دانشگاه علم و صنعت ایران - دانشکده مهندسی کامپیوتر

موضوع: پروژه جبر خطی

استاد: دکتر امید نیکان

تهیه کنندگان:

ایلیا صادق‌احمدی

محمد منتظری

شاهین شاپورزاده

سیدعلیرضا حسینی

سؤال ۱

از روش‌های ارزش‌گذاری کلمات در متون، روش‌های TF و IDF را توضیح دهید. چرا مطالعهٔ مستقل این دو ارزش ممکن است گمراه‌کننده باشد؟

پاسخ

Term Frequency (TF)

میزان تکرار یک کلمه در یک متن مشخص را اندازه می‌گیرد. رایج‌ترین تعریف آن به‌صورت نرمالیزه‌شده است:

$$TF(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}}$$

این نرمال‌سازی باعث می‌شود متن‌های بلندتر به‌طور مصنوعی امتیاز بالاتری نگیرند.

Inverse Document Frequency (IDF)

IDF میزان «کمیابی» یک کلمه در کل مجموعه متون را اندازه می‌گیرد:

$$IDF(t) = \log\left(\frac{N}{1 + df_t}\right)$$

که در آن:

N تعداد کل متون است.

عدد ۱ در مخرج برای جلوگیری از تقسیم بر صفر اضافه می‌شود. هرچه کلمه‌ای در متون کمتری ظاهر شود، مقدار IDF آن بزرگ‌تر است و در نتیجه آن کلمه قدرت تفکیک‌کنندگی بیشتری دارد.

ترکیب TF و IDF

در عمل معمولاً از حاصل‌ضرب این دو معیار استفاده می‌شود:

$$TFIDF(t,d) = TF(t,d) \times IDF(t)$$

چرا بررسی مستقل TF و IDF می‌تواند گمراه‌کننده باشد؟

TF به‌تنهایی نمی‌تواند بین کلمات پرتکرارِ مهم و کلمات پرتکرارِ بی‌اهمیت تمایز قائل شود.
مثلاً کلماتی مانند «is»، «the» و «and» در بسیاری از متون تکرار زیادی دارند، اما لزوماً اطلاعات موضوعی مهمی منتقل نمی‌کنند.

IDF به‌تنهایی فقط نادر بودن یک کلمه را در کل مجموعه متون می‌سنجد و توجهی به تعداد تکرار آن در متن مورد نظر ندارد.
بنابراین ممکن است کلمه‌ای در کل مجموعه بسیار کمیاب باشد، اما اصلاً در متن مورد نظر ظاهر نشده باشد؛ در این صورت نباید برای آن متن امتیاز مهمی بگیرد. همچنین کلمات ناشی از غلط املایی یا نویز هم ممکن است IDF بالایی داشته باشند، بدون اینکه واقعاً مفید باشند.

بنابراین فقط ترکیب TF و IDF معنادار است؛ یعنی کلمه‌ای برای یک متن مهم تلقی می‌شود که:

در همان متن زیاد تکرار شده باشد (TF بالا)، و

در کل مجموعه متون نسبتاً کمیاب باشد (IDF بالا).

در نتیجه، TF-IDF کلماتی را برجسته می‌کند که هم در متن مهم‌اند و هم آن متن را از سایر متون متمایز می‌کنند.

سؤال ۲

یکی از روش‌های تعیین آستانه‌ی مقادیر تکین برای کاهش مرتبه در Truncated SVD، رویکرد نقطه زانویی است. این رویکرد را شرح دهید.

پاسخ

پس از انجام SVD روی ماتریس داده، مقادیر تکین را به‌صورت زیر داریم:

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r \geq 0$$

این مقادیر معمولاً در ابتدا با سرعت زیادی کاهش پیدا می‌کنند و سپس روند افت آن‌ها آهسته‌تر و تقریباً تخت می‌شود.

ایده‌ی روش نقطه زانویی (Elbow Point)

رسم می‌شود. در این نمودار معمولاً یک نقطه‌ی iبر حسب شماره مؤلفه σ_iدر این روش، نمودار مقادیر تکین شکست یا «زانویی» دیده می‌شود؛ یعنی نقطه‌ای که:

قبل از آن مقادیر تکین با شیب زیاد افت می‌کنند و مؤلفه‌ها بخش عمده‌ای از ساختار اصلی داده را حمل می‌کنند.

بعد از آن مقادیر تکین با شیب کم افت می‌کنند و سهم آن‌ها بیشتر به نویز مربوط می‌شود تا اطلاعات مهم.

بنابراین نقطه‌ی زانو را می‌توان به‌عنوان آستانه‌ی مناسبی برای انتخاب رتبه‌ی Cدر Truncated SVD در نظر گرفت.

تفسیر شهودی

اگر رتبه را تا نقطه‌ی زانو نگه داریم، در واقع:

بیشترین بخش سیگنال و ساختار مهم داده را حفظ کرده‌ایم،

و مؤلفه‌های کم‌اهمیت‌تر یا نویزی را حذف کرده‌ایم.

به این ترتیب، با تعداد نسبتاً کمی مؤلفه، بخش بزرگی از اطلاعات ماتریس حفظ می‌شود.

روش خودکار متداول برای یافتن نقطه زانو

یکی از روش‌های رایج این است که:

یک خط مستقیم بین اولین و آخرین نقطه‌ی نمودار، رسم می‌کنیم.

فاصله‌ی عمود هر نقطه از نمودار تا این خط را محاسبه می‌کنیم.

نقطه‌ای که بیشترین فاصله را از این خط داشته باشد، به‌عنوان نقطه‌ی زانو انتخاب می‌شود.

این نقطه معمولاً همان جایی است که تغییر نرخ افت مقادیر تکین بیشترین مقدار را دارد.

سؤال ۳

نحوه‌ی محاسبه‌ی خطای بازسازی در Truncated SVD را بیان کنید.

پاسخ

فرض کنید ماتریسAبا ابعاد dxnدارای تجزیه‌ی کامل SVD به‌صورت زیر باشد:

$$A = U\Sigma V^T = \sum_{i=۱}^{r} \sigma_i \; u_i v_i^T$$

است.$r = \text{rank}(A)$که در آن

در Truncated SVD با رتبه‌ی ، تقریب رتبه- از ماتریس زیر ساخته می‌شود:

$$A_c = \sum_{i=۱}^{c} \sigma_i \; u_i v_i^T = \widetilde{U}\widetilde{\Sigma}\widetilde{V}^T$$

تعریف خطای بازسازی

خطای بازسازی معمولاً با نرم فروبنیوس اختلاف بین ماتریس اصلی و تقریب آن اندازه‌گیری می‌شود:

$$E = \| A - A_c \|_F = \sqrt{\sum_{i,j} \left(A_{ij} - (A_c)_{ij}\right)^2}$$

این مقدار نشان می‌دهد تقریب رتبه-c تا چه حد با ماتریس اصلی فاصله دارد.

رابطه‌ی خطا با مقادیر تکین

طبق قضیه‌ی Eckart–Young، خطای بازسازی در Truncated SVD رابطه‌ی بسیار مهمی با مقادیر تکین حذف‌شده دارد:

$$\| A - A_c \|_F = \sqrt{\sum_{i=c+1}^{r} \sigma_i^2}$$

حذف cیعنی خطای بازسازی دقیقاً برابر است با ریشه‌ی مجموع مربعات مقادیر تکینی که بعد از مؤلفه‌ی شده‌اند.

خطای نسبی

برای مقایسه‌ی خطا بین ماتریس‌هایی با مقیاس متفاوت، معمولاً از خطای نسبی استفاده می‌شود:

$$E_{\text{rel}} = \frac{\| A - A_c \|_F}{\| A \|_F}$$

با توجه به اینکه:

$$\| A \|_F = \sqrt{\sum_{i=1}^{r} \sigma_i^2}$$

می‌توان نوشت:

$$E_{\text{rel}} = \sqrt{\frac{\sum_{i=c+1}^{r} \sigma_i^2}{\sum_{i=1}^{r} \sigma_i^2}}$$

نتیجه

پس در Truncated SVD هرچه تعداد بیشتری از مقادیر تکین بزرگ را نگه داریم، خطای بازسازی کمتر می‌شود. از طرف دیگر، اگر هدف کاهش بُعد باشد، باید بین کم بودن خطا و کم بودن رتبه تعادل برقرار کنیم.

سؤال ۴

دو معیار شباهت بردارهای عددی، Cosine Similarity و Euclidean Distance را ضمن ارائه‌ی روابط ریاضی توضیح دهید. بیشینه و کمینه‌ی این دو معیار را تفسیر کنید.

پاسخ

۱) Cosine Similarity

شباهت کسینوسی زاویه‌ی بین دو بردار vوuرا اندازه می‌گیرد و به اندازه‌ی بردارها حساس نیست؛ بلکه بیشتر به جهت آن‌ها توجه می‌کند.

فرمول آن به‌صورت زیر است:

$$\text{CosineSimilarity}(u, v) = \cos(\theta) = \frac{u \cdot v}{\| u \|_2 \quad \| v \|_2} = \frac{\sum_i u_i v_i}{\sqrt{\sum_i u_i^2} \sqrt{\sum_i v_i^2}}$$

بازه‌ی مقدار

به‌طور کلی:

$$-1 \leq \text{CosineSimilarity}(u, v) \leq 1$$

اما اگر مؤلفه‌های بردارها غیرمنفی باشند (مثل بردارهای فراوانی کلمات)، این بازه معمولاً به صورت زیر محدود می‌شود:

$$0 \leq \text{CosineSimilarity}(u, v) \leq 1$$

تفسیر بیشینه و کمینه

بیشینه = 1
دو بردار کاملاً هم‌جهت هستند و زاویه‌ی بین آن‌ها صفر است. در این حالت، از نظر الگوی توزیع مؤلفه‌ها بسیار مشابه‌اند.

مقدار = 0
دو بردار متعامد هستند و از نظر جهتی ارتباطی ندارند.

کمینه = -1
فقط در حالت کلی و زمانی که مؤلفه‌ها می‌توانند منفی باشند رخ می‌دهد. در این حالت دو بردار کاملاً در خلاف جهت هم قرار دارند.

مزیت اصلی Cosine Similarity

این معیار به طول بردارها حساس نیست. به همین دلیل برای داده‌های متنی، بردارهای embedding و مقایسه‌ی اسناد، بسیار مناسب است؛ چون ممکن است دو متن از نظر طول متفاوت باشند اما از نظر موضوعی مشابه باشند.

۲) Euclidean Distance

فاصله‌ی اقلیدسی فاصله‌ی مستقیم دو نقطه در فضا را اندازه می‌گیرد و هم به جهت و هم به اندازه‌ی بردارها حساس است.

فرمول آن به‌صورت زیر است:

$$\text{EuclideanDistance}(u, v) = \| u - v \|_2 = \sqrt{\sum_i (u_i - v_i)^2}$$

بازه‌ی مقدار

فاصله‌ی اقلیدسی همیشه نامنفی است:

$$0 \leq \text{EuclideanDistance}(u, v) < \infty$$

تفسیر بیشینه و کمینه

کمینه = 0

زمانی رخ می‌دهد که دو بردار کاملاً یکسان باشند.

بیشینه

کران بالای مشخصی ندارد و به مقیاس داده‌ها بستگی دارد. هرچه مقدار فاصله بیشتر باشد، دو بردار از هم دورتر و متفاوت‌تر هستند.

مقایسه‌ی مفهومی این دو معیار

Cosine Similarity روی جهت بردارها تمرکز دارد و نسبت به مقیاس مقاوم‌تر است.

Euclidean Distance هم اختلاف در جهت و هم اختلاف در اندازه را لحاظ می‌کند.

بنابراین اگر بردارها نرمال‌سازی یا استانداردسازی نشده باشند، فاصله‌ی اقلیدسی ممکن است تحت تأثیر بزرگی مؤلفه‌ها قرار بگیرد و نتیجه‌ی معنایی مطلوبی ندهد؛ در حالی که شباهت کسینوسی در بسیاری از مسائل متنی انتخاب مناسب‌تری است.

سؤال ۵

ابتدا درباره‌ی استاندارد کردن بردارهای عددی جست‌وجو کنید سپس ضمن ارائه‌ی رابطه‌ی ریاضی، درباره‌ی لزوم استاندارد کردن ماتریس داده‌ها پیش از تجزیه‌ی آن بحث کنید.

پاسخ

استانداردسازی (Standardization) چیست؟

استانداردسازی یک ویژگی (Feature) یا ستون از ماتریس داده به این معناست که مقادیر آن ستون طوری تبدیل شوند که:

میانگین آن برابر صفر شود،

و انحراف معیار آن برابر یک شود.

این کار معمولاً با نرمال‌سازی Z-Score انجام می‌شود:

$$x'_{i,j} = \frac{x_{i,j} - \mu_j}{\sigma_j}$$

که در آن:

است، iدر نمونه‌ی jمقدار ویژگی $x_{i,j}$

در کل داده‌ها، jمیانگین ستون μ_j

است. jانحراف معیار ستون σ_j

چرا استانداردسازی قبل از تجزیه مهم است؟

روش‌هایی مانند SVD و PCA به واریانس و مقیاس داده حساس هستند. این یعنی اگر بعضی ستون‌ها دامنه‌ی عددی بزرگ‌تری داشته باشند، همان ستون‌ها تأثیر بیشتری روی مؤلفه‌های اصلی خواهند گذاشت؛ حتی اگر از نظر مفهومی خیلی مهم نباشند.

در ماتریس‌های متنی مثل Bag of Words یا ماتریس فراوانی کلمات، برخی کلمات بسیار پرتکرارند و به‌طور طبیعی مقادیر بزرگ‌تری دارند. اگر داده را بدون استانداردسازی وارد SVD کنیم، این مشکل‌ها پیش می‌آید:

ستون‌هایی با مقیاس عددی بزرگ‌تر، بیش از حد روی مقادیر تکین و بردارهای تکین اثر می‌گذارند.

مؤلفه‌های اصلی به‌جای اینکه الگوهای واقعی هم‌رخدادی و ساختار معنایی داده را نشان دهند، بیشتر بازتاب‌دهنده‌ی تفاوت مقیاس خام کلمات می‌شوند.

در نتیجه، فضای نهان به‌دست‌آمده ممکن است بیشتر تحت سلطه‌ی کلمات پرتکرار عمومی باشد تا کلمات واقعاً معنادار.

فایده‌ی استانداردسازی پیش از SVD

با استانداردسازی:

هر ویژگی سهم منصفانه‌تری در تحلیل خواهد داشت،

اثر تفاوت در مقیاس خام ستون‌ها کاهش می‌یابد،

و SVD بهتر می‌تواند ساختار همبستگی و الگوهای پنهان واقعی را کشف کند.

در نتیجه، استانداردسازی معمولاً باعث می‌شود فضای نهان استخراج‌شده از داده‌ها معنادارتر، پایدارتر و قابل‌تفسیرتر باشد.

سؤال ۶

در برخورد با ماتریس‌های بزرگ، یکی از روش‌های تجزیه‌ی کارآمد، Randomized SVD است. ضمن ارائه‌ی شبه‌کد این الگوریتم، نحوه‌ی عملکرد و کاربردهای آن را تشریح کنید.

پاسخ

وقتی ماتریس داده بسیار بزرگ باشد، محاسبه‌ی SVD کامل بسیار پرهزینه است. اگر ماتریسی با ابعاد m x n باشد، هزینه‌ی محاسبه‌ی SVD کامل معمولاً در حد زیر است:

$$O(mn \cdot \min(m,n))$$

برای همین در مسائل بزرگ‌مقیاس از Randomized SVD استفاده می‌شود که یک روش تقریبی اما سریع برای محاسبه‌ی مؤلفه‌های اصلی SVD است.

ایده‌ی اصلی Randomized SVD

به‌جای اینکه مستقیماً روی ماتریس بزرگ تجزیه انجام دهیم، ابتدا با استفاده از فرافکنی تصادفی یک زیرفضای کوچک پیدا می‌کنیم که تقریب خوبی از فضای ستونی باشد. سپس SVD را روی یک ماتریس کوچک‌تر انجام می‌دهیم. این کار باعث کاهش شدید هزینه‌ی محاسباتی می‌شود.

شبه کد:

```
Input:
A : ماتریس m×n
k : رتبه‌ی هدف
p : تعداد oversampling (مثلاً 10)
q : تعداد power iterations (مثلاً 2)

1) Ω ← یک ماتریس تصادفی گاوسی با ابعاد n×(k+p)
2) Y ← AΩ

3) for i = 1 to q:
      Y ← A(A^T Y)

4) Q, _ ← QR(Y)

5) B ← Q^T A

6) Ũ, Σ, V^T ← SVD(B)

7) U ← QŨ

Output:
U[:, :k], Σ[:k], V^T[:k, :]
```

توضیح مراحل الگوریتم

۱) نمونه‌برداری تصادفی از فضای ستونی

تولید می‌شود و سپس ماتریس زیر محاسبه می‌شود: $n \times (k+p)$ با ابعاد Ω ابتدا یک ماتریس تصادفی

$$Y = A\Omega$$

ایده این است که ستون‌های Y زیرفضایی را پوشش دهند که بخش مهمی از فضای ستونی A را در خود دارد.

۲) Power Iteration برای افزایش دقت

اگر افت مقادیر تکین آهسته باشد، برای بهتر جدا کردن مؤلفه‌های مهم از مؤلفه‌های ضعیف‌تر، چند بار از تکرار توانی استفاده می‌شود:

$$Y \leftarrow A(A^T Y)$$

این کار باعث می‌شود مؤلفه‌های متناظر با مقادیر تکین بزرگ، پررنگ‌تر شوند و تقریب دقیق‌تر شود.

۳) ساخت پایه‌ی متعامد

سپس با تجزیه‌ی QR، یک پایه‌ی متعامد برای زیرفضای به‌دست‌آمده ساخته می‌شود:

$$Q, _ = QR(Y)$$

ماتریس Q تقریب خوبی از فضای ستونی اصلی A می‌دهد.

۴) فرافکنی ماتریس اصلی به فضای کوچک‌تر

در ادامه، ماتریس بزرگ روی این زیرفضای کوچک فرافکنی می‌شود:

$$B = Q^T A$$

ماتریس B ابعاد بسیار کوچک‌تری نسبت به A دارد و کار با آن ارزان‌تر است.

۵) انجام SVD روی ماتریس کوچک

انجام می‌دهیم: B حالا SVD را روی

$$B = \tilde{U} \Sigma V^T$$

و سپس بردارهای تکین چپ را به فضای اصلی برمی‌گردانیم:

$$U = Q\widetilde{U}$$

در نهایت تقریب رتبه-k از SVD به‌صورت زیر به‌دست می‌آید:

$$U_k, \Sigma_k, V_k^T$$

مزیت Randomized SVD

مهم‌ترین مزیت این روش این است که به‌جای انجام SVD روی یک ماتریس بسیار بزرگ، آن را روی یک ماتریس کوچک‌تر انجام می‌دهد. بنابراین:

سرعت بسیار بالاتری دارد،

حافظه‌ی کمتری مصرف می‌کند،

و برای داده‌های بزرگ‌مقیاس کاملاً مناسب است.

در بسیاری از کاربردهای عملی، دقت آن به‌اندازه‌ای خوب هست که جایگزین مناسبی برای SVD کامل باشد.

کاربردها

Randomized SVD در مسائل زیادی استفاده می‌شود، از جمله:

کاهش بُعد داده‌های بزرگ

Latent Semantic Indexing (LSI) در تحلیل متون

سیستم‌های توصیه‌گر

فشرده‌سازی تصویر و داده

تحلیل ماتریس‌های بزرگ در یادگیری ماشین

تقریب کم‌رتبه‌ی ماتریس‌ها در مسائل علمی و صنعتی