مبانی ریاضی علوم‌داده

گزارش پروژه عملی دوم

پوریا صامتی

۴۰۴۰۹۷۱۴

زمستان ۱۴۰۴

فهرست مطالب

گام صفر ۴

گام اول ۶

نتایج پیاده‌سازی گام اول ۶

پاسخ پرسش اول ۸

پاسخ پرسش دوم ۸

گام دوم ۹

نتایج پیاده‌سازی گام دوم ۹

پاسخ پرسش اول ۱۰

پاسخ پرسش دوم ۱۰

گام سوم ۱۱

نتایج پیاده‌سازی گام سوم ۱۱

پاسخ پرسش اول ۱۳

پاسخ پرسش دوم ۱۳

گام چهارم ۱۴

نتایج پیاده‌سازی گام چهارم - بخش اول و دوم ۱۴

پاسخ پرسش اول ۱۷

پاسخ پرسش دوم ۱۷

گام پنجم ۱۸

نتایج پیاده‌سازی گام پنجم ۱۸

پاسخ پرسش اول ۱۹
پاسخ پرسش دوم ۱۹
گام ششم ۲۰
نتایج پیاده‌سازی گام ششم ۲۰
گام هفتم ۲۱
نتایج پیاده‌سازی گام هفتم ۲۱
پاسخ پرسش اول ۲۳
پاسخ پرسش دوم ۲۳

گام صفر

در این گام، ابتدا سه مجموعه داده را بررسی می‌کنیم. از این سه مجموعه داده برای بررسی داده‌های فیلم‌های مختلف استفاده می‌کنیم. این مجموعه‌ها هر کدام شامل بخشی از داده‌های فیلم‌های مختلف هستند و با استفاده از آنها در گام‌های بعدی، مجموعه‌های مختلفی طراحی می‌گردد.

نمونه‌ای از مجموعه‌داده movielens.csv:

	userId	movieId	rating	timestamp	title	genres
0	0	0	4.0	964982703	Toy Story (1995)	Adventure|Animation|Children|Comedy|Fantasy
1	0	2	4.0	964981247	Grumpier Old Men (1995)	Comedy|Romance
2	0	4	4.0	964982224	Heat (1995)	Action|Crime|Thriller
3	0	23	5.0	964983815	Seven (a.k.a. Se7en) (1995)	Mystery|Thriller
4	0	25	5.0	964982931	Usual Suspects; The (1995)	Crime|Mystery|Thriller

در مجموعه دوم، داده‌های صحت‌سنجی قرار گرفته اند. از این مجموعه‌داده برای ارزیابی مدلهای خود در گام‌های بعدی استفاده می‌کنیم. نمونه‌ای از مجموعه‌داده movielens_val_indices.csv:

	row_inds	col_inds
0	541	743
1	306	387
2	153	958
3	554	649
4	253	719

در مجموعه سوم، داده‌های هر فیلم قرار گرفته اند. در این مجموعه‌داده که بصورت Dictionary می‌باشد، آیدی (id) و نام هر فیلم قرار گرفته است. نمونه‌ای از داده‌های مجموعه movie_map.pkl به شرح زیر است:

```
[(0, 'Toy Story (1995)'),
 (2, 'Grumpier Old Men (1995)'),
 (4, 'Heat (1995)'),
 (23, 'Seven (a.k.a. Se7en) (1995)'),
 (25, 'Usual Suspects; The (1995)')]
```

در ادامه ابعاد، ستون‌ها و سایر ویژگی‌های این سه مجموعه‌داده را بررسی میکنیم. این بررسی موجب می‌گردد که در گام‌های بعدی بتوانیم برای مجموعه‌داده‌های تولید شده از این سه مجموعه، ابعاد درست آنها را بررسی کنیم.

	Data	Value
0	movielens size	(62518, 6)
1	movielens columns	Index(['userId', 'movieId', 'rating', 'timestamp', 'title', 'genres'], dtype='object')
2	movielens_val_indices size	(6252, 2)
3	movielens_val_indices columns	Index(['row_inds', 'col_inds'], dtype='object')
4	data_map size	1050
5	data_map type	<class 'dict'>

گام اول

نتایج پیاده‌سازی گام اول

در این گام، هدف پیاده‌سازی ماتریس سیستم توصیه‌گر است. در نهایت، ماتریسی طراحی می‌شود که **هر سطر آن متناظر با یک کاربر و هر ستون آن متناظر با یک فیلم** باشد و مقدار هر خانه نشان‌دهنده‌ی امتیاز داده‌شده توسط آن کاربر به آن فیلم است.

از آن‌جا که در عمل هر کاربر تنها بخش محدودی از فیلم‌ها را مشاهده و امتیازدهی کرده است، بخش قابل توجهی از این ماتریس شامل مقادیر مشاهده‌نشده (خالی) خواهد بود. بنابراین، برای مدیریت این مقادیر گمشده، دو سازوکار مختلف برای پر کردن خانه‌های خالی در نظر گرفته می‌شود:

- **پر کردن با صفر(Zero-fill)** : تمامی خانه‌های بدون امتیاز با مقدار صفر جایگزین می‌شوند. در این روش، صفر صرفاً به‌عنوان یک مقدار نمایشی برای نبود امتیاز در نظر گرفته می‌شود و به‌معنای نارضایتی کاربر نیست.

- **پر کردن با میانگین کاربر(Mean-fill)** : هر خانه‌ی خالی با میانگین امتیازهایی که همان کاربر به سایر فیلم‌ها داده است پر می‌شود. در صورتی که کاربری هیچ امتیازی به هیچ فیلمی نداده باشد، کل سطر مربوط به آن کاربر با **میانگین سراسری امتیازات** در مجموعه‌داده پر خواهد شد.

این دو روش پر کردن داده، در مراحل بعدی برای مقایسه‌ی عملکرد مدل‌های مختلف سیستم توصیه‌گر مورد استفاده قرار می‌گیرند.

```
A shape: (610, 1050)
A nonzeros: 62518
A sum: 227818.5
```

ابعاد ماتریس نهایی به شرح زیر است:

نمونه‌ای از ۵ سطر اول ماتریس به شرح زیر است:

5 rows × 1,050 columns np.ndarray

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33
0	4.0	0.0	4.0	0.0	4.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	5.0	0.0	5.0	0.0	0.0	0.0	0.0	0.0	3.0	0.0	0
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0	0.0	3.0	2.0	0.0	0.0	3.0	3.0	0.0	0.0	0.0	0.0	0.0	0
4	4.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	4.0	0.0	0.0	0.0	0.0	0.0	0.0	4.0	4.0	3.0	0.0	0.0	0.0	0.0	4.0	0.0	5.0	0.0	0.0	0.0	0.0	0.0	0

در گام بعدی، باید مجموعه‌داده صحت‌سنجی (Validation) را از این مجموعه‌داده جدا کنیم. برای این کار از مجموعه‌داده صحت‌سنجی که در گام اول معرفی کردیم استفاده می‌کنیم. در نتیجه این مجموعه‌داده تشکیل شده و با استفاده از معیارهای زیر گزارش می‌شود:

```
All validation entries zeroed: True
A_fit sum: 205049.0
y_val_true (first 5 samples): [4.  2.  5.  3.  4.5]
y_val_true (last 5 samples): [4. 3. 3. 3. 4.]
```

پاسخ پرسش اول

چرا جدا کردن مجموعه‌ی Validation و صفر کردن خانه‌های آن در ماتریس A_fit منطقی است؟

جدا کردن مجموعه‌ی صحت‌سنجی و صفر کردن مقادیر متناظر با آن در ماتریس A_fit به این دلیل ضروری است که از نشت اطلاعات (Data Leakage) جلوگیری شود. اگر امتیازهای مربوط به صحت‌سنجی در فرآیند آموزش مورد استفاده قرار گیرند، مدل عملاً پاسخ صحیح را از قبل دیده است و ارزیابی انجام‌شده خوش‌بینانه و غیرواقعی خواهد بود (یعنی مقدار خطا قطعا بطور غیرواقعی برای مقادیر صحت‌سنجی کم خواهد بود).

با صفر کردن خانه‌های صحت‌سنجی در مجموعه‌داده مدل تنها بر اساس داده‌های آموزشی یاد می‌گیرد و مقادیر صحت‌سنجی به‌عنوان **اطلاعات ندیده‌شده** تلقی می‌شوند.

در نتیجه، خطای محاسبه‌شده بر مجموعه صحت‌سنجی بیانگر **توان واقعی مدل در تعمیم (Generalization)** است، نه صرفاً بازتولید داده‌های دیده‌شده.

پاسخ پرسش دوم

اگر برخی کاربران یا فیلم‌ها در مجموعه‌ی Validation هیچ نمونه‌ای نداشته باشند، چه اثری بر ارزیابی‌های «سطح کاربر» می‌گذارد؟

اگر برخی کاربران یا فیلم‌ها در مجموعه‌ی صحت‌سنجی هیچ نمونه‌ای نداشته باشند، ارزیابی‌های «سطح کاربر» دچار محدودیت می‌شوند. برای این کاربران امکان محاسبه‌ی معیارهایی مانند RMSE وجود ندارد، زیرا مقدار واقعی در دسترس نیست. در نتیجه، ارزیابی **تنها بر کاربران فعال‌تر** انجام می‌شود که می‌تواند باعث سوگیری در نتایج و نادیده گرفته شدن کاربران کم‌تعامل شود. بنابراین، هنگام گزارش نتایج باید مشخص شود ارزیابی بر روی چه تعداد کاربر انجام شده و این محدودیت در تفسیر عملکرد مدل لحاظ گردد.

گام دوم

نتایج پیاده‌سازی گام دوم

در این گام با استفاده از مکانیزم‌هایی که در گام اول بیان کردیم، درایه‌های ماتریس آموزش را پر می‌کنیم.

در حالت اول خانه‌های خالی را با امتیاز 0 پر می‌کنیم. به عبارتی فرض می‌کنیم اگر کاربری فیلمی را ندیده است، پس همانند این است که کاربر به آن فیلم امتیاز 0 داده است.

در حالت دوم، ابتدا میانگین امتیازاتی که کاربر به فیلم‌هایی که دیده است را بدست می‌آوریم. سپس خانه‌های خالی را با این میانگین‌ها پر می‌کنیم. اگر کاربر هیچ فیلمی ندیده بود (سطر آن کاربر خالی بود)، آنگاه از میانگین سراسری استفاده می‌کنیم و درایه‌های خالی را با میانگین سراسری پر می‌کنیم. در نتیجه برای گزارش اعمال این گام، معیارهای زیر گزارش می‌شوند:

```
A_fit_mean shape: (610, 1050)
A_fit_mean nonzeros: 640500
A_fit_mean sum: 2384086.2986536995
Global Mean: 3.6442789606511927
```

همچنین نمونه‌ای از ماتریس آموزش با مقادیر جایگزین شده به شرح زیر هستند:

5 rows × 1,050 columns np.ndarray

	0	1	2	3	4	5	6	7	8	9	10
0	4.000000	4.400000	4.000000	4.400000	4.000000	4.400000	4.400000	4.400000	4.400000	4.400000	4.400000
1	3.909091	3.909091	3.909091	3.909091	3.909091	3.909091	3.909091	3.909091	3.909091	3.909091	3.909091
2	1.736842	1.736842	1.736842	1.736842	1.736842	1.736842	1.736842	1.736842	1.736842	1.736842	1.736842
3	3.552632	3.552632	3.552632	3.552632	3.552632	3.552632	3.552632	3.552632	3.552632	3.552632	3.552632
4	4.000000	3.567568	3.567568	3.567568	3.567568	3.567568	3.567568	3.567568	3.567568	3.567568	3.567568

پاسخ پرسش اول

چرا صفر کردن خانه‌ها در A_fit (Zero-fill) می‌تواند سوگیری وارد کند؟

پر کردن خانه‌های خالی با صفر، به مدل این پیام را می‌دهد که «کاربر به این فیلم‌ها امتیاز صفر داده است». بنابراین فرض می‌کنیم که کاربر این فیلم را دیده و به آن 0 داده است. در صورتیکه چنین فرضی غلط است. به عبارتی درایه خالی به آن معناست که کاربر فیلم را ندیده و هیچ امتیازی به آن نداده است. در نتیجه مدل ممکن است کم‌امتیاز فرض کند و وزن‌دهی در یادگیری را تحت تأثیر قرار دهد و باعث سوگیری سیستماتیک به سمت پیش‌بینی امتیازهای پایین می‌شود. بنابراین، نتایج RMSE روی مجموعه صحت‌سنجی می‌تواند کمی خوش‌بینانه یا بدبینانه باشد، بسته به نحوه پراکندگی داده‌های واقعی.

همچنین ماتریس اولیه شامل تعداد بسیار بالایی خانه خالی بود. به همین دلیل وقتی خانه‌ها را با مقدار 0 پر کنیم، پیش‌بینی مدل نهایی به سمت 0 میل پیدا می‌کند که امری نامطلوب است.

پاسخ پرسش دوم

چرا پر کردن با میانگین کاربر (User-mean fill) گاهی خطا را عجیب کم می‌کند؟

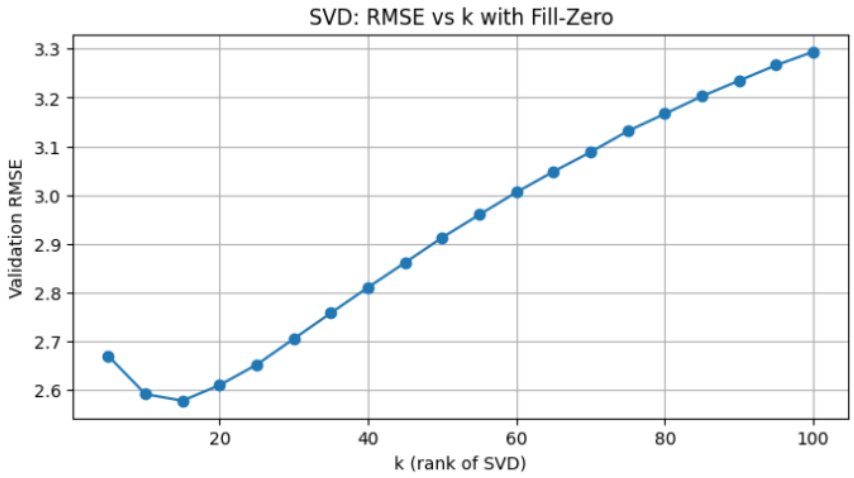
پر کردن خانه‌های خالی با میانگین امتیاز کاربر باعث می‌شود که ماتریس تقریبا متراکم و هموار شود زیرا مدل کمتر با مقادیر صفر مواجه است و کمتر خطا در بازسازی امتیازها ایجاد می‌شود. این کاهش خطا لزوماً به معنای یادگیری واقعی نیست، بلکه بخشی از آن ناشی از تخمین اولیه نزدیک به میانگین است. بنابراین، خطا پایین‌تر ممکن است غیرواقعی و خوش‌بینانه باشد و نشان‌دهنده عملکرد بهتر مدل واقعی نباشد.

گام سوم

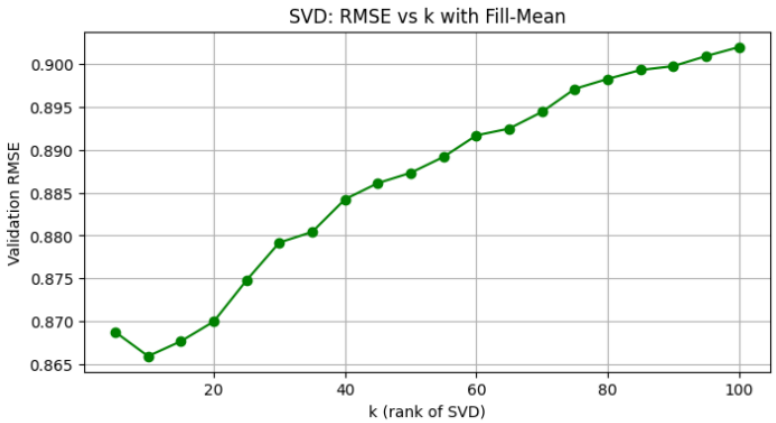
نتایج پیاده‌سازی گام سوم

در این گام میخواهیم بر روی دو ماتریس A_fit_mean و A_fit_zero تجزیه مقادیر ویژه را اعمال کنیم. این تجزیه را با رنک‌های مختلف بر روی این دو روش اعمال می‌کنیم. در نتیجه انتظار داریم که با رنک‌های مختلف ماتریس را بازسازی کنیم و سپس مشاهده کنیم که وضعیت خطا بر روی مجموعه صحت‌سنجی چه مقدار می‌باشد.

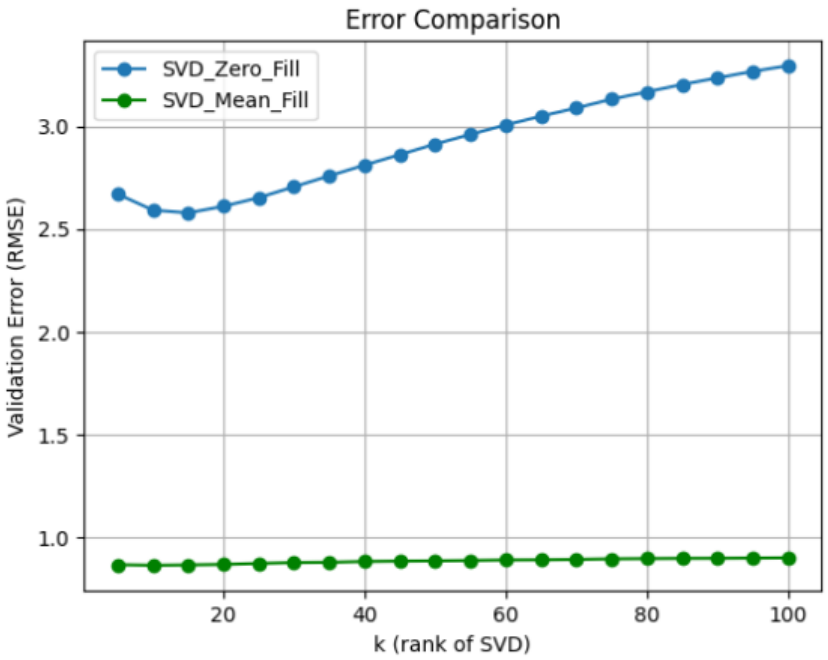
نتیجه میزان خطا پیش و پس از بازسازی با استفاده از تجزیه مقادیر ویژه بر روی ماتریس بر پایه جایگزینی 0 به شرح زیر است:



نتیجه میزان خطا پیش و پس از بازسازی با استفاده از تجزیه مقادیر ویژه بر روی ماتریس بر پایه جایگزینی میانگین به شرح زیر است:



اما مشاهده می‌شود که بازه خطا پس از بازسازی، در روش مبتنی بر جایگزینی میانگین کمتر است. به عبارتی اگر SVD را بر ماتریس A-fit-mean (که در این ماتریس مقادیر گمشده را با میانگین جایگزین کردیم) اعمال کنیم و سپس همان ماتریس را بازسازی کنیم و نتیجه خطای بازسازی را بررسی کنیم، خطای کمتری نسبت به استفاده از ماتریس A-fit-zero مشاهده می‌کنیم. اگر نمودار خطای این دو روش را در کنار هم قرار دهیم، همین مطلب منعکس می‌شود.



پاسخ پرسش اول

آیا همیشه با افزایش k مقدار RMSE بهتر می‌شود؟

خیر، همیشه این‌گونه نیست. با افزایش k، مدل توانایی بازسازی دقیق‌تر داده‌های آموزشی و صحت‌سنجی را پیدا می‌کند، بنابراین خطا در ابتدا کاهش می‌یابد. اما اگر k بیش از حد افزایش یابد، مدل ممکن است به داده‌های آموزش بیش از حد برازش کند (Overfitting) و توانایی تعمیم به داده‌های صحت‌سنجی کاهش یابد. بنابراین، پس از یک مقدار مثل k_{best}، خطا ممکن است دوباره افزایش پیدا کند.

پاسخ پرسش دوم

چرا منحنی fill(user-mean) ممکن است خیلی پایین‌تر از fill (0) باشد؟

پر کردن خانه‌های خالی با میانگین امتیازات همان کاربر باعث می‌شود ماتریس تقریباً هموارتر و متراکم‌تر شود. این تخمین اولیه نزدیک به امتیاز واقعی کاربر است و بنابراین مدل راحت‌تر می‌تواند مقادیر Validation را پیش‌بینی کند. در مقابل، zero-fill مقادیر خانه‌های خالی را صفر فرض می‌کند، که معمولاً با رفتار واقعی کاربران فاصله دارد و باعث افزایش خطا می‌شود. به همین دلیل است که بر مجموعه صحت‌سنجی در این حالت شاهد افزایش خطا هستیم زیرا بر اساس یک فرض نادرست برای مقادیر خالی جایگزینی را انجام میدهد.

گام چهارم

نتایج پیاده‌سازی گام چهارم - بخش اول و دوم

در این بخش، روش **PCA مبتنی بر ماتریس کوواریانس** برای کاهش بعد و بازسازی داده‌ها پیاده‌سازی می‌شود. ابتدا برای ماتریس ورودی (X) میانگین هر ستون محاسبه شده و از داده‌ها کم می‌شود تا ماتریس Center شده (X_c) به دست آید. سپس ماتریس کوواریانس $\left(C = \frac{1}{n-1} X_c^T X_c\right)$ محاسبه می‌شود و با انجام **تجزیه مقدار ویژه(EVD)**، مقادیر ویژه و بردارهای ویژه متناظر استخراج می‌شوند. بردارهای ویژه مربوط به بزرگ‌ترین مقادیر ویژه محورهای اصلی را تشکیل می‌دهند. با انتخاب (k) مولفه اصلی، داده‌ها تقلیل بعد یافته و سپس با ترکیب آن‌ها ماتریس اولیه تقریب زده و بازسازی می‌شود.

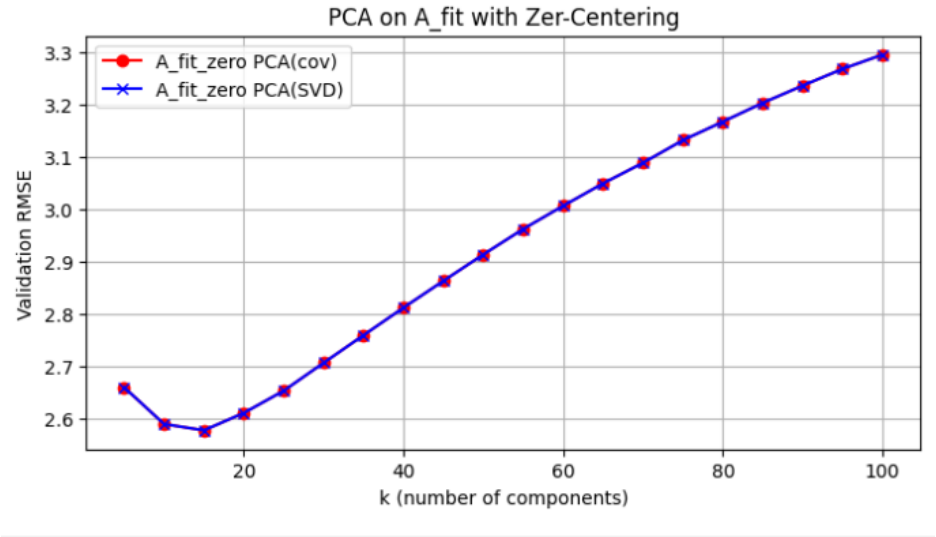
در روش دوم، روش PCA را بر اساس استفاده از **تجزیه مقادیر تکین (SVD)** اعمال می‌کنیم. ابتدا میانگین هر ستون ماتریس ورودی (X) محاسبه و از داده‌ها کم می‌شود تا ماتریس Center شده به دست آید. سپس با انجام SVD، ماتریس به شکل USV^T تجزیه می‌شود، که U و V^T بردارهای منفرد و S مقادیر منفرد (تناظر با مقدار ویژه) را دربردارد. برای هر مقدار k از مولفه‌های اصلی، ماتریس بازسازی شده با ترکیب k مقدار و بردارهای منفرد برتر ساخته می‌شود:

$$A_{\text{pred}} = U_k S_k V_k^T + \text{mean}$$

پس از بازسازی داده‌ها با رنک‌های مختلف، سپس خطا بر روی مجموعه صحت‌سنجی محاسبه شود. این فرآیند برای مقادیر مختلف k تکرار شده و دقت بازسازی با تغییر تعداد مولفه‌های اصلی بررسی می‌شود، که نشان‌دهنده توانایی PCAدر کاهش بعد بدون از دست دادن اطلاعات مهم است.

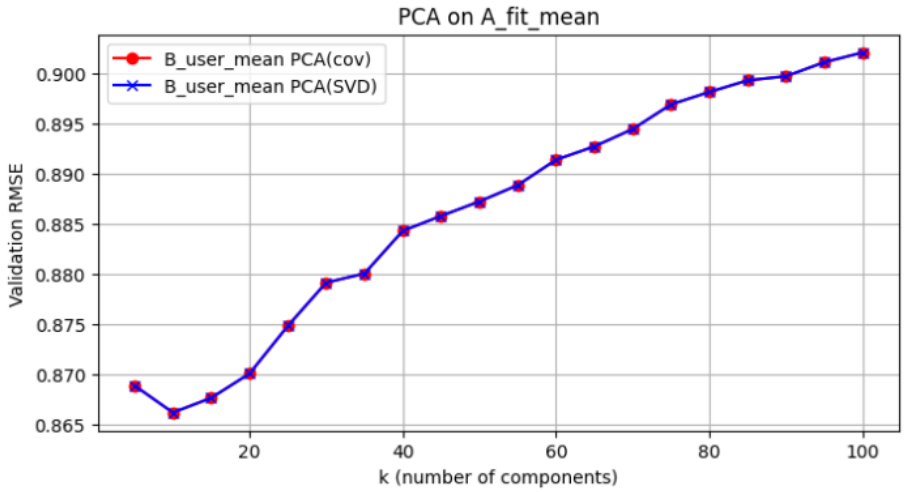
این دو روش برای دو ماتریس (A_fit_mean) و (A_fit_zero) به عنوان ماتریس ورودی اعمال شده و نتایج بازسازی با داده‌های اصلی مقایسه می‌شود تا اثر مرکز کردن داده‌ها و تعداد مولفه‌های اصلی بر دقت بازسازی بررسی گردد.

در بخش بخش اول، اگر ورودی ماتریس (A_fit_zero) باشد، آنگاه اگر PCA بر اساس دو روش گفته شده اعمال شود، خطای پیش‌بینی مجموعه صحت‌سنجی به شکل زیر اعمال نمایش داده خواهد شد:



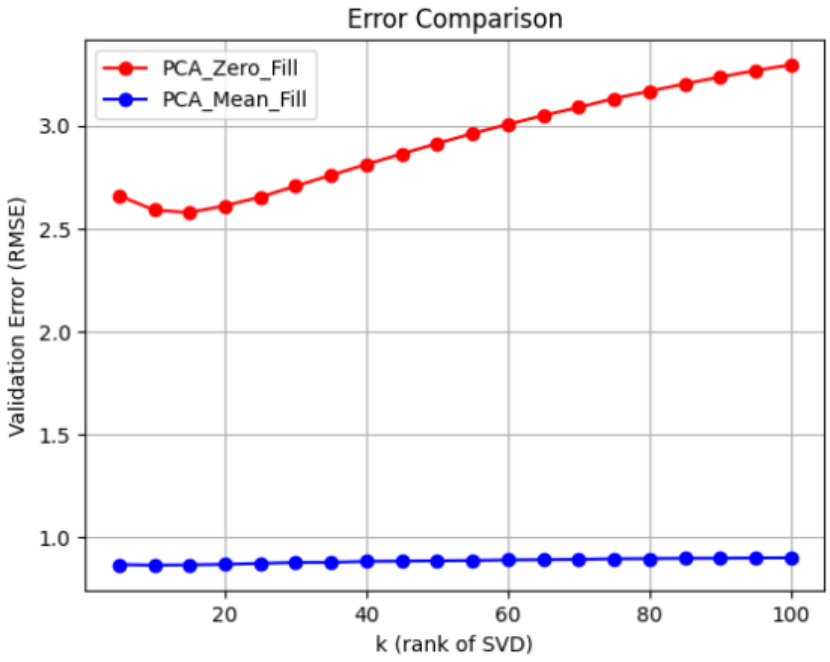
همانطور که مشخص است، اعمال PCA با هر دو روش، دارای نتایج بسیار بسیار مشابهی پس از بازسازی است.

در بخش بخش دوم، اگر ورودی ماتریس (A_fit_mean) باشد، آنگاه اگر PCA بر اساس دو روش گفته شده اعمال شود، خطای پیش‌بینی مجموعه صحت‌سنجی به شکل زیر اعمال نمایش داده خواهد شد:



مجددا مشاهده می‌شود که اعمال PCA با هر دو روش، دارای نتایج بسیار بسیار مشابهی پس از بازسازی است.

طبق مشاهدات مرحله سوم، زمانی که ماتریس ورودی PCA با **مقادیر میانگین ستون‌ها پر شده باشد**، خطای بین ماتریس بازسازی شده و مقادیر واقعی به‌طور قابل توجهی کمتر است، در مقایسه با حالتی که ماتریس ورودی با **مقدار صفر** پر شده باشد. این نکته نشان می‌دهد که **پر کردن داده‌های خالی با میانگین هر سطر** تأثیر قابل توجهی بر دقت بازسازی داده‌ها دارد و استفاده از میانگین به عنوان مقدار اولیه، خطای بازسازی را کاهش می‌دهد. نمودار زیر بر همین موضوع دلالت دارد:



پاسخ پرسش اول

چرا انتظار داریم نتایج(PCA(cov و PCA(SVD) (روی داده‌ها Center شده‌اند) نتایج بسیار نزدیکی بدهند؟

الگوریتم PCA مبتنی بر ماتریس کوواریانس و PCA مبتنی بر SVD در واقع دو راه مختلف برای رسیدن به همان تجزیه خطی داده‌ها هستند. وقتی داده‌ها center شده باشند (ستون‌ها از میانگین خود کم شده باشند):

- تجزیه مقدار ویژه ماتریس کوواریانس $C = \frac{1}{n-1} X_c^T X_c$
- تجزیه مقدار منفرد ماتریس Center شده $X_c = USV^T$

هر دو روش همان محورهای اصلی (Principal Components) و همان مقادیر مرتبط با واریانس را استخراج می‌کنند. بنابراین انتظار می‌رود بازسازی داده‌ها و خطاهای مربوطه بسیار نزدیک باشند، به ویژه زمانی که از تعداد مولفه‌های اصلی یکسان استفاده شود.

پاسخ پرسش دوم

اگر دقیقا یکسان نشدند، دو دلیل ممکن چیست؟

- **خطای عددی و دقت محاسباتی** SVD و EVD: هر دو به روش‌های عددی متفاوتی داده‌ها را تجزیه می‌کنند و مقادیر حاصل به دلیل تقریب عددی ممکن است کمی متفاوت باشند. این اختلاف معمولاً بسیار کوچک است و فقط در چند رقم اعشار مشاهده می‌شود.
- **ترتیب مقادیر ویژه یا بردارهای ویژه**: برخی پیاده‌سازی‌های SVD و EVD ممکن است بردارهای ویژه را با **ضریب منفی** یا با ترتیب اندکی متفاوت بازگردانند. این تغییر ترتیب یا علامت بردارها می‌تواند باعث اختلاف جزئی در بازسازی داده‌ها شود، اما ماهیت اصلی PCA و واریانس توضیح داده شده توسط مولفه‌ها تقریباً بدون تغییر باقی می‌ماند.

گام پنجم

نتایج پیاده‌سازی گام پنجم

حال در گام‌های قبلی، ۶ روش مختلف را بررسی کرده‌ایم. در این گام قصد داریم، بهترین روش را با توجه به میزان خطا پس از بازسازی بر مجموعه Validation و براساس رنک (Rank) بازسازی انتخاب کنیم. برای این منظور از بین ۶ روش و رنک‌های مختلف و نتایج آنها، جست و جو انجام دادیم و در نهایت جدول زیر را تهیه کردیم:

6 rows | 6 rows × 3 columns pd.Dataframe

	method	Best_k	Best_RMSE
0	SVD + fill(0)	15	2.578572
1	SVD + fill(mean)	10	0.865916
2	PCA(cov) + fill(0)	15	2.578147
3	PCA(cov) + fill(mean)	10	0.866142
4	PCA(svd) + fill(zero)	15	2.578147
5	PCA(svd) + fill(mean)	10	0.866142

در نهایت، روش SVD با ماتریس ورودی A-fit-mean با رنک $R = 10$ به عنوان بهترین روش انتخاب می‌شود.

پاسخ پرسش اول

چرا خطای بازسازی در حالت mean-fill با تعداد مولفه‌های بهینه ($best\_k$) کوچکتر است؟

زمانی که ماتریس ورودی با مقادیر میانگین ستون‌ها پر شده باشد(mean-fill) ، داده‌ها پیش از انجام الگوریتم حول میانگین خود Center شده‌اند و توزیع واقعی داده‌ها بهتر حفظ می‌شود. بنابراین، انتخاب بهترین تعداد مولفه‌ها ($best\_k$) باعث می‌شود که بیشترین واریانس داده‌ها توسط مولفه‌های اصلی گرفته شود و بخش زیادی از اطلاعات مفید در بازسازی حفظ گردد. نتیجه این است که خطای بازسازی (RMSE) کمتر و دقت بازسازی بالاتر خواهد بود.

همچنین معیار ارزیابی RMSE است که به میانگین خطا توجه دارد. وقتی که ماتریس مرجع با مقدار میانگین پر شود، این معیار ارزیابی بین ماتریس بازسازی شده و ماتریس مرجع، فاصله کمتری را محاسبه میکند. زیرا ماتریس مرجع به میانگین نزدیک‌تر است.

پاسخ پرسش دوم

اگر best_k در fill(0) خیلی بزرگتر/کوچکتر شد، تفسیر کنید.

وقتی ماتریس ورودی با صفر پر شود(zero-fill)، داده‌ها به طور مصنوعی از مرکز واقعی خود فاصله دارند و بخش زیادی از توزیع اصلی داده‌ها در ماتریس اولیه از دست رفته است. حتی با انتخاب $best\_k$، مولفه‌های اصلی استخراج شده از داده‌های zero-fill نماینده واریانس واقعی نیستند و بازسازی داده‌ها دقیق نخواهد بود. به همین دلیل یا خیلی بزرگ خواهد بود یا خیلی کوچک خواهد بود. بنابراین، خطای بازسازی معمولاً بزرگ‌تر است و افزایش تعداد مولفه‌ها نمی‌تواند به طور کامل اطلاعات از دست رفته ناشی از صفرپر شدن را جبران کند.

گام ششم

نتایج پیاده‌سازی گام ششم

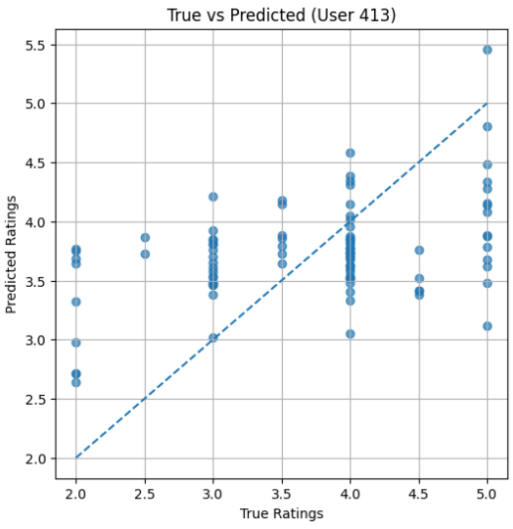
حال در این گام قصد داریم تا کاربری را از مجموعه صحت‌سنجی انتخاب کنیم و به این کاربر فیلم‌هایی را پیشنهاد دهیم. این کار را با توجه به بهترین مدلی که در گام پنجم انتخاب کردیم انجم می‌دهیم. بر این اساس، جدول زیر به کاربر پیشنهاد می‌شود:

	movieId	predicted_rating	title
0	256	4.209631	English Patient; The (1996)
1	286	4.197935	Manhattan (1979)
2	868	4.178226	Corpse Bride (2005)
3	321	4.173752	Birds; The (1963)
4	294	4.162800	Shining; The (1980)
5	457	4.156792	Strangers on a Train (1951)
6	302	4.143391	Arsenic and Old Lace (1944)
7	280	4.141230	Annie Hall (1977)
8	209	4.089251	My Fair Lady (1964)
9	312	4.070009	Ben-Hur (1959)

همچنین ما کاربری را انتخاب کردیم که بیشترین تعداد نمونه را در مجموعه صحت‌سنجی دارد. ایم کاربر عبارت است از:

```
Selected user: 413
Validation count for user 413: 93
```

نمودار امتیازات واقعی و امتیازات پیش‌بینی شده کاربر به شرح زیر است:

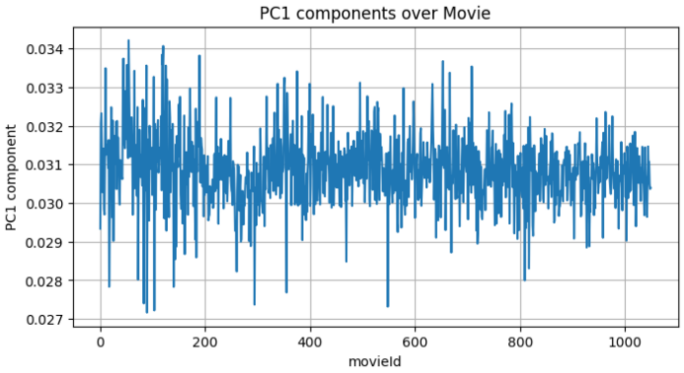


گام هفتم

نتایج پیاده‌سازی گام هفتم

حال در این گام قصد داریم تا فیلم‌های تاثیرگذار بر کاربران را بررسی کنیم. این امر از طریق **PC1** بررسی می‌شود. یعنی **بردار ویژه متناظر با بزرگترین مقدار ویژه**. این بردار دارای تعداد مولفه‌هایی به اندازه تعداد ستون‌های ماتریس A (تعداد فیلم‌ها) می‌باشد. هر **مولفه‌ای که بزرگتر باشد، بدان معنا است که فیلم متناظر با آن مؤلفه، دارای اهمیت بیشتری در بین کاربران می‌باشد و هرچه مؤلفه متناظر با فیلمی کمتر باشد، بدان معناست که آن فیلم در بین کاربران کم اهمیت‌تر بوده است.**

این نمودار با استفاده از PC1 بدست آمده است:



حال با استفاده از داده‌های PC1 ۱۰ فیلم پرنفوذ و ۱۰ فیلم کم نفوذ در بین کاربران را معرفی می‌کنیم:

فیلم‌های پرنفوذ:

	movieId	PC1_Value	title
0	54	0.034210	Net; The (1995)
1	120	0.034063	Coneheads (1993)
2	118	0.033842	City Slickers II: The Legend of Curly's Gold (1994)
3	189	0.033817	Nutty Professor; The (1996)
4	44	0.033738	Casper (1995)
5	653	0.033673	Hollow Man (2000)
6	51	0.033576	Johnny Mnemonic (1995)
7	119	0.033570	Cliffhanger (1993)
8	88	0.033559	Santa Clause; The (1994)
9	125	0.033557	Free Willy (1993)

فیلم‌های کم نفوذ

	movieId	PC1_Value	title
0	89	0.027162	Shawshank Redemption; The (1994)
1	103	0.027219	Forrest Gump (1994)
2	548	0.027315	American Beauty (1999)
3	294	0.027368	Shining; The (1980)
4	83	0.027399	Pulp Fiction (1994)
5	355	0.027682	Fifth Element; The (1997)
6	140	0.027829	Schindler's List (1993)
7	17	0.027833	Twelve Monkeys (a.k.a. 12 Monkeys) (1995)
8	809	0.027995	Eternal Sunshine of the Spotless Mind (2004)
9	72	0.028015	Star Wars: Episode IV - A New Hope (1977)

پاسخ پرسش اول

معنی وزن مثبت/منفی در PC1 چیست؟

در PC-1، هر **مولفه اصلی** ترکیبی خطی از ویژگی‌های اولیه را نشان می‌دهند که هر ویژگی چه سهمی در آن مولفه دارد . **علامت مثبت یا منفی وزن‌ها** صرفاً جهت تغییر ویژگی‌ها را نسبت به محور مولفه نشان می‌دهد:

- **وزن مثبت**: افزایش مقدار آن ویژگی، مولفه اصلی را افزایش می‌دهد.
- **وزن منفی**: افزایش مقدار آن ویژگی، مولفه اصلی را کاهش می‌دهد.

بنابراین علامت وزن‌ها **معنای رتبه‌بندی یا ارزش مطلق ویژگی‌ها را تغییر نمی‌دهد**، بلکه جهت اثر آن ویژگی در مولفه اصلی را مشخص می‌کند. اما مقدار هر مولفه در بردار PC-1 نشان‌دهنده تاثیرگذاری آن مولفه می‌باشد.

پاسخ پرسش دوم

چرا برخی از فیلم‌ها قدرمطلق بزرگتری دارند؟ این چه چیزی درباره "فیلم تعیین‌کننده در سلیقه‌ها" می‌گوید؟

ویژگی‌هایی که در یک مولفه اصلی قدر مطلق وزن بزرگ‌تری دارند، بیشترین تأثیر را بر آن مولفه دارند و به اصطلاح «ویژگی‌های تعیین‌کننده» آن مولفه هستند. در مثال فیلم‌ها، این یعنی فیلم‌هایی که وزن مطلق بیشتری دارند، بیشترین نقش را در تفاوت سلیقه‌ها بازی می‌کنند و تفاوت افراد در این مولفه بیش از سایر فیلم‌ها تحت تأثیر آن‌هاست. به عبارتی این فیلم‌های با بیشترین مقدار در PC-1 هستند که سلیقه کاربران را نشان می‌دهند زیرا در بین کاربران بیشترین تاثیر را دارند.