مبانی ریاضی علوم‌داده

گزارش پروژه اول

پوریا صامتی

۴۰۴۰۹۷۱۴

پائیز ۱۴۰۴

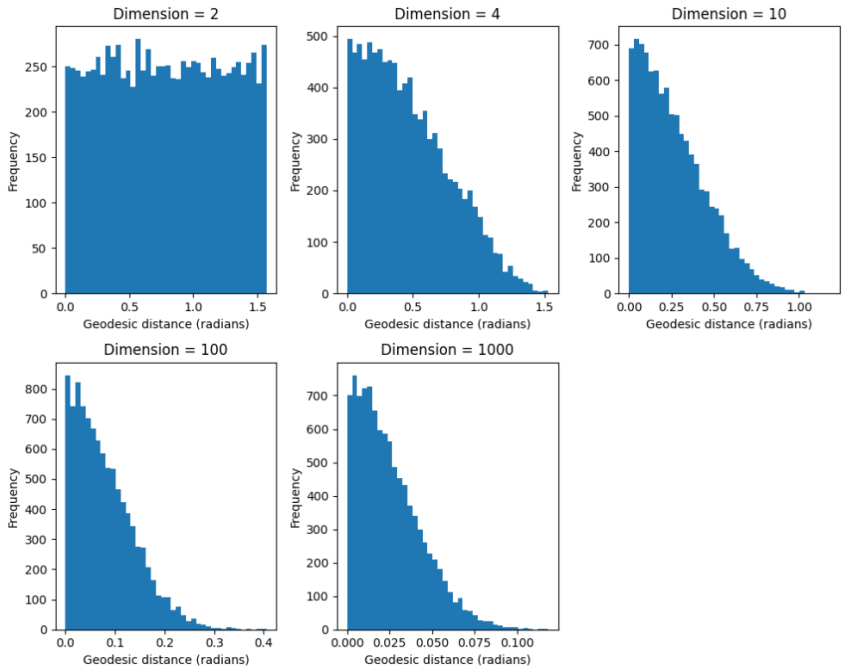
فهرست مطالب

پرسش اول

در این سؤال هدف آن است که در یک کره واحد، مجموعه‌ای از نقاط تصادفی را در ابعاد مختلف تولید کنیم. سپس رفتار این نقاط را در فضاهای با بُعدهای گوناگون بررسی کنیم. برای هر نقطه، فاصله ژئودزیک آن تا استوای کره واحد و همچنین فاصله‌اش تا یک بردار نرمال تصادفی در صفحه مورد مطالعه قرار می‌گیرد. در ادامه نیز زاویه بردارهای متناظر با این نقاط تحلیل می‌شود.

بخش a)

در این بخش، ۱۰۰۰ نقطه تصادفی روی کره‌های واحد در ابعاد ۲، ۴، ۱۰، ۱۰۰ و ۱۰۰۰ تولید کردیم. پس از تولید نقاط، فاصله ژئودزیک هر نقطه تا استوای کره مربوطه اندازه‌گیری شد. هیستوگرام این فواصل برای هر بُعد رسم شده است و رفتار توزیع نقاط نسبت به استوا را نشان می‌دهد.

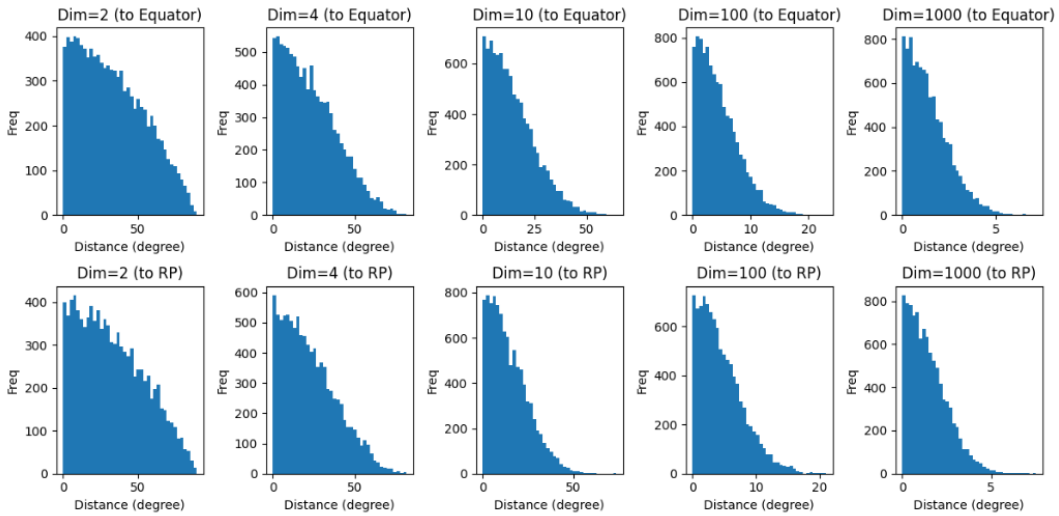


همانطور که مشخص است در ابعاد پایین، میزان این فاصله بسیار پراکنده است و از استوا دور هستند. اما با افزایش بعد، این فواصل همگرا شده و از شدت پراکندگی آنها کاسته می‌شود و به شکل متناظر شاهد نزدیک شدن این فواصل به صفر هستیم. نتایج این آزمایش نشان می‌دهد که با افزایش بُعد کره واحد، پدیده «تمرکز اندازه» به‌صورت چشم‌گیری ظاهر می‌شود؛ به‌طوری‌که در ابعاد پایین مانند ۲ و ۴، نقاط روی کره نسبتاً پراکنده‌اند و فاصله ژئودزیک آن‌ها تا استوا گستره

وسیعی دارد، اما با افزایش بعد به ۱۰، ۱۰۰ و به‌ویژه ۱۰۰۰، این فاصله‌ها به‌شدت کوچک می‌شوند و تقریباً تمام نقاط در ناحیه‌ای بسیار باریک پیرامون استوا جمع می‌شوند. بنابراین هرچه بُعد بیشتر می‌شود، **حجم مؤثر کره به‌طور فزاینده‌ای پیرامون استوا متمرکز شده و توزیع نقاط به شکل یک قله باریک نزدیک صفر** در می‌آید. این مشاهده نیز در **راستای مشاهده دوم نتایج را به ما نشان می‌دهد که دلالت بر درستی این آزمایش** دارد.

بخش b)

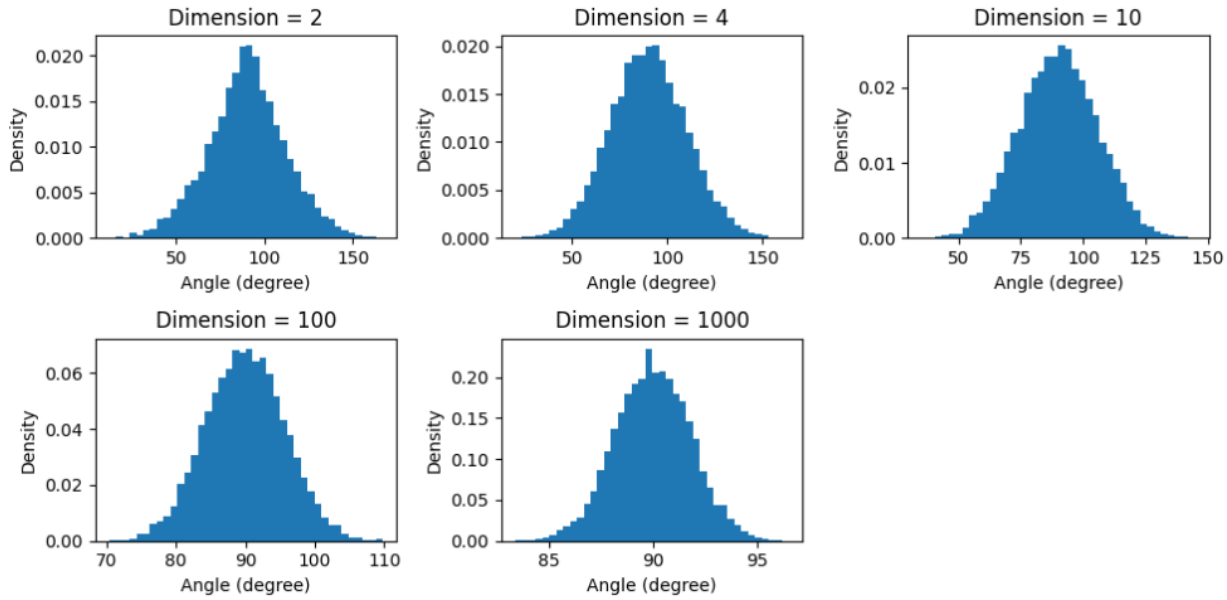
در این مرحله با استفاده از یک بردار تصادفی، صفحه‌ای را در نظر گرفتیم که از مبدا می‌گذرد. حال فاصله نقاط تولید شده را تا این صفحه محاسبه کرده‌ایم. هیستوگرام فواصل به شرح زیر هستند:



در این مرحله نیز با محاسبه فاصله نقاط روی کره‌های واحد با ابعاد مختلف نسبت به یک صفحه عبوری از مبدأ (تعریف‌شده توسط یک بردار نرمال تصادفی)، همان الگوی آشنای تمرکز اندازه در ابعاد بالا مشاهده شد: در ابعاد کم، فاصله نقاط از صفحه توزیعی گسترده دارد و نقاط در دو طرف صفحه پراکنده‌اند، اما با افزایش بعد به ۱۰، ۱۰۰ و ۱۰۰۰ تقریباً تمام نقاط در فاصله‌ای بسیار کوچک از صفحه قرار می‌گیرند (تا حداکثر ۱۰ درجه) و هیستوگرام‌ها شکلی تیز و متمرکز نزدیک صفر پیدا می‌کنند. این رفتار کاملاً مشابه بخش قبل است که در آن فاصله نقاط تا استوا بررسی شد؛ چراکه **چه استوا و چه یک صفحه تصادفی، هر دو زیرساختارهایی از کره‌اند که بعدشان یک واحد کمتر از بعد فضاست، و تمرکز اندازه باعث می‌شود تقریباً همه نقاط در همسایگی چنین زیرساختارهایی جمع** شوند. بنابراین، نتایج هر دو بخش بیانگر این واقعیت هستند که در فضاهای بعد بالا، نقاط روی کره به‌شدت به ساختارهای با یک بعد کمتر (استوا یا هر صفحه تصادفی) نزدیک می‌شوند.

بخش c)

در این بخش، دوباره ۱۰۰۰ نقطه تازه روی کره واحد در ابعاد ۲، ۴، ۱۰، ۱۰۰ و ۱۰۰۰ تولید کردیم و این بار زاویه بین هر دو نقطه را محاسبه و تحلیل کردیم.



نتایج نشان می‌دهند که در ابعاد پایین (مانند ۲ و ۴) توزیع زاویه‌ها گسترده است و مقادیر متنوعی از زوایا مشاهده می‌شود؛ یعنی نقاط می‌توانند با جهت‌گیری‌های مختلف نسبت به یکدیگر قرار گیرند.

اما هرچه بعد افزایش می‌یابد، توزیع زاویه‌ها به‌سرعت حول مقدار ۹۰ درجه متمرکز می‌شود. در ابعاد ۱۰۰ و ۱۰۰۰ تقریباً تمام زاویه‌ها بسیار نزدیک به ۹۰ درجه هستند (با حدود ۵ درجه اختلاف) و هیستوگرام‌ها قله‌ای تیز و باریک در همین حوالی تشکیل می‌دهند. این نتیجه بیانگر یکی از پدیده‌های کلیدی فضاهای بُعد بالا است که **تقریباً تمام بردارهای واحد در بُعدهای بزرگ نسبت به یکدیگر تقریباً متعامد می‌شوند**. این پدیده ناشی از تمرکز اندازه و افزایش آزادی جهت‌گیری در ابعاد بالا است و نشان می‌دهد.

این مشاهده نیز دلالتی بر **تعامد تقریبی نقاط در کره واحد در ابعاد بالا** می‌باشد.

بخش d)

حال می‌خواهیم میانگین فواصل نقاط تا استوا و میانگین زاویه بین نقاط را بررسی کنیم.

- **میانگین فاصله تا استوا**: از نتایج شبیه‌سازی مشخص است که میانگین فاصله ژئودزیک تا استوا با افزایش بُعد کاهش می‌یابد. دلیل آن این است که در ابعاد بالا بیشتر جرم اندازه روی نقاطی متمرکز می‌شود که مؤلفه اول آن‌ها کوچک است و به «استوا» نزدیک‌ترند .

	dimension	point	Mean Distance To Equator
0	2	10000	0.776446
1	4	10000	0.468695
2	10	10000	0.267582
3	100	10000	0.081592
4	1000	10000	0.025123

- **میانگین زاویه بین نقاط**: در بخش زوایا دیدیم که با افزایش ابعاد، میانگین زاویه بین دو نقطه تصادفی روی کره واحد به‌سرعت به ۹۰ درجه نزدیک می‌شود.

	dimension	point	Mean angle
0	2	10000	90.299995
1	4	10000	89.456626
2	10	10000	89.679869
3	100	10000	89.979764
4	1000	10000	90.009036

- **رفتار d × E[distance]** : حاصل‌ضرب این ثابت نمی‌ماند؛ بلکه تقریباً مانند $\sqrt{d}$ افزایش می‌یابد. اینکه این مقدار رشد می‌کند و ثابت نمی‌ماند، ناشی از پدیده تمرکز اندازه در کره‌های با بُعد بالا است؛ یعنی اینکه مؤلفه‌های یک نقطه یکنواخت روی کره به شدت کوچک در مقیاس $1/\sqrt{d}$ می‌شوند. پس در اصل همانند این است که نرم بردارها به یک مقدار ثابت میل می‌کند در صورتیکه بعد در این رابطه افزایش یافته و نتیجه ضرب نیز صعودی خواهد شد.

	dimension	Mean Distance to Equator	d × Mean Distance
0	2	0.789153	1.578306
1	4	0.464556	1.858226
2	10	0.267452	2.674518
3	100	0.080448	8.044829
4	1000	0.025238	25.238175

پرسش دوم

حال در این بخش قصد داریم تا در کره و مکعب واحد نقاطی را تولید کنیم. سپس با استفاده از روش منت کارلو، حجم کره را تخمین بزنیم و دیگر خواص این فضاها را بررسی کنیم.

بخش a)

در این بخش تعداد ۱۰۰۰ نقطه را در ابعاد ۲، ۳، ۵، ۱۰ و ۲۰ در مکعب واحد در بازه 0.5- تا 0.5+ تولید کرده‌ایم. برای اینکار به آسانی از توزیع یکنواخت در بازه 5- تا 0.5+ بهره می‌بریم.

بخش b)

حال می‌خواهیم با استفاده از تقریب مونت‌کارلو، حجم کره به شعاع ۰.۵ را تقریب بزنیم. پس از بدست آوردن حجم کره در ابعاد مختلف، حال به بررسی نتایج این آزمایش می‌پردازیم.

	Dim	Points	Sphere Volume (MC)	Mean Distance	Mean Nearest Distance
0	2	1000	0.788	0.533893	0.015839
1	3	1000	0.538	0.666605	0.057862
2	5	1000	0.160	0.882778	0.178935
3	10	1000	0.004	1.264372	0.499553
4	20	1000	0.000	1.815232	1.047879

- **بررسی حجم**: با افزایش بعد، حجم کره به 0 میل می‌یابد. این پدیده یکی از **نمودهای مهم نفرین ابعاد** است.

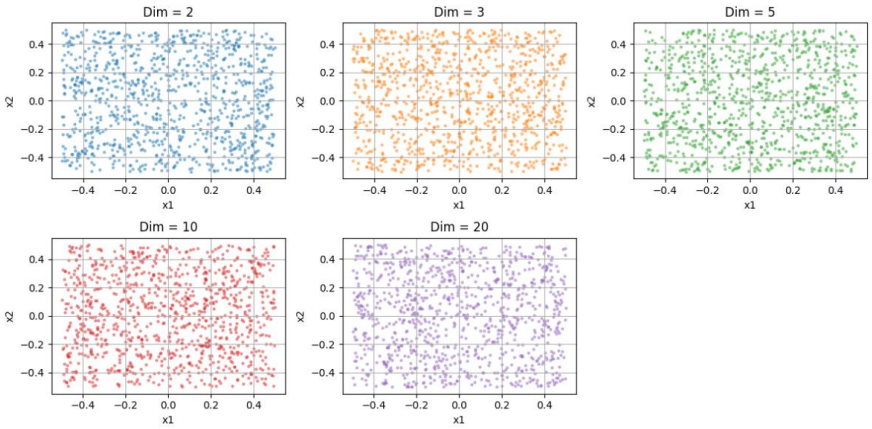
- **بررسی میانگین فاصله نقاط**: با افزایش بُعد، نقاط در فضا بسیار پراکنده‌تر می‌شوند و بردارها نسبت به یکدیگر عمودتر می‌گردند؛ به همین دلیل میانگین فاصله میان تمام نقاط افزایش می‌یابد و تقریباً با $\sqrt{d}$رشد می‌کند.

- **بررسی میانگین فاصله تا نزدیک‌ترین همسایه**: از سوی دیگر، فاصله تا نزدیک‌ترین همسایه نیز افزایش پیدا می‌کند زیرا فضا در ابعاد بالا بسیار خلوت‌تر می‌شود؛ با این حال، نرخ رشد فاصله نزدیک‌ترین همسایه کندتر از افزایش فاصله میانگین نقاط است، چون هنوز تعدادی نقاط نسبتاً نزدیک وجود دارند. در نتیجه، در بعدهای بالا حتی نزدیک‌ترین نقطه نیز چندان نزدیک نیست و کل فضا به‌شدت «کشیده» و «خالی» به نظر می‌رسد.

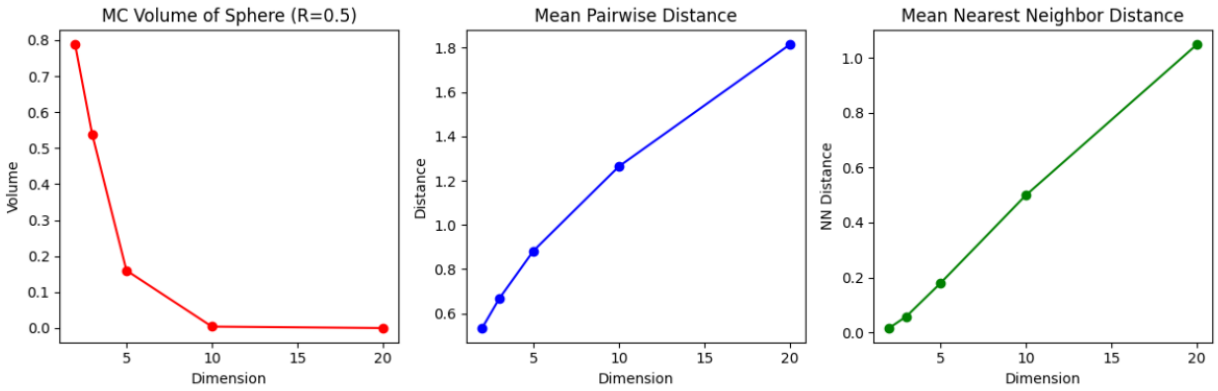
بخش c)

در این بخش ابتدا نمودارهای مطلوب را نمایش می‌دهیم.

- **پراکنده‌سازی دو بعد اول نقاط**: در نمودار پراکنده‌سازی دو مؤلفه اول، هیچ تفاوت چشمگیری بین ابعاد مختلف مشاهده نمی‌شود و نقاط تقریباً یکسان به نظر می‌رسند؛ زیرا تفاوت‌های اساسی ناشی از افزایش بعد در ساختار هندسی کل فضاست، نه در توزیع دوبعدی مؤلفه‌های اول.



- **نمودارهای حجم کره در ابعاد مختلف، میانگین فاصله بین نقاط، میانگین فاصله به نزدیک‌ترین همسایه**: با افزایش بعد، حجم کره واحد به‌سرعت کاهش می‌یابد و در ابعاد بالا تقریباً به صفر نزدیک می‌شود، در حالی که میانگین فاصله بین تمام نقاط به‌طور یکنواخت افزایش یافته و نقاط نسبت به هم بسیار دورتر می‌شوند. میانگین فاصله تا نزدیک‌ترین همسایه نیز افزایش می‌یابد، اما با سرعتی کمتر از میانگین فاصله زوجی، به این معنی که حتی نزدیک‌ترین نقطه نیز در فضای ابعاد بالا چندان نزدیک نیست. این رفتارها با هم نشان‌دهنده اثر «**نفرین ابعاد**» هستند: فضای ابعاد بالا بسیار خالی می‌شود و فواصل بین نقاط همگرا و بزرگ می‌شوند، در حالی که حجم قابل استفاده کره در مقایسه با مکعب محیطی ناچیز است.



با افزایش بعد، حجم کره واحد به‌سرعت کاهش می‌یابد و در مقایسه با حجم مکعب تقریباً به صفر میل می‌کند، بنابراین کره در فضای ابعاد بالا بخش ناچیزی از فضا را اشغال می‌کند.

نقاط نیز از مرکز فاصله می‌گیرند و بیشتر **در یک «پوسته» دور از مرکز قرار می‌گیرند** (مشاهده اول)، به‌طوری که فاصله متوسط از مرکز تقریباً با ابعاد رشد می‌کند. میانگین فاصله بین نقاط و میانگین فاصله تا نزدیک‌ترین همسایه هر دو افزایش می‌یابند، اما تفاوت نسبی آن‌ها کاهش می‌یابد و فاصله‌ها به‌طور کلی به یکدیگر نزدیک می‌شوند؛ این پدیده همان «تمرکز فاصله‌ها» است که باعث کاهش کارایی الگوریتم‌های مبتنی بر فاصله همانند KMeans یا K-Nearest Neighbor می‌شود. حتی اگر بازه مختصات کوچک‌تر شود، رفتار مشابه است و فواصل با ابعاد رشد می‌کنند مگر آنکه مقیاس هر بعد متناسب با افزایش ابعاد کاهش داده شود.

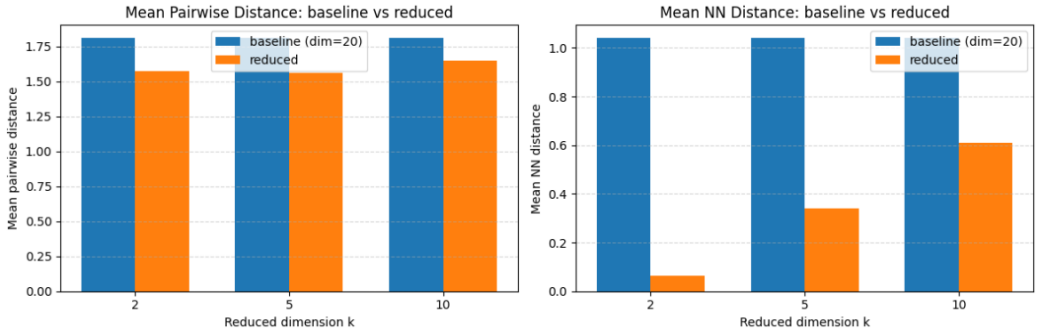
در نتیجه، در ابعاد بالا بیشتر مشاهدات با شهود دوبعدی فاصله می‌گیرند و بدون کاهش ابعاد یا روش‌های تصحیح فاصله، **عملکرد الگوریتم‌های مبتنی بر فاصله ضعیف** خواهد بود.

بخش d)

حال که متوجه وجود نفرین ابعاد در داده‌های مراحل قبل شدیم، در این بخش از یک الگوریتم کاهش بعد به نام Random Projection استفاده می‌کنیم. حال می‌خواهیم وضعیت داده‌های خود را قبل و بعد از اعمال این الگوریتم بررسی کنیم. ما داده ۲۰ بعدی خود را به ۲، ۵ و ۱۰ با این روش کاهش می‌دهیم.

	New Dim	Mean Distance	Mean NN Distance	Mean Distance Change (percent)	Mean NN Change (percent)
0	2	1.573837	0.064896	-13.157341	-93.770089
1	5	1.564221	0.340362	-13.687924	-67.325669
2	10	1.651489	0.609359	-8.872572	-41.502235

در این جدول، میانگین فواصل نقاط و میانگین فاصله تا نزدیک‌ترین نقطه مشخص شده اند. همچنین درصد اختلاف آن با حالت قبل از کاهش بعد مقایسه شده است. نمودارهای زیر با توجه به این جدول تولید شده اند. این نتایج را تحلیل می‌کنیم.

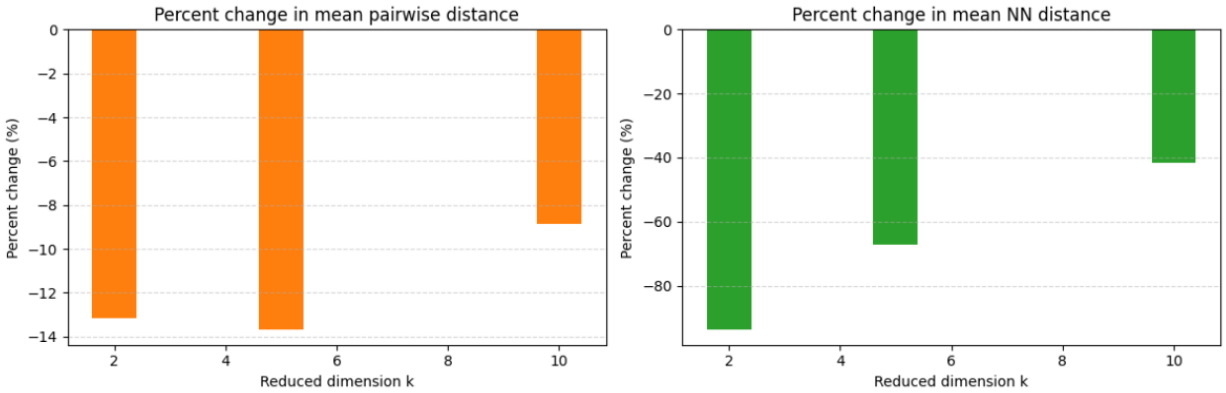


- **اثر کاهش بعد بر میانگین فاصله‌ها**: داده اصلی ۲۰ بعدی شامل نقاطی با فاصله‌های بزرگ و پراکنده است. کاهش بعد به ۱۰، ۵ یا ۲ بعد باعث می‌شود که نقاط بر روی فضای کم‌بعدی فشرده شوند. با استفاده از الگوریتم پروجکشن تصادفی فواصل تقریباً حفظ می‌شوند، اما نسبت دقیق فاصله‌ها کمی تغییر می‌کند. میانگین فاصله بین نقاط بعد از کاهش ابعاد کمی کوچک‌تر یا نزدیک‌تر می‌شود، زیرا برخی اطلاعات بعدی حذف شده است.

- **اثر کاهش بعد بر میانگین فاصله تا نزدیک‌ترین همسایه**: میانگین فاصله تا نزدیک‌ترین همسایه کاهش شدیدی داشته است. در نتیجه **اثر نفرین ابعاد تا حدی کاهش یافته است** .

- **مقایسه با حالت اصلی**: بدون کاهش بعد، فواصل بین نقاط و فاصله نزدیک‌ترین همسایه بزرگ‌تر و پراکنده‌تر است. با کاهش بعد، فواصل کمی جمع‌تر می‌شوند و تغییر نسبی بین میانگین فاصله زوجی و میانگین فاصله به نزدیک‌ترین همسایه کمتر می‌شود. به این ترتیب، کاهش بعد باعث می‌شود اثر نفرین ابعاد تا حدی کاهش یابد و الگوریتم‌های مبتنی بر فاصله در فضای کم‌بعدی بهتر عمل کنند.

- **نتیجه‌گیری تحلیلی**: کاهش بعد، اگر با روش مناسب مانند پروجکشن تصادفی با انجام شود، می‌تواند فاصله‌ها را تقریباً حفظ کند. این کار باعث می‌شود داده‌ها کمتر پراکنده شوند و فضای کارایی الگوریتم‌های فاصله‌محور بهبود یابد. در نمودارهای میانگین فاصله قبل و بعد از کاهش بعد، معمولاً مشاهده می‌شود که میانگین فاصله‌ها کمی کاهش یافته اما الگوی نسبی فاصله‌ها حفظ شده است. زیرا پروجکشن تصادفی فاصله‌ها را تقریباً حفظ می‌کند ولی فضای عملیاتی کوچکتر باعث می‌شود الگوریتم‌های مبتنی بر فاصله بتوانند روابط بین نقاط را بهتر مدیریت کنند و اثرات ناشی از خالی بودن شدید فضای ابعاد بالا کاهش یابد. در نمودار زیر نیز مقدار کاهش فاصله به درصد نیز مشخص شده است که نشان‌دهنده مدیریت فاصله در ابعاد کمتر می‌باشد.



پرسش سوم

در این پرسش از مجموعه داده متنی 20Newsgroups استفاده می‌کنیم. ابتدا داده‌های متنی را با استفاده از روش TF-IDF به بردارهای عددی تبدیل می‌کنیم. طبق مشاهده اولیه، پس از بردارسازی داده‌های متنی، به مجموعه‌داده‌ای می‌رسیم که نماینده یک ماتریس خلوت می‌باشد. اما بحث اصلی این است که این مجموعه‌داده دارای ابعاد بسیار بزرگی نسبت به تعداد نمونه‌های موجود در مجموعه‌داده می‌باشد. پس احتمال بروز پدیده نفرین ابعاد بسیار بالا می‌باشد. در ادامه این موضوع را بررسی می‌کنیم.

بخش a)

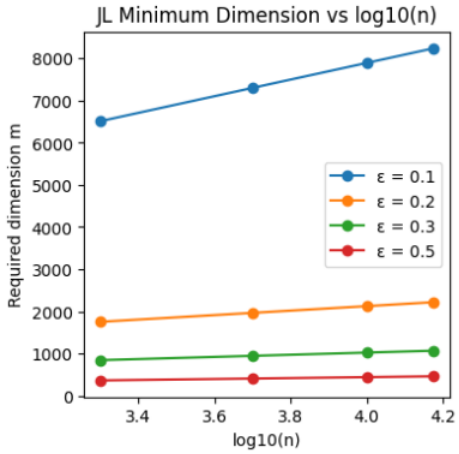
در این بخش میخواهیم درباره لم **Johnson–Lindenstrauss (JL)** صحبت کنیم که بیان می‌کند که می‌توان مجموعه‌ای از n نقطه در فضای با ابعاد بالا را به یک فضای کم‌بعدی با بعد $m = O(\frac{\log n}{\epsilon^2})$ پروجکت کرد به طوری که فاصله‌های اقلیدسی بین نقاط تقریباً حفظ شوند، با خطای نسبی حداکثر ϵ. به بیان ساده، این لم امکان کاهش ابعاد داده‌های بزرگ را فراهم می‌کند بدون آنکه فواصل بین نقاط تغییر زیادی کنند و پایه تکنیک‌های کاهش بعد تصادفی مانند Gaussian یا Sparse Random Projection است. در ابتدا درباره پارامترهای موجود در این بخش صحبت می‌کنیم.

- **بعد کمینه (m)**: حداقل بعدی است که برای نگه داشتن فواصل بین بردارها با دقت مشخص ϵ لازم است. به عبارت دیگر، بعدی که پس از کاهش بعد، فواصل بین بردارها تقریباً بدون تغییر باقی بمانند.
- **خطای مجاز در فاصله‌ها (ϵ)**: خطای قابل قبول در حفظ فاصله‌ها پس از کاهش بعد است. اگر ϵ کوچک باشد، فواصل دقیق‌تر حفظ می‌شوند و نیاز به بعد بزرگ‌تری داریم؛ اگر ϵبزرگ‌تر باشد، خطای مجاز بیشتر است و بعد کمتری کافی است.

جدول زیر حاصل از اعمال این لم بر داده‌های ما به ازای پارامترهای مذکور می‌باشد:

÷	0.1 ÷	0.2 ÷	0.3 ÷	0.5 ÷
2000	6515	1754	844	364
5000	7300	1965	946	408
10000	7894	2125	1023	442
15000	8242	2219	1068	461

اگر این جدول را در نمودار زیر نمایش دهیم، رابطه لگاریتمی که در توضیحات معرفی شد، خود را نمایش می‌دهد.



بخش b)

حال با توجه به مقادیر m الگوریتم Gaussian Random Project بر مجموعه‌داده اعمال می‌شود. سپس نتیجه نهایی را با توجه به ϵ مختلف بررسی می‌کنیم.

بخش c)

نتایج بخش b با نمودارهایی از نسبت فاصله‌ها به بعد m و ϵ نمایش داده شده‌اند که شامل میانگین، میانه و صدک‌های ۵ و ۹۵ نسبت‌ها هستند. این نمودارها نشان می‌دهند که کاهش بعد تصادفی با Gaussian Random Projection فواصل بین بردارها را تقریباً حفظ می‌کند و اثر نفرین ابعاد کاهش می‌یابد. هرچه بعد کاهش یافته بزرگ‌تر باشد، درصد زوج‌های با نسبت فاصله دقیق‌تر افزایش می‌یابد. این فرآیند باعث می‌شود الگوریتم‌های مبتنی بر فاصله زیرا فواصل نسبی تقریباً ثابت باقی می‌مانند و خطای آن‌ها محدود است؛ به‌عبارت دیگر، هم کاهش بعد و هم حفظ فاصله‌ها به طور همزمان امکان‌پذیر است. نتایج در جدول زیر نمایش داده شده اند.

	m	epsilon	mean_ratio	median_ratio	p05_ratio	p95_ratio	percent_inside_1±eps
0	50	0.02	0.998367	0.996813	0.836291	1.166965	15.712
1	50	0.05	0.998367	0.996813	0.836291	1.166965	38.106
2	50	0.10	0.998367	0.996813	0.836291	1.166965	67.880
3	100	0.02	0.999598	0.999219	0.885189	1.116560	22.380
4	100	0.05	0.999598	0.999219	0.885189	1.116560	52.494
5	100	0.10	0.999598	0.999219	0.885189	1.116560	84.490
6	200	0.02	0.997614	0.997359	0.915999	1.079555	31.184
7	200	0.05	0.997614	0.997359	0.915999	1.079555	68.520
8	200	0.10	0.997614	0.997359	0.915999	1.079555	95.522
9	500	0.02	0.998565	0.998336	0.947276	1.050514	47.578
10	500	0.05	0.998565	0.998336	0.947276	1.050514	88.770
11	500	0.10	0.998565	0.998336	0.947276	1.050514	99.856
12	1000	0.02	0.999115	0.999096	0.962974	1.035757	63.240
13	1000	0.05	0.999115	0.999096	0.962974	1.035757	97.598
14	1000	0.10	0.999115	0.999096	0.962974	1.035757	99.998
15	2000	0.02	0.999320	0.999324	0.973362	1.025310	79.488
16	2000	0.05	0.999320	0.999324	0.973362	1.025310	99.890
17	2000	0.10	0.999320	0.999324	0.973362	1.025310	100.000
18	5000	0.02	0.999546	0.999481	0.983174	1.016011	95.558
19	5000	0.05	0.999546	0.999481	0.983174	1.016011	100.000
20	5000	0.1	0.999546	0.999481	0.983174	1.016011	100.0