SVD

Inteligência na Web e Big Data

Fabricio Olivetti de França e Thiago Ferreira Covões
folivetti@ufabc.edu.br, thiago.covoes@ufabc.edu.br

Centro de Matemática, Computação e Cognição
Universidade Federal do ABC

UFABC

CMCC
Centro de Matemática, Computação e Cognição
Center for Mathematics, Computation and Cognition

Decomposição de Matrizes

SVD

Uma outra forma de reduzir a dimensionalidade é decompondo a matriz.

Digamos que exista duas matrizes U e V de tal forma que:

$$X = U \cdot V^T$$

SVD

Sendo:

$$X \in \mathbb{R}^{n \times d}$$

$$U \in \mathbb{R}^{n \times f}$$

$$V \in \mathbb{R}^{d \times f}$$

Poderíamos representar X através de duas matrizes muito menores, caso $f << d$.

Recomendação

Suponha que você seja dono de uma locadora de filmes.

No ato da devolução você dá um bônus de $0,50$ caso o cliente preencha um formulário atribuindo uma nota, de 1 a 5, para o filme que acaba de assistir.

Recomendação

Se for possível fazer isso através da web podemos coletar notas de milhares, talvez milhões de usuários.

Com base nesses dados é possível aprender a função:

$F(u, f) = n$

Ou, dado um usuário e um filme, retorne a nota que ele dará para esse filme.

Recomendação

De posse dessa função é possível recomendar todo filme f que o usuário u não assistiu ainda e que:

$F(u, f) = 5$

Sistema de Recomendação

Encontrar uma forma de predizer se um cliente irá gostar ou não de certo produto,

ou

predizer apenas a lista de produtos que ele irá gostar

compõe um Sistema de Recomendação.

Recomendação Colaborativa

Quando isso é feito baseado na experiência acumulada de diversos clientes (colaborativamente), dizemos que é um sistema de Recomendação Colaborativa.

O processo para extrair essa informação é chamado Filtragem Colaborativa.

Sistemas de Recomendação

Os Sistemas de Recomendação são compostos por duas entidades:

- Usuários: com a informação dos itens que eles tem preferência.
- Itens: com a informação de quais usuários os consumiram.

Sistemas de Recomendação

A informação que temos pode ser representada em forma matricial pela Matriz de Utilidade.

Nessa matriz as linhas representam os usuários e as colunas os itens.

O valor de cada elemento da matriz representa o quanto o usuário gosta de determinado item.

Sistemas de Recomendação

Mas, na prática, não teremos todos esse valores.

O objetivo principal é preencher esses espaços com valores próximos dos reais.

Ou…preencher apenas os espaços que são mais relevantes para o negócio em questão.

Filtragem Colaborativa

Dada a matriz de utilidade precisamos verificar quais usuários tem o gosto mais similar entre si.

Uma possibilidade é representar cada usuário como um conjunto de itens que ele gostou.

Filtragem Colaborativa

Para isso basta formar o conjunto com os itens que ele atribuiu nota superior a um limiar.

Com isso podemos usar a Similaridade de Jaccard vista anteriormente.

Filtragem Colaborativa

Mas, estaremos ignorando a informação dos itens que o usuário não gostou.

Além disso, a ausência de um item pode tanto significar que ele não conhece ainda ou não quer consumir.

SVD – Singular-value Decompostion

A ideia dessa decomposição, no nosso caso, é que os usuários tendem a classificar produtos que os interessam.

Isso torna a matriz esparsa.

Além disso cada usuário tem um subconjunto bem definido de preferências (embora não saibamos qual).

SVD – Singular-value Decompostion

Com o SVD nós decompomos a matriz em uma matriz de usuários (U) e uma matriz de itens (V).

A matriz de usuários relaciona cada usuário com “f” atributos escondidos em nossa matriz original.

SVD – Singular-value Decompostion

Tomando o exemplo de uma locadora de filmes, esses “f” atributos poderiam corresponder ao gênero do filme.

$F = \{ao, comdia, romance, aventura\}$

Um certo usuário pode ser representado pelo vetor:

$u = [0.9, 0.3, -0.6, 0.8]$

Indicando sua preferência por filmes de ação e aventura.

SVD – Singular-value Decompostion

A matriz V, por sua vez, relaciona os itens aos “f” atributos escondidos.

Ou seja, se um filme pode ser representado pelo vetor:

$v = [-0.9, 0.6, 0.8, -0.4]$

Indicando que é um filme de comédia-romântica.

SVD – Singular-value Decompostion

Esses atributos escondidos não são bem definidos.

Eles surgem da correlação existente entre os diferentes usuários e os diferentes itens.

SVD

Para encontrar os valores dessas duas matrizes, podemos estimar em um processo iterativo.

SVD

Inicialmente “chutamos” valores para elas.

HP1	HP2	HP3	TW	SW4	SW5	SW6	
A	0.39			0.98	-1.37		
B	0.71	0.71	-1.41				
C				-1.34	0.27	1.07	
D		0					0

SVD

U	d1	d2
A	1	1
B	1	1
C	1	1
D	1	1

V	HP1	HP2	HP3	TW	SW4	SW5	SW6
d1	1	1	1	1	1	1	1
d2	1	1	1	1	1	1	1

SVD

Precisamos agora de uma função de avaliação que nos indique o quão próximo estamos da matriz original.

SVD

Vamos utilizar a Raiz do Erro Quadrático Médio que é calculada da seguinte maneira:

- Calcula-se o quadrado da diferença entre todos os elementos não-nulos de M e os correspondentes de U x V.
- Calcula a média desses valores.
- Calcula a raiz dessa média.

SVD

HP1	HP2	HP3	TW	SW4	SW5	SW6	
A	0.39			0.98	-1.37		
B	0.71	0.71	-1.41				
C				-1.34	0.27	1.07	
D		0					0

SVD

UxV	HP1	HP2	HP3	TW	SW4	SW5	SW6
A	2	2	2	2	2	2	2
B	2	2	2	2	2	2	2
C	2	2	2	2	2	2	2
D	2	2	2	2	2	2	2

SVD

M - UxV	HP1	HP2	HP3	TW	SW4	SW5	SW6
A	-1,61			-1,02	-3,37		
B	-1,29	-1,29	-3,41				
C				-3,34	-1,73	-0,93	
D		-2					-2

SVD

$(M–UxV)^2$	HP1	HP2	HP3	TW	SW4	SW5	SW6
A	2,6			1	11,4		
B	1,7	1,7	11,6				
C				11,1	3	0,86	
D		4					4

SVD

HP1	HP2	HP3	TW	SW4	SW5	SW6	
A	0.39			0.98	-1.37		
B	0.71	0.71	-1.41				
C				-1.34	0.27	1.07	
D		0					0

$MSE = 4,81$

$RMSE = 2,19$

SVD

Agora, para cada variável, vamos estimar um valor que reduz nosso erro quadrático.

SVD

Vamos tomar o elemento (A,d1) da matriz U.

U	d1	d2
A	X	1
B	1	1
C	1	1
D	1	1

SVD

Ao torná-lo uma variável nossa matriz resultante fica:

UxV	HP1	HP2	HP3	TW	SW4	SW5	SW6
A	X+1	X+1	X+1	X+1	X+1	X+1	X+1
B	2	2	2	2	2	2	2
C	2	2	2	2	2	2	2
D	2	2	2	2	2	2	2

SVD

Com isso basta compararmos o erro quadrático da primeira linha em relação a primeira linha da matriz original.

UxV	HP1	HP2	HP3	TW	SW4	SW5	SW6
A	X+1	X+1	X+1	X+1	X+1	X+1	X+1

HP1	HP2	HP3	TW	SW4	SW5	SW6
A	0.39			0.98	-1.37	

SVD

O erro quadrático fica:

$$(X + 1 - 0,39)^2 + (X + 1 - 0,98)^2 + (X + 1 + 1,37)^2$$

SVD

Queremos o valor de x que minimiza a Eq.:

$$(X + 1 - 0,39)^2 + (X + 1 - 0,98)^2 + (X + 1 + 1,37)^2$$

Derivamos e igualamos a zero:

$$2 * (X + 1 - 0,39 + X + 1 - 0,98 + X + 1 + 1,37) = 0$$

Resolvendo temos que:

$(3X + 3) = 0; X = -1$

SVD

E a decomposição fica:

U	d1	d2
A	-1	1
B	1	1
C	1	1
D	1	1

SVD

Repetimos o processo para cada variável e por várias iterações, até que o cálculo estabilize e a matriz resultante pouco se altere.

SVD

Vamos utilizar o gradiente descendente para resolver nosso problema, minimizar:

$$MSE = \frac{1}{2} \sum (M - U \times V)^2$$

SVD

Equivalente a:

$$MSE = \frac{1}{2} \sum_{i=0..m, j=0..n} (m_{i,j} - \sum_{k=0}^{f} u_{i,k} \cdot v_{k,j})^2$$

SVD

Temos que encontrar a direção para cada elemento $u_{i,k}$:

$$\frac{\partial MSE}{\partial u_{i,k}} = -(m_{i,j} - \sum_{k=0}^{f} u_{i,k} \cdot v_{k,j}) \cdot v_{k,j}$$

SVD

Temos que encontrar a direção para cada elemento $v_{k,j}$:

$$\frac{\partial MSE}{\partial u_{i,k}} = -(m_{i,j} - \sum_{k=0}^{f} u_{i,k} \cdot v_{k,j}) \cdot u_{i,k}$$

SVD

Definindo:

$$erro(i,j) = m_{i,j} - \sum_{k=0}^{f} u_{i,k} \cdot v_{k,j}$$

SVD

Da mesma forma que na regressão linear, podemos regularizar a otimização das matrizes U e V:

$$u_{i,k} = (1 - \alpha\lambda) \cdot u_{i,k} + \alpha \cdot (erro(i,j) \cdot v_{k,j}$$

$$v_{k,j} = (1 - \alpha\lambda) \cdot v_{k,j} + \alpha \cdot (erro(i,j) \cdot u_{i,k}$$

SVD - Spark

u’, v’ são as matrizes de decomposição, u”, v”as matrizes corrigidas, u, i o usuário e item de interesse e r a nota:

Mapper:

- Recebe (u, i, u’[u], v’[i], r)
- Emite (u+”$_u$”, u”[u]) e (v+”$_v$”, v”[u])

Combiner / Reducer:

- Recebe (z, [z”])
- Retorna (z, soma [z”])

SVD - Spark

Em seguida substitui cada linha de u’ e v’ de acordo

As matrizes U e V são RDDs em que cada elemento tem a chave i (j) e o valor uma lista com f elementos

O erro é calculado em cada tupla (i,j) que existe na base de notas

SVD - Spark

Implemente esse algoritmo no Apache Spark como exercício!