳.۳.۵ از کتاب mining massive datasets عنوان شده است را شبیه سازی کنیم.

الف) فرض کنید یک ستون n تایی با m مقدار ۱ و درنتیجه n-m مقدار صفر داریم و به صورت تصادفی k سطر در هنگام محاسبه minhash value انتخاب می شوند. در چه صورت نتیجه minhash برابر با "don't know" می شود؟ حال ثابت کنید احتمال گرفتن "don't know" به عنوان minhash value برای این ستون حداکثر $\left(\frac{n-k}{n}\right)^m$ بدست می آید.

ب) در چه صورتی الگوریتم انتخابی به صورت ناموفق عمل می کند؟ فرض کنید که می خواهیم احتمال "don't know" برابر با حداکثر e^{-10} شود؛ در نظر بگیرید n و m مقادیر بسیار بزرگی دارند (اما n بسیار بزرگتر از m یا k است)، به صورت تقریبی کوچکترین مقدار k را به نوعی بیابید که این اتفاق رخ دهد.

(راهنمایی: برای مقادیر بسیار بزرگ x خواهیم داشت: $\left(1-\frac{1}{x}\right)^x \approx \frac{1}{e}$)

ج) در هنگام minihashing می توان انتظار داشت که ما میتوانیم Jaccard similarity را بدون استفاده از تمامی جایگشت های سطرها تخمین بزنیم؛ برای مثال می توانیم تنها جایشگت های تناوبی[1] را اجازه دهیم. در ابتدا یک مثال برای این حالت ارائه دهید؛ آیا این جایگشت ها برای تخمین Jaccard similarity مناسب هستند؟ (راهنمایی: پاسخ خود را با ارائه یک مثال از یک ماتریس دو ستونه که هر کدام از ستون های آن مربوط به دو مجموعه ی S1 و S2 هستند بیان کنید، Jaccard similarity مجموعه های S1 و S2 را محاسبه کنید و در پایان احتمال اینکه یک جایگشت تناوبی تصادفی برای دو مجموعه ی S1 و S2، minhash value یکسان بدست آورد را بدست آورید).

سوال دوم)

در این مسئله قصد داریم کاربرد LSH را برای پیدا کردن approximate nearest neighbors بررسی کنیم. فرض کنید یک دیتاست A با n نقطه در یک فضای متریک با معیار فاصله d(., .) داریم. c را یک عدد ثابت بزرگتر از یک درنظر بگیرید؛ بنابراین مسئله (c, λ)-Approximate Near Neighbor (ANN) به صورت زیر تعریف می شود:

فرض کنید یک query point بنام z داریم؛ در نظر بگیرید یک نقطه x در این دیتاست با شرط $d(x,z) \leq \lambda$ نقطه دیگری بنام x' با شرط $d(x',z) \leq c\lambda$ را برمیگرداند (این نقطه (c, λ)-ANN نامیده می شود). پارامتر c در این مسئله maximum approximation factor را نشان می دهد. حال یک خانواده LSH بنام H از توابع هش را در نظر بگیرید به گونه ای که برای معیار فاصله d(., .)، $(\lambda, c\lambda, p_1, p_2)$-sensitive باشد. حال فرض کنید:

$$G = H^k = \{g = (h_1, \dots, h_k) | \, h_i \in H, \forall\, 1 \leq i \leq k\} \; where \; k = \log_{1/p_2}(n)$$

حال پروسه زیر را در نظر بگیرید:

[1] Cyclic permutation

۱. $L = n^p$ عضو تصادفی $g_1, \dots, g_L$ از G را به گونه ای انتخاب کنید که $\rho = \frac{\log(1/p_1)}{\log(1/p_2)}$
۲. تمامی نقاط داده به همراه query point را با استفاده از تمامی $(1 \le i \le L)$ g_i هش کنید.
۳. حداکثر 3L نقطه که به صورت تصادفی یکنواخت انتخاب شده اند را از مجموعه ی L باکتی که query point به آن هش می شود را بازیابی کنید.
۴. از میان تمام نقاطی که در مرحله ۳ انتخاب شده اند، یکی را که از بین همه به query point نزدیک تر است را به عنوان یک (c, λ)-Approximate Near Neighbor (ANN) گزارش کنید.

هدف از اولین بخش از این مسئله این است که نشان دهید این پروسه در نهایت به یک پاسخ درست با یک احتمال ثابت منجر می شود.

الف) فرض کنید $(1 \le j \le L)$ $W_j = \{x \in A | g_j(x) = g_j(z)\}$ مجموعه ای از نقاط داده x است که همانند query point z به یک مقدار یکسان با استفاده از تابع هش g_j مپ می شوند. ثابت کنید: (راهنمایی: نامساوی مارکوف)

$$\Pr\left[\sum_{j=1}^{L} |T \cap W_j| \ge 3L\right] \le \frac{1}{3} \qquad where\ T = \{x \in A | d(x,z) > c\lambda\}$$

ب) فرض کنید $x^* \in A$ نقطه باشد به گونه ای که $d(x^*, z) \le \lambda$ ؛ ثابت کنید:

$$\Pr\left[\forall\ 1 \le j \le L, \quad g(x^*)_j \ne g(z)_j\right] < \frac{1}{e}$$

ج) نتیجه گیری کنید با احتمالی بزرگتر از یک مقدار ثابت، نقطه گزارش داده شده یک (c, λ)- ANN واقعی است.

سوال سوم)

در این بخش قصد داریم خودروهایی را شناسایی کنیم که مسیر مشابهی را پیموده اند. در سایت cw درس، لینک درایو مجازی با عنوان اطلاعات پروژه و تمرین درس موجود است. نمونه کد برای خوانش و پردازش اولیه برروی دیتا نیز در این درایو موجود است. می توانید در ابتدا برروی SampleData.csv کار کنید سپس برروی دیتای اصلی سوال.

الف) مسیر عبوری هر خودرو را به تفکیک روز در یک rdd مشخص کنید. به عنوان مثال key = (Plate, Date) و value = [Device Code List].

ب) یک مسیر فرضی به صورت [Device Code List] را در نظر بگیرید و با محاسبه شباهت کسینوسی بین این مسیر و مسیر خودروها در rdd بخش الف، ۵ مسیر و خودرو با بیشترین شباهت را گزارش کنید.

پ) در این بخش با استفاده از LSH بخش ب را حل کنید. چند hyperplane در نظر بگیرید و محاسبات مربوط به نحوه استفاده از آن ها را انجام دهید (and or). سپس مسیر های مشابه با مسیر فرضی را گزارش کنید. در صورت افزایش تعداد hyperplaneها دقت به چه صورت افزایش می یابد؟

موفق باشید