Sistema de Recomendação de filmes por meio do método PCA

Trabalho Final - Álgebra Linear e Aplicações (SME0142)

Grupo 31 - Allan Garcia, Eduarda Neumann e Teo Sobrino

ICMC-USP

4 de dezembro de 2023

1. Introdução
2. Referencial Teórico
3. Organização dos dados
4. Exemplo

1 Introdução

2 Referencial Teórico

3 Organização dos dados

4 Exemplo

- Matriz de preferências
- Reduzir a dimensionalidade com PCA
- Dois métodos para selecionar as k principais componentes: número absoluto e porcentagem
- Reconstruir a matriz de preferências apenas com as informações das k principais componentes
- Usar a matriz reconstruída para fazer recomendações

1 Introdução

2 Referencial Teórico

3 Organização dos dados

4 Exemplo

Suponha que existam u usuários avaliadores e f filmes avaliados representados na matriz de preferências $X \in M_{u \times f}$. Como realizar PCA nessa matriz usando k autovetores?

Passo 1

Calcula-se a matriz centralizada de preferências, $\widehat{X} \in M_{uxf}$, subtraindo de cada linha da matriz de preferências sua média. Isso busca resolver o problema de que alguns usuários possuem avaliações mais críticas do que outros.

Passo 2

Calcula-se a matriz de covariância $\Sigma = \frac{1}{f-1}(\widehat{X}^T\widehat{X}) \in M_{f\times f}$, e obtém-se uma matriz $P \in M_{f\times k}$, onde cada coluna é um autovetor de Σ.

Passo 3

Projeta-se a matriz X no espaço $M_{u \times k}$, sendo $Y = X^T P$ essa projeção. Tal projeção possui apenas as características mais relevantes da matriz original.

Passo 4

Utilizando apenas as informações contidas nos k autovetores, projetou-se a matriz Y no espaço $M_{u \times f}$, obtendo a matriz de previsões $Z = PY^T = PP^TX$.

Interpretação

Na matriz Z, as entradas possuem os valores extrapolados do espaço M_{uxk} para o espaço M_{uxf} com o auxílio dos k autovetores. Por isso, as entradas da matriz Z correspondentes às entradas da matriz X com valores nulo (representando um filme não avaliado ainda), podem ser usadas como uma previsão dos gostos do usuário para esses filmes.

1 Introdução

2 Referencial Teórico

3 Organização dos dados

4 Exemplo

- Utilizou-se o dataset MovieLens 25M, que contém as avaliações de 162.000 usuários sobre 62.000 filmes (não é obrigatório que um usuário tenha avaliado todos os filmes).
- Como cada execução no Google Colab pode usar, no máximo, 12.7 GB de memória RAM, e a matriz de preferências tem tamanho $u \cdot f$, limitou-se a quantidade de usuários para 1000, de modo que $u = 1000$ e $f = 11920$. A redução no número de filmes ocorre pois os filmes que não foram avaliados por nenhum dos usuários selecionados também foram descartados, já que não seria possível fazer uma previsão sobre estes filmes.
- Para a manipulação desses dados foram usadas as bibliotecas numpy e pandas.

1 Introdução

2 Referencial Teórico

3 Organização dos dados

4 Exemplo

Matrizes Iniciais

- Dataset MovieLens 25M

userId	movieId	rating
38	68954	3.50
52	3617	2.50
402	91529	4.50
556	7153	5.00
705	8644	3.00
727	2571	4.50
871	71033	4.50
871	5630	3.50
899	6350	0.50
...	...	...

Matrizes Iniciais

- Matriz de Preferências ($X \in M_{uxf}$)

userId \ movieId	1	2	3	⋯	206499	206523
1	0	0	0	⋯	0	0
2	3.5	0	0	⋯	0	0
3	4	0	0	⋯	0	0
4	3	0	0	⋯	1.3	0
...	...	...	...	...	...	...
997	4.5	3.5	0	⋯	0	0
998	0	0	4	⋯	0	0
999	0	0	0	⋯	0	0
1000	0	0	0	⋯	0	3.4