Single Value Decomposition

SVD

Lema. Toda matriz A, com dimensões nxd, admite uma fatoração A= UDV^T.

U é uma matriz com n linhas e r colunas
(r dimensão de A)
D é uma matriz diagonal r x r;
V é uma matriz com d linhas e r colunas;
U e V são ortonormais

SVD: Exemplo

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 3 & 3 & 3 & 0 & 0 \\ 4 & 4 & 4 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 2 & 0 & 4 & 4 \\ 0 & 0 & 0 & 5 & 5 \\ 0 & 1 & 0 & 2 & 2 \end{bmatrix} = \underbrace{\begin{bmatrix} .13 & .02 & -.01 \\ .41 & .07 & -.03 \\ .55 & .09 & -.04 \\ .68 & .11 & -.05 \\ .15 & -.59 & .65 \\ .07 & -.73 & -.67 \\ .07 & -.29 & .32 \end{bmatrix}}_{U} \underbrace{\begin{bmatrix} 12.4 & 0 & 0 \\ 0 & 9.5 & 0 \\ 0 & 0 & 1.3 \end{bmatrix}}_{\Sigma} \underbrace{\begin{bmatrix} .56 & .59 & .56 & .09 & .09 \\ .12 & -.02 & .12 & -.69 & -.69 \\ .40 & -.80 & .40 & .09 & .09 \end{bmatrix}}_{V^{\mathrm{T}}}$$

SVD: Motivação

Lema. Toda matriz A, com dimensões nxd, admite uma fatoração A= UDV^T

Aplicações

Enxergando A como uma matriz de dados (e.g. cada linha como um dado) temos aplicações em

- Compressão/ visualização
- Ranking
- Clustering

SVD: Interpretação

	Matrix	Alien	Star Wars	Casablanca	Titanic
Joe	1	1	1	0	0
Jim	3	3	3	0	0
John	4	4	4	0	0
Jack	5	5	5	0	0
Jill	0	0	0	4	4
Jenny	0	0	0	5	5
Jane	0	0	0	2	2

Figure 11.6: Ratings of movies by users

SVD: Interpretação

	Matrix	Alien	Star Wars	Casablanca	Titanic
Joe	1	1	1	0	0
Jim	3	3	3	0	0
John	4	4	4	0	0
Jack	5	5	5	0	0
Jill	0	0	0	4	4
Jenny	0	0	0	5	5
Jane	0	0	0	2	2

Figure 11.6: Ratings of movies by users

- Garotos gostam de Sci-Fi
- Meninas gostam de romance

SVD: Interpretação

Hipótese

- O rating de um usuário ***u*** para um filme ***f*** pode ser explicado pelo *perfil* de ***u*** (gosto em relação aos diferentes gêneros) e pelo pefil do filme ***f*** *(*aderência em relação aos diferentes gêneros)
- Generos não precisam ser conhecidos/escolhidos a priori

SVD: Interpretação

	Matrix	Alien	Star Wars	Casablanca	Titanic
Joe	1	1	1	0	0
Jim	3	3	3	0	0
John	4	4	4	0	0
Jack	5	5	5	0	0
Jill	0	0	0	4	4
Jenny	0	0	0	5	5
Jane	0	0	0	2	2

Figure 11.6: Ratings of movies by users

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 3 & 3 & 3 & 0 & 0 \\ 4 & 4 & 4 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 0 & 0 & 4 & 4 \\ 0 & 0 & 0 & 5 & 5 \\ 0 & 0 & 0 & 2 & 2 \end{bmatrix} = \begin{bmatrix} .14 & 0 \\ .42 & 0 \\ .56 & 0 \\ .70 & 0 \\ 0 & .60 \\ 0 & .75 \\ 0 & .30 \end{bmatrix} \begin{bmatrix} 12.4 & 0 \\ 0 & 9.5 \end{bmatrix} \begin{bmatrix} .58 & .58 & .58 & 0 & 0 \\ 0 & 0 & 0 & .71 & .71 \end{bmatrix}$$

SVD: Interpretação

	Matrix	Alien	Star Wars	Casablanca	Titanic
Joe	1	1	1	0	0
Jim	3	3	3	0	0
John	4	4	4	0	0
Jack	5	5	5	0	0
Jill	0	0	0	4	4
Jenny	0	0	0	5	5
Jane	0	0	0	2	2

Figure 11.6: Ratings of movies by users

Perfil de Joe em relação a Sci-Fi e Romance

Peso do conceito Sci-Fi

Perfil de Star Wars em relação a Sci-Fi e romance

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 3 & 3 & 3 & 0 & 0 \\ 4 & 4 & 4 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 0 & 0 & 4 & 4 \\ 0 & 0 & 0 & 5 & 5 \\ 0 & 0 & 0 & 2 & 2 \end{bmatrix} = \begin{bmatrix} \boxed{.14 \quad 0} \\ .42 \quad 0 \\ .56 \quad 0 \\ .70 \quad 0 \\ 0 \quad .60 \\ 0 \quad .75 \\ 0 \quad .30 \end{bmatrix} \begin{bmatrix} \boxed{12.4} & 0 \\ 0 & 9.5 \end{bmatrix} \begin{bmatrix} .58 & .58 & \boxed{.58} & 0 & 0 \\ 0 & 0 & \boxed{0} & .71 & .71 \end{bmatrix}$$

SVD: Interpretação

Exemplo com ratings mais misturados

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 3 & 3 & 3 & 0 & 0 \\ 4 & 4 & 4 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 2 & 0 & 4 & 4 \\ 0 & 0 & 0 & 5 & 5 \\ 0 & 1 & 0 & 2 & 2 \end{bmatrix} = \underbrace{\begin{bmatrix} .13 & .02 & -.01 \\ .41 & .07 & -.03 \\ .55 & .09 & -.04 \\ .68 & .11 & -.05 \\ .15 & -.59 & .65 \\ .07 & -.73 & -.67 \\ .07 & -.29 & .32 \end{bmatrix}}_{U} \underbrace{\begin{bmatrix} 12.4 & 0 & 0 \\ 0 & 9.5 & 0 \\ 0 & 0 & 1.3 \end{bmatrix}}_{\Sigma} \underbrace{\begin{bmatrix} .56 & .59 & .56 & .09 & .09 \\ .12 & -.02 & .12 & -.69 & -.69 \\ .40 & -.80 & .40 & .09 & .09 \end{bmatrix}}_{V^{T}}$$

SVD: Interpretação

Se A está contido em um subespaço de dimensão **r** então temos que A= UDV^T e

$$A = \sum_{i=1}^{r} \sigma_i \mathbf{u_i} \mathbf{v_i}^T.$$

- v_i é a i-ésima direção 'mais importante' de A
- u_i são os componentes dos pontos de A na direção v_i
- σ_i =D(i,i) é usado para dar um peso a direção v_i

SVD: interpretação

Se truncarmos a soma abaixo em k termos conseguimos aproximações/versões compactas da matriz de dados A

$$A = \sum_{i=1}^{r} \sigma_i \mathbf{u_i} \mathbf{v_i}^T.$$

SVD: Redução de dimensionalidade

$$\begin{bmatrix} \mathbf{1} & \mathbf{1} & \mathbf{1} & 0 & 0 \\ \mathbf{3} & \mathbf{3} & \mathbf{3} & 0 & 0 \\ \mathbf{4} & \mathbf{4} & \mathbf{4} & 0 & 0 \\ \mathbf{5} & \mathbf{5} & \mathbf{5} & 0 & 0 \\ 0 & \mathbf{2} & 0 & \mathbf{4} & \mathbf{4} \\ 0 & 0 & 0 & \mathbf{5} & \mathbf{5} \\ 0 & \mathbf{1} & 0 & \mathbf{2} & \mathbf{2} \end{bmatrix} = \begin{bmatrix} \mathbf{0.13} & 0.02 & -0.01 \\ \mathbf{0.41} & 0.07 & -0.03 \\ \mathbf{0.55} & 0.09 & -0.04 \\ \mathbf{0.68} & 0.11 & -0.05 \\ 0.15 & \mathbf{-0.59} & \mathbf{0.65} \\ 0.07 & \mathbf{-0.73} & \mathbf{-0.67} \\ 0.07 & \mathbf{-0.29} & \mathbf{0.32} \end{bmatrix} \times \begin{bmatrix} \mathbf{12.4} & 0 & 0 \\ 0 & \mathbf{9.5} & 0 \\ 0 & 0 & \mathbf{1.3} \end{bmatrix} \times \begin{bmatrix} \mathbf{0.56} & \mathbf{0.59} & \mathbf{0.56} & 0.09 & 0.09 \\ 0.12 & -0.02 & 0.12 & \mathbf{-0.69} & \mathbf{-0.69} \\ 0.40 & \mathbf{-0.80} & 0.40 & 0.09 & 0.09 \end{bmatrix}$$

J. Leskovec, A. Rajaraman, J. Ullman: Mining of Massive Datasets, http://www.mmds.org

SVD: Redução de dimensionalidade

$$\begin{bmatrix} \mathbf{1} & \mathbf{1} & \mathbf{1} & 0 & 0 \\ \mathbf{3} & \mathbf{3} & \mathbf{3} & 0 & 0 \\ \mathbf{4} & \mathbf{4} & \mathbf{4} & 0 & 0 \\ \mathbf{5} & \mathbf{5} & \mathbf{5} & 0 & 0 \\ 0 & \mathbf{2} & 0 & \mathbf{4} & \mathbf{4} \\ 0 & 0 & 0 & \mathbf{5} & \mathbf{5} \\ 0 & \mathbf{1} & 0 & \mathbf{2} & \mathbf{2} \end{bmatrix} = \begin{bmatrix} \mathbf{0.13} & 0.02 & -0.01 \\ \mathbf{0.41} & 0.07 & -0.03 \\ \mathbf{0.55} & 0.09 & -0.04 \\ \mathbf{0.68} & 0.11 & -0.05 \\ 0.15 & \mathbf{-0.59} & \mathbf{0.65} \\ 0.07 & \mathbf{-0.73} & \mathbf{-0.67} \\ 0.07 & \mathbf{-0.29} & \mathbf{0.32} \end{bmatrix} \times \begin{bmatrix} \mathbf{12.4} & 0 & 0 \\ 0 & \mathbf{9.5} & 0 \\ 0 & 0 & \cancel{\mathbf{1.3}} \end{bmatrix} \times \begin{bmatrix} \mathbf{0.56} & \mathbf{0.59} & \mathbf{0.56} & 0.09 & 0.09 \\ 0.12 & -0.02 & 0.12 & \mathbf{-0.69} & \mathbf{-0.69} \\ 0.40 & \mathbf{-0.80} & 0.40 & 0.09 & 0.09 \end{bmatrix}$$

J. Leskovec, A. Rajaraman, J. Ullman: Mining of Massive Datasets, http://www.mmds.org

SVD: Redução de dimensionalidade

$$\begin{bmatrix} \mathbf{1} & \mathbf{1} & \mathbf{1} & 0 & 0 \\ \mathbf{3} & \mathbf{3} & \mathbf{3} & 0 & 0 \\ \mathbf{4} & \mathbf{4} & \mathbf{4} & 0 & 0 \\ \mathbf{5} & \mathbf{5} & \mathbf{5} & 0 & 0 \\ 0 & \mathbf{2} & 0 & \mathbf{4} & \mathbf{4} \\ 0 & 0 & 0 & \mathbf{5} & \mathbf{5} \\ 0 & \mathbf{1} & 0 & \mathbf{2} & \mathbf{2} \end{bmatrix} \approx \begin{bmatrix} \mathbf{0.13} & 0.02 & -0.01 \\ \mathbf{0.41} & 0.07 & -0.03 \\ \mathbf{0.55} & 0.09 & -0.04 \\ \mathbf{0.68} & 0.11 & -0.05 \\ 0.15 & \mathbf{-0.59} & \mathbf{0.65} \\ 0.07 & \mathbf{-0.73} & \mathbf{-0.67} \\ 0.07 & \mathbf{-0.29} & \mathbf{0.32} \end{bmatrix} \times \begin{bmatrix} \mathbf{12.4} & 0 & 0 \\ 0 & \mathbf{9.5} & 0 \\ 0 & 0 & \cancel{\mathbf{1.3}} \end{bmatrix} \times \begin{bmatrix} \mathbf{0.56} & \mathbf{0.59} & \mathbf{0.56} & 0.09 & 0.09 \\ 0.12 & -0.02 & 0.12 & \mathbf{-0.69} & \mathbf{-0.69} \\ 0.40 & \mathbf{-0.80} & 0.40 & 0.09 & 0.09 \end{bmatrix}$$

J. Leskovec, A. Rajaraman, J. Ullman: Mining of Massive Datasets, http://www.mmds.org

SVD: Redução de dimensionalidade

$$\begin{bmatrix} \mathbf{1} & \mathbf{1} & \mathbf{1} & 0 & 0 \\ \mathbf{3} & \mathbf{3} & \mathbf{3} & 0 & 0 \\ \mathbf{4} & \mathbf{4} & \mathbf{4} & 0 & 0 \\ \mathbf{5} & \mathbf{5} & \mathbf{5} & 0 & 0 \\ 0 & \mathbf{2} & 0 & \mathbf{4} & \mathbf{4} \\ 0 & 0 & 0 & \mathbf{5} & \mathbf{5} \\ 0 & \mathbf{1} & 0 & \mathbf{2} & \mathbf{2} \end{bmatrix} \approx \begin{bmatrix} \mathbf{0.13} & 0.02 & \cancel{-0.01} \\ \mathbf{0.41} & 0.07 & \cancel{-0.03} \\ \mathbf{0.55} & 0.09 & \cancel{-0.04} \\ \mathbf{0.68} & 0.11 & \cancel{-0.05} \\ 0.15 & \mathbf{-0.59} & \cancel{\mathbf{0.65}} \\ 0.07 & \mathbf{-0.73} & \cancel{\mathbf{-0.67}} \\ 0.07 & \mathbf{-0.29} & \cancel{\mathbf{0.32}} \end{bmatrix} \times \begin{bmatrix} \mathbf{12.4} & 0 & 0 \\ 0 & \mathbf{9.5} & 0 \\ 0 & 0 & \cancel{\mathbf{1.3}} \end{bmatrix} \times \begin{bmatrix} \mathbf{0.56} & \mathbf{0.59} & \mathbf{0.56} & 0.09 & 0.09 \\ 0.12 & -0.02 & 0.12 & \mathbf{-0.69} & \mathbf{-0.69} \\ \cancel{0.40} & \cancel{\mathbf{-0.80}} & \cancel{0.40} & \cancel{0.09} & \cancel{0.09} \end{bmatrix}$$

J. Leskovec, A. Rajaraman, J. Ullman: Mining of Massive Datasets, http://www.mmds.org

SVD: Redução de dimensionalidade

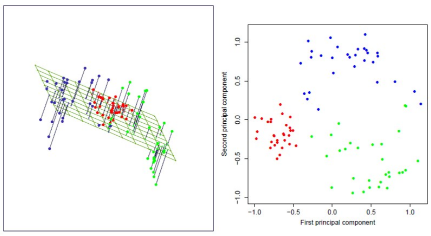
$$
\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 3 & 3 & 3 & 0 & 0 \\ 4 & 4 & 4 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 2 & 0 & 4 & 4 \\ 0 & 0 & 0 & 5 & 5 \\ 0 & 1 & 0 & 2 & 2 \end{bmatrix} \approx \begin{bmatrix} 0.13 & 0.02 \\ 0.41 & 0.07 \\ 0.55 & 0.09 \\ 0.68 & 0.11 \\ 0.15 & -0.59 \\ 0.07 & -0.73 \\ 0.07 & -0.29 \end{bmatrix} \times \begin{bmatrix} 12.4 & 0 \\ 0 & 9.5 \end{bmatrix} \times \begin{bmatrix} 0.56 & 0.59 & 0.56 & 0.09 & 0.09 \\ 0.12 & -0.02 & 0.12 & -0.69 & -0.69 \end{bmatrix}
$$

SVD: Redução de dimensionalidade

$$
\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 3 & 3 & 3 & 0 & 0 \\ 4 & 4 & 4 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 2 & 0 & 4 & 4 \\ 0 & 0 & 0 & 5 & 5 \\ 0 & 1 & 0 & 2 & 2 \end{bmatrix} \approx \begin{bmatrix} 0.92 & 0.95 & 0.92 & 0.01 & 0.01 \\ 2.91 & 3.01 & 2.91 & -0.01 & -0.01 \\ 3.90 & 4.04 & 3.90 & 0.01 & 0.01 \\ 4.82 & 5.00 & 4.82 & 0.03 & 0.03 \\ 0.70 & 0.53 & 0.70 & 4.11 & 4.11 \\ -0.69 & 1.34 & -0.69 & 4.78 & 4.78 \\ 0.32 & 0.23 & 0.32 & 2.01 & 2.01 \end{bmatrix}
$$

SVD: Construção

- Qual o conjunto de k direções que melhor *'aproxima/representa'* um conjunto de n pontos em R^d ?

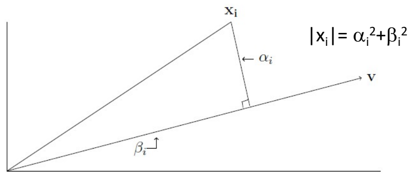


SVD: Construção

- Qual o conjunto de k direções que melhor *'aproxima/representa'* um conjunto de n pontos em R^d ?
- Qualidade da aproximação
  - Soma do quadrado das distâncias dos pontos ao hiperplano gerado pelas direções (deve ser minimizada)
  - Soma dos quadrados das projeções no subespaço gerado pelas direções (deve ser maximizada)

SVD: Construção

- Com qual qualidade o vetor unitário **v** representa um ponto $\mathbf{x}_i$ em R^d ?



$|x_i| = \alpha_i^2+\beta_i^2$

- Maximizar o quadrado da norma da projeção é equivalente a minimizar o quadrado da distância perpendicular a **v**

SVD: Construção

Qual conjunto de **k** direções (vetores unitários) que melhor representa A?

SVD: Construção

Método guloso (número de direções fixas)

1. Encontrar vetor unitário v_1 tal que $|Av_1|$ é maximizado
2. Enquanto i<k definir v_{i+1} como o vetor unitário ortogonal a $v_1,\ldots,v_i$ tal que

$$|Av_{i+1}| \geq |Av|$$

para todo v ortogonal a $v_1,\ldots,v_i$

SVD: Construção

Método guloso (todas as direções)

1. Encontrar vetor unitário v_1 tal que $|Av_1|$ é maximizado
2. Enquanto $|Av_i| > 0$ definir v_{i+1} como o vetor unitário ortogonal a $v_1,\ldots,v_i$ tal que

$$|Av_{i+1}| \geq |Av|$$

para todo v ortogonal a $v_1,\ldots,v_i$. Incrementar i

3. Definir r como i

SVD: Construção

Propriedade. Ao término do procedimento todos o n pontos da matriz A pertencem ao subespaço V_r gerado pelas direções $v_1,\dots, v_r$

Prova. Os pontos **a** que não pertencem a V_r podem ser escritos como

$$\mathbf{a}=v_a+v',$$

onde v_a pertence a V_r e v' é ortogonal a V_r

Neste caso **a**v'> 0 para todo ponto **a** que não está em V_r. Logo, teríamos |Av'|>0. Portanto, o algoritmo não teria parado

SVD: Construção

Processo guloso leva a melhor representação?

SVD: Construção

Definição. v_i é i-ésimo vetor singular a direita da matriz A

SVD: Construção

Teorema. Seja A uma matriz nxd com vetores singulares $v_1,\dots, v_r$. Para $1 \leq k \leq r$, o espaço V_k gerado por $v_1,\dots, v_k$ é o espaço k-dimensional que maximiza a soma dos quadrados das projeções dos n pontos

Prova

- Para k=1, segue da definição do método.
- Passo de Indução: Se V_k é ótimo para dim k então V_{k+1} é ótimo para dim k+1

SVD: Construção

Prova (cont'd)

Seja W um subespaço ótimo de dim k+1 e seja w_{k+1} um vetor de w ortogonal a V_k

Seja $w_1,\ldots, w_{k+1}$ uma base ortonormal para W tal que w_{k+1} é ortogonal a V_k

SVD: Construção

Prova (cont'd)

Por hipótese

$$|Aw_1|^2+\ldots+|Aw_k|^2 \leq |Av_1|^2+\ldots+|Av_k|^2$$

Além disso, por construção

$$|Aw_{k+1}|^2 \leq |Av_{k+1}|^2$$

Somando as desigualdades provamos o resultado

SVD: Construção

Observação. Av_i é a lista dos comprimentos das projeções dos n pontos na direção v_i

Definição . O i-ésimo valor singular da matriz A é $\sigma_i=|Av_i|$

SVD: Construção

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 3 & 3 & 3 & 0 & 0 \\ 4 & 4 & 4 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 0 & 0 & 4 & 4 \\ 0 & 0 & 0 & 5 & 5 \\ 0 & 0 & 0 & 2 & 2 \end{bmatrix} = \begin{bmatrix} .14 & 0 \\ .42 & 0 \\ .56 & 0 \\ .70 & 0 \\ 0 & .60 \\ 0 & .75 \\ 0 & .30 \end{bmatrix} \begin{array}{c} \sigma_1 \\ \begin{bmatrix} \boxed{12.4} & 0 \\ 0 & 9.5 \end{bmatrix} \end{array} \begin{array}{c} \mathbf{v_1} \\ \begin{bmatrix} \boxed{.58 \quad .58 \quad .58 \quad 0 \quad 0} \\ 0 \quad 0 \quad 0 \quad .71 \quad .71 \end{bmatrix} \end{array}$$

Exemplo

- Se v_1 corresponde a Sci-Fi então
  - $\sigma_1=|Av_1|$ é o peso de Sci-Fi

SVD: Construção

- 5 pontos em R^5

$$A = \begin{bmatrix} 2 & 0 & 8 & 6 & 0 \\ 1 & 6 & 0 & 1 & 7 \\ 5 & 0 & 7 & 4 & 0 \\ 7 & 0 & 8 & 5 & 0 \\ 0 & 10 & 0 & 0 & 7 \end{bmatrix}$$

- Os valores singulares da matriz acima são 17.92, 15.17, 3.56, 1.98, 0.35

SVD: Construção

- σ_i pode ser vista como a componente de A na direção v_i
- Para essa interpretação fazer sentido deve ser possível recuperar o conteúdo da matriz A a partir dos valores singulares

SVD

Propriedade. Seja a_i é a i-ésima linha da matriz A. Então

$$\sum_{j=1}^{n} |\mathrm{a_j}|^2 = \sum_{j=1}^{n}\sum_{i=1}^{r} (\mathrm{a_j} \cdot \mathrm{v_i})^2 = \sum_{i=1}^{r}\sum_{j=1}^{n} (\mathrm{a_j} \cdot \mathrm{v_i})^2 = \sum_{i=1}^{r} |A\mathrm{v_i}|^2 = \sum_{i=1}^{r} \sigma_i^2(A).$$

- A primeira igualdade abaixo decorre de que todo vetor a_i pode ser escrita como uma combinação linear dos vetores ortonormais $v_1,\ldots,v_r$
- A terceira e quarta igualdade seguem das definições

SVD

- σ_i dá uma idéia da variância capturada pela direção v_i

$$A = \begin{bmatrix} 2 & 0 & 8 & 6 & 0 \\ 1 & 6 & 0 & 1 & 7 \\ 5 & 0 & 7 & 4 & 0 \\ 7 & 0 & 8 & 5 & 0 \\ 0 & 10 & 0 & 0 & 7 \end{bmatrix}$$

- Os valores singulares da matriz acima são 17.92, 15.17, 3.56, 1.98, 0.35
- Os dois componentes principais já explicam 97% da variância: $(17.92^2+15.17^2)/564$

SVD

$$
\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 3 & 3 & 3 & 0 & 0 \\ 4 & 4 & 4 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 2 & 0 & 4 & 4 \\ 0 & 0 & 0 & 5 & 5 \\ 0 & 1 & 0 & 2 & 2 \end{bmatrix} = \begin{bmatrix} 0.13 & 0.02 & -0.01 \\ 0.41 & 0.07 & -0.03 \\ 0.55 & 0.09 & -0.04 \\ 0.68 & 0.11 & -0.05 \\ 0.15 & -0.59 & 0.65 \\ 0.07 & -0.73 & -0.67 \\ 0.07 & -0.29 & 0.32 \end{bmatrix} \times \begin{bmatrix} 12.4 & 0 & 0 \\ 0 & 9.5 & 0 \\ 0 & 0 & 1.3 \end{bmatrix} \times \begin{bmatrix} 0.56 & 0.59 & 0.56 & 0.09 & 0.09 \\ 0.12 & -0.02 & 0.12 & -0.69 & -0.69 \\ 0.40 & -0.80 & 0.40 & 0.09 & 0.09 \end{bmatrix}
$$

- Os valores singulares da a esquerda são 12.4, 9.5 e 1.3.
- O primeiro valor corresponde a 62.5% da variância
- Os dois primeiros correspondem a 99.3% da variância

SVD

Definição. Seja $u_i = Av_i / |Av_i|$. O conjunto de vetores $u_1, \ldots, u_r$ são os vetores singulares a esquerda da matriz A

SVD: Construção

$$
\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 3 & 3 & 3 & 0 & 0 \\ 4 & 4 & 4 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 0 & 0 & 4 & 4 \\ 0 & 0 & 0 & 5 & 5 \\ 0 & 0 & 0 & 2 & 2 \end{bmatrix} = \begin{bmatrix} \boxed{\begin{matrix} .14 \\ .42 \\ .56 \\ .70 \\ 0 \\ 0 \\ 0 \end{matrix}} & \begin{matrix} 0 \\ 0 \\ 0 \\ 0 \\ .60 \\ .75 \\ .30 \end{matrix} \end{bmatrix} \begin{bmatrix} \boxed{12.4} & 0 \\ 0 & 9.5 \end{bmatrix} \begin{bmatrix} \boxed{.58 \quad .58 \quad .58 \quad 0 \quad 0} \\ 0 \quad 0 \quad 0 \quad .71 \quad .71 \end{bmatrix}
$$

u_1 v_1

Exemplo

- Se v_1 corresponde a Sci-Fi então
  - u_1: são os valores dos usuários em relação a Sci-Fi

SVD

Lema. Os vetores $u_1, \ldots, u_r$ formam uma base ortonormal

SVD

Teorema. $A=UDV^T$

A: matriz de dados com n linhas e d colunas

U: matriz com n linhas e r colunas em que as colunas são os vetores singulares a esquerda,

D: matriz diagonal de dimensão (rxr) formada pelos valores singulares

V: é a matriz com d linhas e r colunas em que as colunas são os vetores singulares a direita

SVD

Lema auxiliar. Se A**v**=B**v** para todo vetor **v** então A=B

SVD

Teorema. $A=UDV^T$

Prova.

Por causa do lema, basta mostrar que Av =(UDV^T)v para todo vetor v. Para uma direção principal v_j

$$\left(\sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^T\right) \mathbf{v}_j = \sigma_j \mathbf{u}_j = A\mathbf{v}_j$$

A primeira igualdade segue da ortonomalidade dos vetores v_j e a segunda da definição de u_j

SVD

Teorema. $A=UDV^T$

Prova (cont'd).

Se **v'** é ortogonal a V_r então

A **v'** = (UDV^T) **v'** = 0

Como qualquer vetor **v** pode ser escrito como combinação linear dos vetores em V_r somado a um vetor ortogonal a V_r o resultado segue.

Best k rank approximation

Seja

$$A_k = \sum_{i=1}^{k} \sigma_i \mathbf{u_i} \mathbf{v_i}^T$$

a aproximação para matriz A dada pelas k primeiras direções (vetores singulares)

Best k rank approximation

Lema (3.5). As linhas de A_k são as projeções dos pontos (linhas de A) no subespaço V_k gerado pelas k direções principais $v_1,..,v_k$

Prova. Como os v_i's formam uma base ortonormal, a projeção do ponto **a** em V_k é dada por

$$\sum_{i=1}^{k} (\mathbf{a} \cdot \mathbf{v_i}) \mathbf{v_i}^T,$$

Portanto, a projeção das linhas da matriz A é dada por

$$\sum_{i=1}^{k} A\mathbf{v_i}\mathbf{v_i}^T = \sum_{i=1}^{k} \sigma_i \mathbf{u_i}\mathbf{v_i}^T = A_k.$$

Best k rank approximation

Definição: Para uma matriz C a norma $||C||_F$ (norma de Frobenius) é a soma dos quadrados das entradas da matriz.

Best k rank approximation

Teorema. Para qualquer matriz B de rank no máximo k temos

$$||A-A_k||_F \leq ||A - B||_F$$

Prova (sketch)

- Seja B uma matriz que minimiza $||A - B||_F$
- Seja V o subespaço gerado pelas linhas de B. Temos que cada linha de B é a projeção de A em V, caso contrário trocando as linhas de B pelas projeções das linhas de A em V diminuíriamos $||A - B||_F$.

Best k rank approximation

Teorema. Para qualquer matriz B de rank no máximo k temos

$||A-A_k||_F \le || A - B||_F$

Prova (sketch)

Segue do Teorema (análise do guloso) que

$||A_k||_F \ge ||B||_F$.

Como

$||A-A_k||_F + ||A_k||_F = ||A|_F$

e

$||A-B||_F + ||B||_F = ||A||_F$

temos que

$||A-A_k||_F \le ||A - B||_F$

Best k rank approximation

Interpretação/Consequência

- A matriz A_k provê a melhor representação possível dos n pontos em termos de erro quadrático.
- Se tentarmos representar os n pontos de A em um subespaço de dim ≤k o erro quadrático será maior ou igual ao de A_k

SVD: Método de potências

Como calcular a decomposição SVD?

SVD: Método de potências

Seja $B=A^TA$. Temos que

$$B = A^T A = \left(\sum_i \sigma_i \mathbf{v_i} \mathbf{u_i}^T\right)\left(\sum_j \sigma_j \mathbf{u_j} \mathbf{v_j}^T\right)$$
$$= \sum_{i,j} \sigma_i \sigma_j \mathbf{v_i}(\mathbf{u}_i^T \cdot \mathbf{u_j})\mathbf{v_j}^T = \sum_i \sigma_i^2 \mathbf{v_i} \mathbf{v_i}^T.$$

SVD: Método de potências

Seja $B=A^TA$. Possível mostrar que

$$B^k = \sum_{i=1}^{r} \sigma_i^{2k} \mathrm{v_i v_i}^T.$$

Note que os valores singulares de $B^{k/2}$ são

$\sigma_1^k, \ldots, \sigma_r^k$

Portanto,

$$||B^{k/2}||_F = \sigma_1^{2k} + \ldots + \sigma_r^{2k}$$

SVD: Método de potências

Note que por construção $\sigma_1 \geq \sigma_2 \geq \ldots.\ \sigma_r > 0$
Se $\sigma_1 > \sigma_2$

$$\frac{1}{\sigma_1^{2k}} B^k \to \mathrm{v_1 v_1}^T.$$

e

$$\frac{B^k}{||B^{k/2}||_F} \to v_1\ v_1^T$$

Para encontrar v_1 basta avaliar B^k, para k 'grande' e normalizar uma linha da matriz

$$\frac{B^k}{||B^{k/2}||_F}$$

SVD: Método de potências

Complexidade

- Convergência do método depende de $\sigma_1 - \sigma_2$
- Método envolve multiplicação de matrizes, pode ser caro computacionalmente

  Exemplo
  - A tem tamanho $10^8 \times 10^8$ com 10^9 entradas não nulas (esparsa)
  - AxA tem tamanho 10^{16} e não é esparsa. Inviável!

SVD: Método de potências

1. Seja $(v_1, v_2, \ldots, v_d)$ uma base ortonormal para R^d obtida estendendo a base $v_1, v_2, \ldots, v_r$ formada pelos vetores singulares. Além disso, seja $x = c_1 v_1 + c_2 v_2 + \ldots. + c_d v_d$ com c_1 <>0
2. Se $\sigma_1 > \sigma_2$ então $B^k x$ converge para $\sigma_1^{2k} c_1 v_1$
3. Como sabemos que v_1 tem norma 1, podemos obter v_1 normalizando $\sigma_1^{2k} c_1 v_1$
4. Para calcular $B^k x$ fazemos $Bx = A^T(Ax)$ e $B^k x = A^T(AB^{k-1}x)$
5. Multiplicação entre matrizes são evitadas
   - Apenas multiplicamos matrizes por vetores O(nd k)

SVD: Método de potências

Entrada: Matriz A de dimensões nxd
Saída: vetor de dimensão d

1. Escolha um vetor aleatório **x** com norma 1 utilizando uma distribuição gaussiana
2. Repita

   $x_{old} \leftarrow x$
   $x \leftarrow A^T(Ax)$
   $y \leftarrow$ x normalizado ; $y_{old} \leftarrow x_{old}$ normalizado

   Até que $|y-y_{old}| <$ err
3. Devolva y

SVD: Método de potências

Propriedades do método

1. Se o maior valor singular for maior que o segundo maior singular de A então o método converge para v_1
2. Se o maior valor singular for igual ao segundo maior valor singular o método converge para um vetor no espaço de gerado pelas direções principais associadas aos 'maiores' valores singulares (Teorema 3.11 e Lema 3.12 de FDS)

SVD: Método de potências

1. Método apresentado apenas permite obter v_1
2. Para obter v_{k+1} dado que já obtemos $v_1, v_2, \ldots, v_k$

utilizar o método anterior em A - A_k ,onde

$$A_k = \sum_{i=1}^{k} \sigma_i \mathbf{u_i} \mathbf{v_i}^T$$