Singular Value Decomposition (SVD)/Singulärwertzerlegung

Katja Markert, einige Folien von Julian Hitschler

Institut für Computerlinguistik
Uni Heidelberg
markert@cl.uni-heidelberg.de

June 11, 2019

Bisher und heute

1. Hintergrund zu den mathematischen Begriffen und Methoden, die für Matrizenzerlegung notwendig ist
2. Jetzt: der SVD-Hauptsatz (mit Konstruktion, aber ohne Beweis)
3. Die Intuition dahinter
4. Reduced SVD $\rightarrow$ Dichte Embeddings
5. Performanz von diesen dichten SVD-Embeddings in NLP

Übersicht

Übersicht

SVD

- Eine reellwertige $m \times n$ - Matrix A lässt sich wie folgt faktorisieren:

$$A = U \Sigma V^T$$

- U ist eine orthonormale $m \times m$ - Matrix
- Σ ist eine “diagonale” $m \times n$ - Matrix
- V^T ist die Transponierte einer orthonormalen $n \times n$ - Matrix

Singulärwertzerlegung

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix} =$$

$$\begin{pmatrix} u_{11} & u_{12} & \dots & u_{1m} \\ u_{21} & u_{22} & \dots & u_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ u_{m1} & u_{m2} & \dots & u_{mm} \end{pmatrix} \begin{pmatrix} \sigma_{11} & 0 & 0 & \dots & 0 \\ 0 & \sigma_{22} & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \sigma_{mn} \end{pmatrix} \begin{pmatrix} v_{11} & v_{12} & \dots & v_{1n} \\ v_{21} & v_{22} & \dots & v_{2n} \\ v_{31} & v_{32} & \dots & v_{3n} \\ \vdots & \vdots & \ddots & \vdots \\ v_{n1} & v_{n2} & \dots & v_{nn} \end{pmatrix}^T$$

- Diagonale Einträge in Σ werden als Singulärwerte von A bezeichnet
- Spalten in U werden als linke Singulärvektoren von A bezeichnet
- Spalten in V werden als rechte Singulärvektoren von A bezeichnet

Konstruktionsmethodologie I

1. Berechne die quadratische $m \times m$ Matrix $A \cdot A^T$
2. Berechne die Eigenwerte σ_i^2 und Eigenvektoren u_i von $A \cdot A^T$
3. Die Vektoren u_i werden orthonormalisiert (Gram-Schmidt) und bilden die Spaltenvektoren von U
4. Berechne die quadratische $n \times n$ Matrix $A^T \cdot A$
5. Berechne die Eigenwerte σ_j^2 und Eigenvektoren v_j von $A^T \cdot A$. Tip: die positiven Eigenwerte sind gleich zu denen von $A \cdot A^T$
6. Die Vektoren v_j werden orthonormalisiert (Gram-Schmidt) und bilden die Spaltenvektoren von V
7. Die Wurzeln der σ^2-Werte werden in absteigender Reihenfolge die Einträge von Σ
8. Eventuell Korrektur der Vorzeichen von U bzw. V

Konstruktionsmethodologie I

1. Berechne die quadratische $m \times m$ Matrix $A \cdot A^T$
2. Berechne die Eigenwerte σ_i^2 und Eigenvektoren u_i von $A \cdot A^T$
3. Die Vektoren u_i werden orthonormalisiert (Gram-Schmidt) und bilden die Spaltenvektoren von U
4. Berechne die quadratische $n \times n$ Matrix $A^T \cdot A$
5. Berechne die Eigenwerte σ_j^2 und Eigenvektoren v_j von $A^T \cdot A$. Tip: die positiven Eigenwerte sind gleich zu denen von $A \cdot A^T$
6. Die Vektoren v_j werden orthonormalisiert (Gram-Schmidt) und bilden die Spaltenvektoren von V
7. Die Wurzeln der σ^2-Werte werden in absteigender Reihenfolge die Einträge von Σ
8. Eventuell Korrektur der Vorzeichen von U bzw. V

Konstruktionsmethodologie I

1. Berechne die quadratische $m \times m$ Matrix $A \cdot A^T$
2. Berechne die Eigenwerte σ_i^2 und Eigenvektoren u_i von $A \cdot A^T$
3. Die Vektoren u_i werden orthonormalisiert (Gram-Schmidt) und bilden die Spaltenvektoren von U
4. Berechne die quadratische $n \times n$ Matrix $A^T \cdot A$
5. Berechne die Eigenwerte σ_j^2 und Eigenvektoren v_j von $A^T \cdot A$. Tip: die positiven Eigenwerte sind gleich zu denen von $A \cdot A^T$
6. Die Vektoren v_j werden orthonormalisiert (Gram-Schmidt) und bilden die Spaltenvektoren von V
7. Die Wurzeln der σ^2-Werte werden in absteigender Reihenfolge die Einträge von Σ
8. Eventuell Korrektur der Vorzeichen von U bzw. V

Konstruktionsmethodologie I

1. Berechne die quadratische $m \times m$ Matrix $A \cdot A^T$
2. Berechne die Eigenwerte σ_i^2 und Eigenvektoren u_i von $A \cdot A^T$
3. Die Vektoren u_i werden orthonormalisiert (Gram-Schmidt) und bilden die Spaltenvektoren von U
4. Berechne die quadratische $n \times n$ Matrix $A^T \cdot A$
5. Berechne die Eigenwerte σ_j^2 und Eigenvektoren v_j von $A^T \cdot A$. Tip: die positiven Eigenwerte sind gleich zu denen von $A \cdot A^T$
6. Die Vektoren v_j werden orthonormalisiert (Gram-Schmidt) und bilden die Spaltenvektoren von V
7. Die Wurzeln der σ^2-Werte werden in absteigender Reihenfolge die Einträge von Σ
8. Eventuell Korrektur der Vorzeichen von U bzw. V

Beispiel (aus Kirk Bakers Tutorial)

Sei

$$A = \begin{pmatrix} 3 & 1 & 1 \\ -1 & 3 & 1 \end{pmatrix}$$

Schritt 1: Berechne $A \cdot A^T$

$$A \cdot A^T = \begin{pmatrix} 3 & 1 & 1 \\ -1 & 3 & 1 \end{pmatrix} \begin{pmatrix} 3 & -1 \\ 1 & 3 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} 11 & 1 \\ 1 & 11 \end{pmatrix}$$

Schritt II: Bestimme die Eigenwerte und Eigenvektoren von $A \cdot A^T$

Setze

$$\begin{pmatrix} 11 & 1 \\ 1 & 11 \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \sigma \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

Damit ergibt sich das Gleichungssystem:

1. $11x_1 + x_2 = \sigma x_1$
2. $x_1 + 11x_2 = \sigma x_2$

Schritt II: Bestimme die Eigenwerte und Eigenvektoren von $A \cdot A^T$

Setze

$$\begin{pmatrix} 11 & 1 \\ 1 & 11 \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \sigma \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

Damit ergibt sich das Gleichungssystem:

1. $11x_1 + x_2 = \sigma x_1$
2. $x_1 + 11x_2 = \sigma x_2$

Schritt II: Bestimme die Eigenwerte und Eigenvektoren von $A \cdot A^T$

1. $11x_1 + x_2 = \sigma x_1$
2. $x_1 + 11x_2 = \sigma x_2$

Ergibt z. B. nach Umformungen (hier Zeilenabzug)

$$10x_1 - 10x_2 = \sigma x_1 - \sigma x_2$$

und damit Eigenwert $\sigma_1 = 10$
Ergibt z. B. nach Umformungen (hier Zeilenaddition)

$$12x_1 + 12x_2 = \sigma x_1 + \sigma x_2$$

und damit Eigenwert $\sigma_2 = 12$

Schritt II: Bestimme die Eigenwerte und Eigenvektoren von $A \cdot A^T$

1. $11x_1 + x_2 = \sigma x_1$
2. $x_1 + 11x_2 = \sigma x_2$

Zwei Eigenwerte $\sigma_1 = 10$ und $\sigma_2 = 12$. Damit ergibt sich für die Eigenvektoren:

1. $11x_1 + x_2 = 10x_1$
2. $x_1 + 11x_2 = 10x_2$ und damit
3. $x_1 = -x_2$
4. Wir wählen $x_1 = 1$ und damit $x_2 = -1$

sowie

1. $11x_1 + x_2 = 12x_1$
2. $x_1 + 11x_2 = 12x_2$ und damit
3. $x_1 = x_2$
4. Wir wählen $x_1 = 1$ und damit $x_2 = 1$

Schritt III: Bestimme U

Starte mit einer Matrix $\tilde{U}$, die die bestimmten Eigenvektoren von $A \cdot A^T$ geordnet nach der Größe der jeweiligen Eigenwerte enthält

1. Für Eigenwert $\sigma_1 = 10$ hatten wir $x_1 = 1$ und damit $x_2 = -1$ gewählt
2. Für Eigenwert $\sigma_1 = 12$ hatten wir $x_1 = 1$ und damit $x_2 = 1$ gewählt
3. Spaltenvektoren nach Eigenwertgröße geordnet

$$\tilde{U} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$$

Schritt III: Bestimme U

Orthonormalisiere mit Gram-Schmidt die Spaltenvektoren von $\tilde{U}$

$$\tilde{U} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$$

1. Normalisiere ersten Spaltenvektor $u^{(1)} = \frac{\tilde{u}^{(1)}}{\|\tilde{u}^{(1)}\|} = \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}$
2. Berechne die Senkrechte $\tilde{u}^{(2)} - \langle u^{(1)}, \tilde{u}^{(2)} \rangle \cdot u^{(1)}$
3. Also $\begin{pmatrix} 1 \\ -1 \end{pmatrix} - \langle \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}, \begin{pmatrix} 1 \\ -1 \end{pmatrix} \rangle \cdot \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix} = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$
4. Normalisiere die Senkrechte $\begin{pmatrix} 1 \\ -1 \end{pmatrix}$ ergibt $u^{(2)} = \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{-1}{\sqrt{2}} \end{pmatrix}$

Schritt III: Bestimme U

Damit ist

$$U = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \end{pmatrix}$$

Schritt IV: Berechne $A^T \cdot A$

$$A^T \cdot A = \begin{pmatrix} 3 & -1 \\ 1 & 3 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} 3 & 1 & 1 \\ -1 & 3 & 1 \end{pmatrix} = \begin{pmatrix} 10 & 0 & 2 \\ 0 & 10 & 4 \\ 2 & 4 & 2 \end{pmatrix}$$

Schritt V: Bestimme die Eigenwerte und Eigenvektoren von $A^T \cdot A$

Setze

$$\begin{pmatrix} 10 & 0 & 2 \\ 0 & 10 & 4 \\ 2 & 4 & 2 \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \sigma \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$$

Damit ergibt sich das Gleichungssystem:

1. $10x_1 + 2x_3 = \sigma x_1$
2. $10x_2 + 4x_3 = \sigma x_2$
3. $2x_1 + 4x_2 + 2x_3 = \sigma x_3$

Schritt V: Bestimme die Eigenwerte und Eigenvektoren von $A^T \cdot A$

Setze

$$\begin{pmatrix} 10 & 0 & 2 \\ 0 & 10 & 4 \\ 2 & 4 & 2 \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \sigma \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$$

Damit ergibt sich das Gleichungssystem:

1. $10x_1 + 2x_3 = \sigma x_1$
2. $10x_2 + 4x_3 = \sigma x_2$
3. $2x_1 + 4x_2 + 2x_3 = \sigma x_3$

Schritt V: Bestimme die Eigenwerte und Eigenvektoren von $A^T \cdot A$

1. $10x_1 + 2x_3 = \sigma x_1$
2. $10x_2 + 4x_3 = \sigma x_2$
3. $2x_1 + 4x_2 + 2x_3 = \sigma x_3$

Gleichung (ii) minus zweimal (i):

$$10x_2 + 4x_3 - 20x_1 - 4x_3 = \sigma x_2 - 2\sigma x_2$$

und damit $\sigma_1 = 10$

Mit ähnlichen Umformungen kann man $\sigma_2 = 12$ sowie $\sigma_3 = 0$ berechnen.

Schritt V: Bestimme die Eigenwerte und Eigenvektoren von $A^T \cdot A$

Drei Eigenwerte $\sigma_1 = 10$, $\sigma_2 = 12$ und $\sigma_3 = 0$. Damit ergibt sich für die Eigenvektoren:

1. $10x_1 + 2x_3 = 10x_1$
2. $10x_2 + 4x_3 = 10x_2$
3. $2x_1 + 4x_2 + 2x_3 = 10x_3$ und damit
4. $x_3 = 0$ aus der ersten Gleichung und damit
5. $2x_1 + 4x_2 = 0$ bzw $x_1 = -2x_2$
6. Wir können die Länge des Eigenvektors wieder frei wählen, da wir danach eh normalisieren. Allerdings sind die Vorzeichen nicht mehr frei wählbar, da ja insgesamt bei $U \cdot \Sigma \cdot V^T$ wieder A herauskommen muss. Darauf macht das Tutorial nicht aufmerksam

   Wir wählen $x_1 = 2, x_2 = -1$, hätten aber auch (fälschlicherweise)$x_1 = -2, x_2 = 1$ wählen können.

Ebenso bestimmen wir für $\sigma_2 = 12$ einen Eigenvektor $(1, 2, 1)$ und für $\sigma_3 = 0$ einen Eigenvektor $(1, 2, -5)$

Schritt VI: Bestimme V

Starte mit einer Matrix $\tilde{V}$, die die bestimmten Eigenvektoren von $A^T \cdot A$ nach der Größe der jeweiligen Eigenwerte als Spaltenvektoren enthält

1. Für Eigenwert $\sigma_1 = 10$ hatten wir $(2, -1, 0)$ gewählt
2. Für Eigenwert $\sigma_2 = 12$ hatten wir $(1, 2, 1)$ bestimmt
3. Für Eigenwert $\sigma_3 = 0$ hatten wir $(1, 2, -5)$ bestimmt
4. Spaltenvektoren nach Eigenwertgröße geordnet

$$\tilde{V} = \begin{pmatrix} 1 & 2 & 1 \\ 2 & -1 & 2 \\ 1 & 0 & -5 \end{pmatrix}$$

Schritt VI: Bestimme V

Orthonormalisiere mit Gram-Schmidt die Spaltenvektoren von $\tilde{V}$

$$\tilde{V} = \begin{pmatrix} 1 & 2 & 1 \\ 2 & -1 & 2 \\ 1 & 0 & -5 \end{pmatrix}$$

1. Normalisiere ersten Spaltenvektor $v^{(1)} = \frac{\tilde{v}^{(1)}}{\|\tilde{v}^{(1)}\|} = \begin{pmatrix} \frac{1}{\sqrt{6}} \\ \frac{2}{\sqrt{6}} \\ \frac{1}{\sqrt{6}} \end{pmatrix}$
2. Berechne die Senkrechte $\tilde{v}^{(2)} - \langle v^{(1)}, \tilde{v}^{(2)} \rangle \cdot v^{(1)}$
3. Also $\begin{pmatrix} 2 \\ -1 \\ 0 \end{pmatrix} - \left\langle \begin{pmatrix} \frac{1}{\sqrt{6}} \\ \frac{2}{\sqrt{6}} \\ \frac{1}{\sqrt{6}} \end{pmatrix}, \begin{pmatrix} 2 \\ -1 \\ 0 \end{pmatrix} \right\rangle \cdot \begin{pmatrix} \frac{1}{\sqrt{6}} \\ \frac{2}{\sqrt{6}} \\ \frac{1}{\sqrt{6}} \end{pmatrix} = \begin{pmatrix} 2 \\ -1 \\ 0 \end{pmatrix}$
4. Normalisiere die Senkrechte $\begin{pmatrix} 2 \\ -1 \\ 0 \end{pmatrix}$ ergibt $v^{(2)} = \begin{pmatrix} \frac{2}{\sqrt{5}} \\ \frac{-1}{\sqrt{5}} \\ 0 \end{pmatrix}$

Schritt VI: Bestimme V

$$\tilde{V} = \begin{pmatrix} 1 & 2 & 1 \\ 2 & -1 & 2 \\ 1 & 0 & -5 \end{pmatrix}$$

Schon bestimmt

$$v^{(1)} = \begin{pmatrix} \frac{1}{\sqrt{6}} \\ \frac{2}{\sqrt{6}} \\ \frac{1}{\sqrt{6}} \end{pmatrix} \text{ sowie } v^{(2)} = \begin{pmatrix} \frac{2}{\sqrt{5}} \\ \frac{-1}{\sqrt{5}} \\ 0 \end{pmatrix}$$

Dann kann man auch noch $v^{(3)}$ berechnen mit der Senkrechten

$$\tilde{v}^{(3)} - \langle v^{(1)}, \tilde{v}^{(3)} \rangle \cdot v^{(1)} - \langle v^{(2)}, \tilde{v}^{(3)} \rangle \cdot v^{(2)}$$

Dies ergibt $\begin{pmatrix} \frac{-2}{3} \\ \frac{-4}{3} \\ \frac{10}{3} \end{pmatrix}$ *Normalisiert* : $v^{(3)} = \begin{pmatrix} \frac{1}{\sqrt{30}} \\ \frac{2}{\sqrt{30}} \\ \frac{-5}{\sqrt{30}} \end{pmatrix}$

Schritt VI: Bestimme V

Damit

$$V = \begin{pmatrix} \frac{1}{\sqrt{6}} & \frac{2}{\sqrt{5}} & \frac{1}{\sqrt{30}} \\ \frac{2}{\sqrt{6}} & \frac{-1}{\sqrt{5}} & \frac{2}{\sqrt{30}} \\ \frac{1}{\sqrt{6}} & 0 & \frac{-5}{\sqrt{30}} \end{pmatrix}$$

und damit

$$V^T = \begin{pmatrix} \frac{1}{\sqrt{6}} & \frac{2}{\sqrt{6}} & \frac{1}{\sqrt{6}} \\ \frac{2}{\sqrt{5}} & \frac{-1}{\sqrt{5}} & 0 \\ \frac{1}{\sqrt{30}} & \frac{2}{\sqrt{30}} & \frac{-5}{\sqrt{30}} \end{pmatrix}$$

Schritt VII: Bestimme Σ

Σ ist eine $m \times n$ (hier 2×3) “Diagonalmatrix” mit den Wurzeln der Eigenwerte (von $A \cdot A^T$ bzw U bzw V auf der Diagonale), nach Größe geordnet:

$$\Sigma = \begin{pmatrix} \sqrt{12} & 0 & 0 \\ 0 & \sqrt{10} & 0 \end{pmatrix}$$

Alles Zusammen

$$A = U\Sigma V^T = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} \sqrt{12} & 0 & 0 \\ 0 & \sqrt{10} & 0 \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{6}} & \frac{2}{\sqrt{6}} & \frac{1}{\sqrt{6}} \\ \frac{2}{\sqrt{5}} & \frac{-1}{\sqrt{5}} & 0 \\ \frac{1}{\sqrt{30}} & \frac{2}{\sqrt{30}} & \frac{-5}{\sqrt{30}} \end{pmatrix} =$$

$$\begin{pmatrix} \frac{\sqrt{12}}{\sqrt{2}} & \frac{\sqrt{10}}{\sqrt{2}} & 0 \\ \frac{\sqrt{12}}{\sqrt{2}} & \frac{-\sqrt{10}}{\sqrt{2}} & 0 \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{6}} & \frac{2}{\sqrt{6}} & \frac{1}{\sqrt{6}} \\ \frac{2}{\sqrt{5}} & \frac{-1}{\sqrt{5}} & 0 \\ \frac{1}{\sqrt{30}} & \frac{2}{\sqrt{30}} & \frac{-5}{\sqrt{30}} \end{pmatrix} =$$

$$\begin{pmatrix} 3 & 1 & 1 \\ -1 & 3 & 1 \end{pmatrix}$$

Möglicher Schritt VIII: Vorzeichenverbesserung

Nehmen wir mal an, wir hätten bei V die falschen Vorzeichen gewählt. Zum Beispiel wir hätten für $\sigma_2 = 10$ anstatt $(2, -1, 0)$ den Vektor $(-2, 1, 0)$ gewählt. Damit ergibt sich

$$\tilde{V} = \begin{pmatrix} 1 & -2 & 1 \\ 2 & 1 & 2 \\ 1 & 0 & -5 \end{pmatrix}$$

und

$$V = \begin{pmatrix} \frac{1}{\sqrt{6}} & \frac{-2}{\sqrt{5}} & \frac{1}{\sqrt{30}} \\ \frac{2}{\sqrt{6}} & \frac{1}{\sqrt{5}} & \frac{2}{\sqrt{30}} \\ \frac{1}{\sqrt{6}} & 0 & \frac{-5}{\sqrt{30}} \end{pmatrix}$$

und

$$V^T = \begin{pmatrix} \frac{1}{\sqrt{6}} & \frac{2}{\sqrt{6}} & \frac{1}{\sqrt{6}} \\ \frac{-2}{\sqrt{5}} & \frac{1}{\sqrt{5}} & 0 \\ \frac{1}{\sqrt{30}} & \frac{2}{\sqrt{30}} & \frac{-5}{\sqrt{30}} \end{pmatrix}$$

Möglicher Schritt VIII: Vorzeichenverbesserung

Leider ist nun

$$U\Sigma V^T = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} \sqrt{12} & 0 & 0 \\ 0 & \sqrt{10} & 0 \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{6}} & \frac{2}{\sqrt{6}} & \frac{1}{\sqrt{6}} \\ \frac{-2}{\sqrt{5}} & \frac{1}{\sqrt{5}} & 0 \\ \frac{1}{\sqrt{30}} & \frac{2}{\sqrt{30}} & \frac{-5}{\sqrt{30}} \end{pmatrix} =$$

$$\begin{pmatrix} \frac{\sqrt{12}}{\sqrt{2}} & \frac{\sqrt{10}}{\sqrt{2}} & 0 \\ \frac{\sqrt{12}}{\sqrt{2}} & \frac{-\sqrt{10}}{\sqrt{2}} & 0 \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{6}} & \frac{2}{\sqrt{6}} & \frac{1}{\sqrt{6}} \\ \frac{-2}{\sqrt{5}} & \frac{1}{\sqrt{5}} & 0 \\ \frac{1}{\sqrt{30}} & \frac{2}{\sqrt{30}} & \frac{-5}{\sqrt{30}} \end{pmatrix} =$$

$$\begin{pmatrix} -1 & 3 & 1 \\ 3 & 1 & 1 \end{pmatrix}$$

Möglicher Schritt VIII: Vorzeichenverbesserung

Die unelegante Lösung: Gehe durch alle Zeilen von V^T (Spalten von V), multizipliere sie systematisch mit -1, bis es passt. Durch die Multiplikation mit -1, bleiben die Vektoren immer noch normal und auch die Orthogonalität ändert sich nicht.

- Da eventuell auch mehrere Zeilen mit -1 multipliziert werden müssen, ist dies leider exponentiell in der Zeilenanzahl von V^T (alle möglichen Kombinationen müssen probiert werden)
- Es ist nicht gerade elegant...
- Man kann natürlich stattdessen auch durch die Spalten von U gehen und diese anpassen.

Konstruktionsmethodologie II

1. Berechne die quadratische $n \times n$ Matrix $A^T \cdot A$
2. Berechne die Eigenwerte σ_j^2 und Eigenvektoren v_j von $A^T \cdot A$.
3. Die Vektoren v_j werden orthonormalisiert (Gram-Schmidt) und bilden die Spaltenvektoren von V
4. Die Wurzeln der σ^2-Werte werden in absteigender Reihenfolge die Einträge von Σ
5. Die Spaltenvektoren von U bis zum Rang r der $A^T \cdot A$ Matrix sind nun eindeutig bestimmt aus der Vorgabe $A = U\Sigma V^T$ und werden daraus berechnet.
6. Wenn $r < m$ müssen noch $m - r$ Vektoren zur Orthonormalbasis ergänzt werden.

Konstruktionsmethodologie II

1. Berechne die quadratische $n \times n$ Matrix $A^T \cdot A$
2. Berechne die Eigenwerte σ_j^2 und Eigenvektoren v_j von $A^T \cdot A$.
3. Die Vektoren v_j werden orthonormalisiert (Gram-Schmidt) und bilden die Spaltenvektoren von V
4. Die Wurzeln der σ^2-Werte werden in absteigender Reihenfolge die Einträge von Σ
5. Die Spaltenvektoren von U bis zum Rang r der $A^T \cdot A$ Matrix sind nun eindeutig bestimmt aus der Vorgabe $A = U\Sigma V^T$ und werden daraus berechnet.
6. Wenn $r < m$ müssen noch $m - r$ Vektoren zur Orthonormalbasis ergänzt werden.

Konstruktionsmethodologie II

1. Berechne die quadratische $n \times n$ Matrix $A^T \cdot A$
2. Berechne die Eigenwerte σ_j^2 und Eigenvektoren v_j von $A^T \cdot A$.
3. Die Vektoren v_j werden orthonormalisiert (Gram-Schmidt) und bilden die Spaltenvektoren von V
4. Die Wurzeln der σ^2-Werte werden in absteigender Reihenfolge die Einträge von Σ
5. Die Spaltenvektoren von U bis zum Rang r der $A^T \cdot A$ Matrix sind nun eindeutig bestimmt aus der Vorgabe $A = U\Sigma V^T$ und werden daraus berechnet.
6. Wenn $r < m$ müssen noch $m - r$ Vektoren zur Orthonormalbasis ergänzt werden.

Konstruktionsmethodologie II

1. Berechne die quadratische $n \times n$ Matrix $A^T \cdot A$
2. Berechne die Eigenwerte σ_j^2 und Eigenvektoren v_j von $A^T \cdot A$.
3. Die Vektoren v_j werden orthonormalisiert (Gram-Schmidt) und bilden die Spaltenvektoren von V
4. Die Wurzeln der σ^2-Werte werden in absteigender Reihenfolge die Einträge von Σ
5. Die Spaltenvektoren von U bis zum Rang r der $A^T \cdot A$ Matrix sind nun eindeutig bestimmt aus der Vorgabe $A = U\Sigma V^T$ und werden daraus berechnet.
6. Wenn $r < m$ müssen noch $m - r$ Vektoren zur Orthonormalbasis ergänzt werden.

Schritt I: Bestimme V^T und Σ wie bisher

Nach dem üblichen Verfahren sei also V^T zum Beispiel

$$\begin{pmatrix} \frac{1}{\sqrt{6}} & \frac{2}{\sqrt{6}} & \frac{1}{\sqrt{6}} \\ \frac{-2}{\sqrt{5}} & \frac{1}{\sqrt{5}} & 0 \\ \frac{1}{\sqrt{30}} & \frac{2}{\sqrt{30}} & \frac{-5}{\sqrt{30}} \end{pmatrix}$$

bzw V

$$V = \begin{pmatrix} \frac{1}{\sqrt{6}} & \frac{-2}{\sqrt{5}} & \frac{1}{\sqrt{30}} \\ \frac{2}{\sqrt{6}} & \frac{1}{\sqrt{5}} & \frac{2}{\sqrt{30}} \\ \frac{1}{\sqrt{6}} & 0 & \frac{-5}{\sqrt{30}} \end{pmatrix}$$

Die Vorzeichen könnten auch anders sein.

Weiterhin ist

$$\Sigma = \begin{pmatrix} \sqrt{12} & 0 & 0 \\ 0 & \sqrt{10} & 0 \end{pmatrix}$$

Schritt II: Bestimme U

Es muss gelten:

$$A = U\Sigma V^T$$

und damit

$$AV = U\Sigma V^T V$$

Da V orthogonal, ist $V^T V$ die Einheitsmatrix und damit

$$AV = U\Sigma$$

Schritt II: Bestimme U

Man bilde nun eine $n \times m$ Matrix Σ', die genau auf den entsprechenden Diagonaleinträgen $1/\sigma_i$ stehen hat. Im Beispiel ist

$$\Sigma' = \begin{pmatrix} \frac{1}{\sqrt{12}} & 0 \\ 0 & \frac{1}{\sqrt{10}} \\ 0 & 0 \end{pmatrix}$$

Es gilt, dass $\Sigma \cdot \Sigma'$ die $m \times m$ Einheitsmatrix ist.
Und damit

$$AV\Sigma' = U\Sigma\Sigma' = U$$

Schritt II: Bestimme U

Also gilt

$$U = AV\Sigma'$$

.

Damit gilt für die Spalten von U, wenn man sich die Matrixmultiplikation genauer anschaut:

$$u^{(j)} = \frac{1}{\sigma_j} A v^{(j)}$$

In unserem Beispiel:

$$u^{(1)} = \frac{1}{\sqrt{12}} \begin{pmatrix} 3 & 1 & 1 \\ -1 & 3 & 1 \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{6}} \\ \frac{2}{\sqrt{6}} \\ \frac{1}{\sqrt{6}} \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}$$

Schritt II: Bestimme U

Und ebenso

$$u^{(2)} = \frac{1}{\sqrt{10}} \begin{pmatrix} 3 & 1 & 1 \\ -1 & 3 & 1 \end{pmatrix} \begin{pmatrix} \frac{-2}{\sqrt{5}} \\ \frac{1}{\sqrt{5}} \\ 0 \end{pmatrix} = \begin{pmatrix} \frac{-1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}$$

Man sieht, dass sich hier das veränderte Vorzeichen von $v^{(2)}$ im entgegengesetzten Vorzeichen von $u^{(2)}$ widerspiegelt. Wir mussten auch nicht zweimal Eigenvektoren berechnen, sondern U ist aus den vorherigen Berechnungen bestimmt

Beispiel mit einer Matrix mit niedrigerem Rang

Da wir durch $\frac{1}{\sigma_j}$ dividieren mussten, geht dies nur für Eigenwerte ungleich Null. Wenn aber die Matrizen niedrigere Ränge haben als m, dann bekommen wir so nicht genügend U-Vektoren. Was machen wir dann?

Beispiel. Sei eine Matrix von Rang 1 gegeben.

$$A = \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix}$$

Berechnung von V

$$A^T A = \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix}$$

Berechnung von V

$$\begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix}$$

hat die Eigenwerte 2 und 0 mit den jeweiligen Eigenvektoren $(1,0)$ sowie $(0,1)$.
Daraus folgt

$$\tilde{V} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

Die ist schon eine Orthonormalmatrix also gilt $V = \tilde{V}$ sowie, da Diagonalmatrix $V^T = V$

Bestimmung von Σ

Es gilt:

$$\Sigma = \begin{pmatrix} \sqrt{2} & 0 \\ 0 & 0 \end{pmatrix}$$

Bestimmung von U

Der erste Spaltenvektor $u^{(1)}$ kann wie üblich bestimmt werden mit

$$u^{(1)} = \frac{1}{\sqrt{2}} A v^{(1)} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

$$= \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}$$

Bestimmung von U

Da wir keinen weiteren Eigenwert $\neq 0$ haben, können wir leider so nicht weitermachen, d.h. es geht nur bis zum Rang von A. Es gilt aber

$$\begin{pmatrix} \frac{1}{\sqrt{2}} & x_2 \\ \frac{1}{\sqrt{2}} & y_2 \end{pmatrix} \begin{pmatrix} \sqrt{2} & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

$$= \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix} = A$$

unabhängig von x_2 und y_2. Das heisst unser halb vollständiges U ist genug für eine Zerlegung.

Bestimmung von U

1. Wir können also jedes x_2 und y_2 wählen, solange normiert und senkrecht auf dem ersten Vektor stehend.
2. Nach dem sogenannten **Basisergänzungssatz** geht dies immer!
3. In unserem Beispiel könnte man zum Beispiel $(\frac{-1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$ wählen.
4. Wie genau man diese weiteren Vektoren von U konstruiert, ist für uns unerheblich, da wir ja sowieso die letzten Spalten von U, die zu den niedrigeren Eigenwerten gehören, “abschneiden” werden...

SVD Zwischenfazit

Wir haben eine Matrix in das Produkt einer Orthonormalmatrix, Diagonalmatrix und Orthogonalmatrix zerlegt

- Die Diagonalmatrix enthält die Wurzeln von den Eigenwerten von $A \cdot A^T$ nach Größe geordnet
- Die erste Orthonormalmatrix enthält die orthonormalisierten Eigenvektoren von $A \cdot A^T$, nach Eigenwertgrösse geordnet, als Spalten
- Die zweite Orthonormalmatrix ist die Transponierte der Matrix, die die orthonormalisierten Eigenvektoren von $A^T \cdot A$, nach Eigenwertgrösse geordnet, als Spalten enthält.

Übersicht

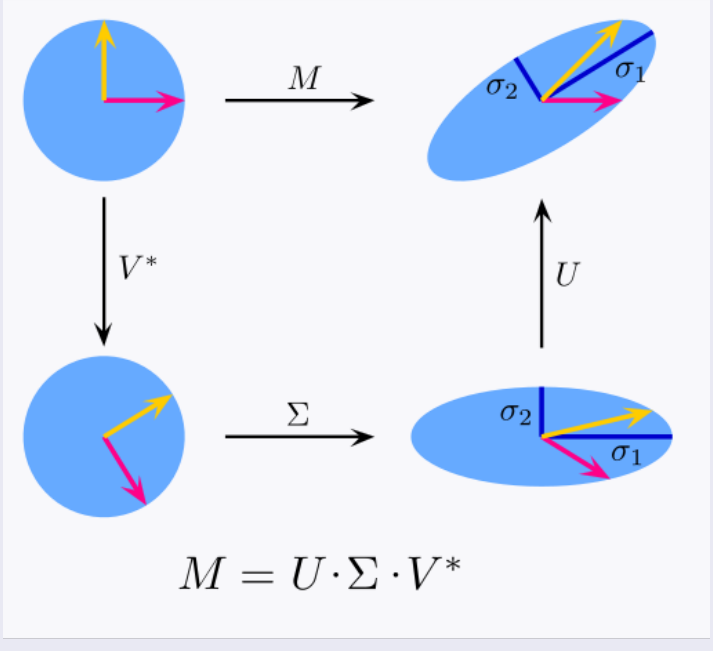
Geometrische Interpretation bei quadratischen Matrizen

Bedeutung der SVD ist, dass sich die Funktion in eine Drehung/Drehspiegelung (V^T), eine Verzerrung (Σ), und eine zweite Drehung/Drehspiegelung (U) zerlegen lsst.

Matrix $\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$

SVD: Geometrische Interpretation[a]

[a] Quelle: https://commons.wikimedia.org/wiki/File:Singular-Value-Decomposition.svg



Noch einige Bemerkungen

- Die Eigenwerte und damit Σ sind eindeutig bestimmt.
- U und V sind nicht eindeutig bestimmt. Es können Drehungen um 180 Grad oder Spiegelungen hinzukommen.
- Wir können in vielen Fällen einen Eigenvektor wählen (Freiheitsgrade), müssen aber dann bei den Vorzeichen der zweiten Matrix aufpassen.

Noch einige Bemerkungen

- Wir haben implizit den Rang der Ursprungsmatrix A bestimmt. Dies kann man klar an $U \cdot \Sigma$ sehen.

$$A = U\Sigma V^T = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} \sqrt{12} & 0 & 0 \\ 0 & \sqrt{10} & 0 \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{6}} & \frac{2}{\sqrt{6}} & \frac{1}{\sqrt{6}} \\ \frac{2}{\sqrt{5}} & \frac{-1}{\sqrt{5}} & 0 \\ \frac{1}{\sqrt{30}} & \frac{2}{\sqrt{30}} & \frac{-5}{\sqrt{30}} \end{pmatrix} =$$

$$\begin{pmatrix} \frac{\sqrt{12}}{\sqrt{2}} & \frac{\sqrt{10}}{\sqrt{2}} & 0 \\ \frac{\sqrt{12}}{\sqrt{2}} & \frac{-\sqrt{10}}{\sqrt{2}} & 0 \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{6}} & \frac{2}{\sqrt{6}} & \frac{1}{\sqrt{6}} \\ \frac{2}{\sqrt{5}} & \frac{-1}{\sqrt{5}} & 0 \\ \frac{1}{\sqrt{30}} & \frac{2}{\sqrt{30}} & \frac{-5}{\sqrt{30}} \end{pmatrix} =$$

$$\begin{pmatrix} 3 & 1 & 1 \\ -1 & 3 & 1 \end{pmatrix}$$

Noch einige Bemerkungen

Deswegen wird manchmal die Singulärwertzerlegung auch als Zerlegung in drei Matrizen $A_{m\times n} = U_{m\times r}\Sigma_{r\times r}V^T_{r\times n}$, wobei r der Rang der Matrix A, ist beschrieben.

$$A = U\Sigma V^T = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} \sqrt{12} & 0 & 0 \\ 0 & \sqrt{10} & 0 \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{6}} & \frac{2}{\sqrt{6}} & \frac{1}{\sqrt{6}} \\ \frac{2}{\sqrt{5}} & \frac{-1}{\sqrt{5}} & 0 \\ \frac{1}{\sqrt{30}} & \frac{2}{\sqrt{30}} & \frac{-5}{\sqrt{30}} \end{pmatrix} =$$

$$\begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} \sqrt{12} & 0 \\ 0 & \sqrt{10} \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{6}} & \frac{2}{\sqrt{6}} & \frac{1}{\sqrt{6}} \\ \frac{2}{\sqrt{5}} & \frac{-1}{\sqrt{5}} & 0 \end{pmatrix} =$$

$$\begin{pmatrix} \frac{\sqrt{12}}{\sqrt{2}} & \frac{\sqrt{10}}{\sqrt{2}} \\ \frac{\sqrt{12}}{\sqrt{2}} & \frac{-\sqrt{10}}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{6}} & \frac{2}{\sqrt{6}} & \frac{1}{\sqrt{6}} \\ \frac{2}{\sqrt{5}} & \frac{-1}{\sqrt{5}} & 0 \end{pmatrix} =$$

$$\begin{pmatrix} 3 & 1 & 1 \\ -1 & 3 & 1 \end{pmatrix}$$

Noch einige Bemerkungen

- Bisher haben wir immer die Ausgangsmatrix einfach reproduziert (mit dem gleichen Rang). Man sieht den Rang besser und es gilt:
- Die “Zerrung” durch die Matrix Σ ist bei den Dimensionen mit den höchsten Eigenwerten am größten. Beispiel:

$$\begin{pmatrix} \sqrt{12} & 0 & 0 \\ 0 & \sqrt{10} & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} =$$

$$\begin{pmatrix} \sqrt{12} \cdot x_1 \\ \sqrt{10} \cdot x_2 \end{pmatrix}$$

- Es macht also Sinn, dass wir, wenn wir eine **Approximation** unserer Ausgangsmatrix mit niedrigerem Rang wollen, dann die unwichtigen Dimensionen, die durch niedrige Eigenwerte gekennzeichnet sind weglassen.

Übersicht

Reduced SVD: Approximation niedrigeren Ranges

- Approximation von $m \times n$ Matrix A mit Rang r durch $m \times n$ - Matrix $\tilde{A}_k$ mit Rang $k < r$:

$$A = U\Sigma V^T \approx \tilde{A}_k = U\tilde{\Sigma}_k V^T$$

- $\tilde{\Sigma}_k$ ist eine “diagonale” $m \times n$ - Matrix, bei der nur auf den ersten k Stellen der Diagonalmatrix Werte $\neq 0$ stehen
- U und V wie zuvor

Beispiel

$$A = U\Sigma V^T = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} \sqrt{12} & 0 & 0 \\ 0 & \sqrt{10} & 0 \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{6}} & \frac{2}{\sqrt{6}} & \frac{1}{\sqrt{6}} \\ \frac{2}{\sqrt{5}} & \frac{-1}{\sqrt{5}} & 0 \\ \frac{1}{\sqrt{30}} & \frac{2}{\sqrt{30}} & \frac{-5}{\sqrt{30}} \end{pmatrix} \approx$$

$$\begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} \sqrt{12} & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{6}} & \frac{2}{\sqrt{6}} & \frac{1}{\sqrt{6}} \\ \frac{2}{\sqrt{5}} & \frac{-1}{\sqrt{5}} & 0 \\ \frac{1}{\sqrt{30}} & \frac{2}{\sqrt{30}} & \frac{-5}{\sqrt{30}} \end{pmatrix} =$$

$$\begin{pmatrix} \frac{\sqrt{12}}{\sqrt{2}} & 0 & 0 \\ \frac{\sqrt{12}}{\sqrt{2}} & 0 & 0 \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{6}} & \frac{2}{\sqrt{6}} & \frac{1}{\sqrt{6}} \\ \frac{2}{\sqrt{5}} & \frac{-1}{\sqrt{5}} & 0 \\ \frac{1}{\sqrt{30}} & \frac{2}{\sqrt{30}} & \frac{-5}{\sqrt{30}} \end{pmatrix} =$$

$$\begin{pmatrix} 1 & 2 & 1 \\ 1 & 2 & 1 \end{pmatrix}$$

Reduced SVD

Wir hätten auch nur die ersten k Spalten von U sowie die ersten k Zeilen von V^T behalten können:

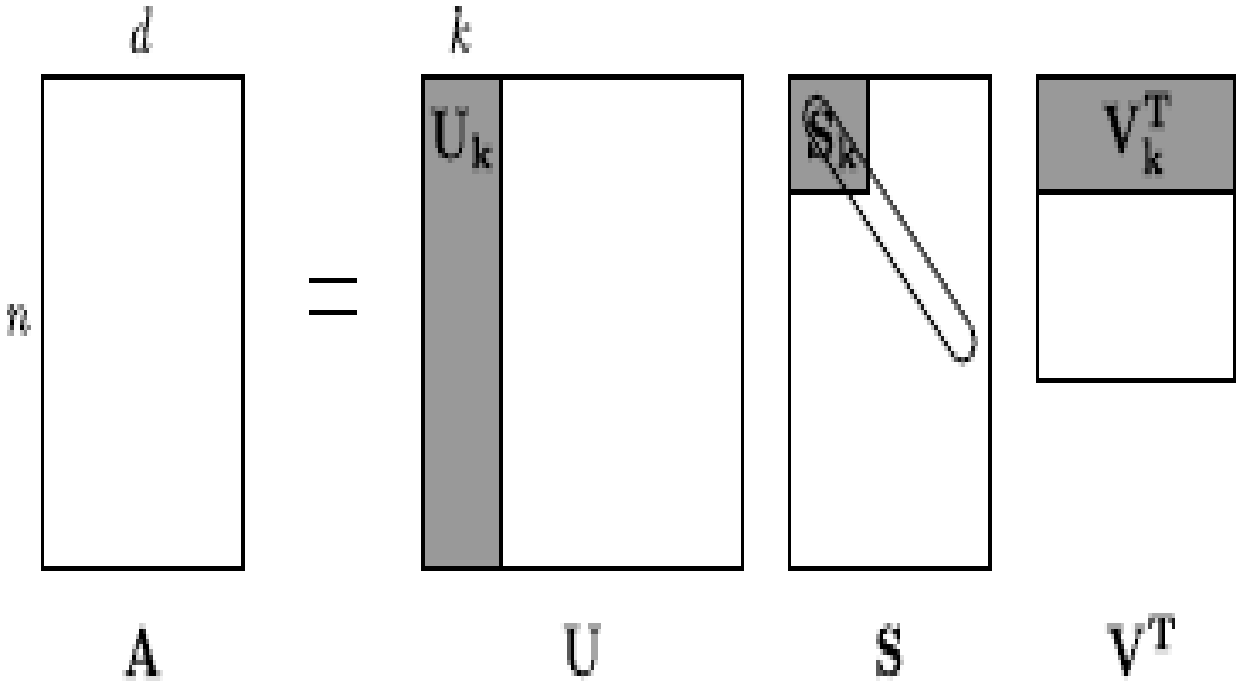
$$A \approx$$

$$\begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix} \left(\sqrt{12}\right) \begin{pmatrix} \frac{1}{\sqrt{6}} & \frac{2}{\sqrt{6}} & \frac{1}{\sqrt{6}} \end{pmatrix} =$$

$$\begin{pmatrix} \frac{\sqrt{12}}{\sqrt{2}} \\ \frac{\sqrt{12}}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{6}} & \frac{2}{\sqrt{6}} & \frac{1}{\sqrt{6}} \end{pmatrix} =$$

$$\begin{pmatrix} 1 & 2 & 1 \\ 1 & 2 & 1 \end{pmatrix}$$

Als Bild



Als Algorithmus

Gegeben $m \times n$ Matrix A und gewünschter Rank k. Wir produzieren eine $Rang - k$ Approximation wie folgt:

1. Wir berechnen eine SVD Zerlegung von A:

$$A_{m \times n} = U_{m \times m} \Sigma_{m \times n} V^T_{n \times n}$$

2. Wir behalten nur die ersten k Spalten von U, resultierend in $U_{m \times k}$
3. Wir behalten nur die Top k, sprich die ersten k Diagonalwerte von Σ, resultierend in $\Sigma_{k \times k}$
4. Wir behalten nur die ersten k Reihen von V^T, resultierend in $V^T_{k \times n}$

Reduced SVD

- Diagonale Einträge in $\tilde{\Sigma}_k$ sind die k größten Eigenwerte von $A \cdot A^T$
- Satz: $U\tilde{\Sigma}_k V^T$ minimiert die Differenz der Frobenius-Norm zwischen A und $\tilde{A}_k$ unter der Bedingung $rang(\tilde{A}_k) = k$
- Das heisst: Für jede $m \times n$ Matrix A mit Rang r, angestrebtem Rang $k < r$, und jede $m \times n$ Matrix B mit Rang k gilt

$$\|A - \tilde{A}_k\|_2 \leq \|A - B\|_2$$

wobei $\tilde{A}_k$ die Rang-k Approximation ist, die man aus der SVD von A ableiten kann.

Was hat man erreicht?

- Unkorrelierte Dimensionen (Orthonormalbasis), die nur noch verzerrt werden
- Wir haben die Dimensionen mit der meisten Variation identifiziert und danach geordnet
- “Unwichtige” Dimensionen werden weggelassen $\rightarrow$ Rauschen vermindert
- Man braucht viel weniger Speicherplatz: $O(k(n+m))$ anstatt $O(nm)$
- Die neuen “versteckten” Dimensionen können zur Ähnlichkeitsberechnung verwendet werden und funktionieren oft besser

Wie wählt man k?

Meist empirisch. Gesehen:

1. In NLP, oft die ersten 200 bis 300 Dimensionen
2. In Recommender Systems: Wähle k so, dass die Summe der ersten k Eigenwerte mindestens 10-mal der Summe aller anderen Eigenwerte

Übersicht

Dichte Embeddings

- A ist Matrix von m Wörtern/Termen und n Kontexten (PPMI, frequency, . . .)
- Reduced SVD resultiert in $\tilde{A}_k$ mit
  - $U_{m \times k}$ ist $m \times k$ - Matrix
  - Zeilen von $U_{m \times k}$ repräsentieren Terme
  - Spalten in $U_{m \times k}$: latente Repräsentation der Kontexte
  - Benutze $U_{m \times k} \cdot \Sigma_{k \times k}$ zur Berechnung von Termähnlichkeiten
- Vorteile gegenüber der vollen Term-Kontext-Matrix
  - Dichte Matrix
  - Dimensionsreduktion: effizientere Verarbeitung
  - De-Noising

Dichte Embeddings

- A ist Matrix von m Wörtern/Termen und n Kontexten (PPMI, frequency, ...)
- Reduced SVD resultiert in $\tilde{A}_k$ mit
  - $U_{m \times k}$ ist $m \times k$ - Matrix
  - Zeilen von $U_{m \times k}$ repräsentieren Terme
  - Spalten in $U_{m \times k}$: latente Repräsentation der Kontexte
  - Benutze $U_{m \times k} \cdot \Sigma_{k \times k}$ zur Berechnung von Termähnlichkeiten
- Vorteile gegenüber der vollen Term-Kontext-Matrix
  - Dichte Matrix
  - Dimensionsreduktion: effizientere Verarbeitung
  - De-Noising

Dichte Embeddings

- A ist Matrix von m Wörtern/Termen und n Kontexten (PPMI, frequency, ...)
- Reduced SVD resultiert in $\tilde{A}_k$ mit
  - $U_{m \times k}$ ist $m \times k$ - Matrix
  - Zeilen von $U_{m \times k}$ repräsentieren Terme
  - Spalten in $U_{m \times k}$: latente Repräsentation der Kontexte
  - Benutze $U_{m \times k} \cdot \Sigma_{k \times k}$ zur Berechnung von Termähnlichkeiten
- Vorteile gegenüber der vollen Term-Kontext-Matrix
  - Dichte Matrix
  - Dimensionsreduktion: effizientere Verarbeitung
  - De-Noising

Das Beispiel vom Anfang

Eine Wort-Dokument-Matrix M aus dem $\mathbb{R}^{5\times 3}$

	$d1$	$d2$	$d3$
ship	1	1	0
boat	0	0	1
ocean	1	0	1
motor	1	0	1
wood	0	1	0

$$cos_{sim}(ship, boat) = 0$$

$$cos_{sim}(ship, ocean) = \frac{1}{2}$$

$$cos_{sim}(boat, ocean) = \frac{1}{\sqrt{2}} = 0.7$$

SVD Zerlegung war:

$$\begin{bmatrix} -0.41 & 0.7 & 0.24 \\ -0.29 & -0.33 & -0.72 \\ -0.61 & -0.2 & 0.14 \\ -0.61 & -0.2 & 0.14 \\ -0.1 & 0.57 & -0.62 \end{bmatrix} \begin{bmatrix} 2.27 & 0 & 0 \\ 0 & 1.49 & 0 \\ 0 & 0 & 0.78 \end{bmatrix} \begin{bmatrix} -0.72 & -0.23 & -0.66 \\ 0.19 & 0.85 & -0.5 \\ 0.67 & -0.48 & -0.56 \end{bmatrix}$$

$$= \begin{bmatrix} -0.93 & 1.04 & 0.187 \\ -0.658 & -0.49 & -0.56 \\ -1.38 & -0.298 & 0.1092 \\ -1.38 & -0.298 & -.1092 \\ -0.227 & 0.84 & -0.483 \end{bmatrix} \begin{bmatrix} -0.72 & -0.23 & -0.66 \\ 0.19 & 0.85 & -0.5 \\ 0.67 & -0.48 & -0.56 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$$

Abweichungen sind Rundungsfehler

Niedrigdimensionale Approximation

Die kleinsten Eigenwerte sind die unwichtigsten. Wir können diese "weglassen" = auf Null setzen $\rightarrow$ eine Matrix mit kleinerem Rang, die aber relativ ähnlich zur Ausgangsmatrix ist.

$$\begin{bmatrix} -0.41 & 0.7 & 0.24 \\ -0.29 & -0.33 & -0.72 \\ -0.61 & -0.2 & 0.14 \\ -0.61 & -0.2 & 0.14 \\ -0.1 & 0.57 & -0.62 \end{bmatrix} \begin{bmatrix} 2.27 & 0 & 0 \\ 0 & 1.49 & 0 \\ 0 & 0 & \mathbf{0} \end{bmatrix} \begin{bmatrix} -0.72 & -0.23 & -0.66 \\ 0.19 & 0.85 & -0.5 \\ 0.67 & -0.48 & -0.56 \end{bmatrix}$$

$$= \begin{bmatrix} -0.93 & 1.04 & 0 \\ -0.658 & -0.49 & 0 \\ -1.38 & -0.298 & 0 \\ -1.38 & -0.298 & 0 \\ -0.227 & 0.84 & 0 \end{bmatrix} \begin{bmatrix} -0.72 & -0.23 & -0.66 \\ 0.19 & 0.85 & -0.5 \\ 0.67 & -0.48 & -0.56 \end{bmatrix}$$

$$= \begin{bmatrix} 0.87 & 1.09 & 0.11 \\ 0.38 & -0.27 & 0.68 \\ 0.93 & 0.05 & 1.06 \\ 0.93 & 0.05 & 1.06 \\ 0.32 & 0.77 & -0.27 \end{bmatrix}$$

Die niedrigdimensionale Approximation

Wir interessieren uns für die Matrix $U_2 = U\Sigma_2$, also die Matrix mit dem niedrigerem Rang:

$$\begin{bmatrix} -0.93 & 1.04 & 0 \\ -0.658 & -0.49 & 0 \\ -1.38 & -0.298 & 0 \\ -1.38 & -0.298 & 0 \\ -0.227 & 0.84 & 0 \end{bmatrix}$$

Man kann diese nun als die Repräsentation unserer 5 Wörter mit zwei versteckten Dimensionen auffassen:

	$h1$	$h2$
ship	-0.93	1.04
boat	-0.658	-0.49
ocean	-1.38	-0.298
motor	-1.38	-0.298
wood	-0.227	0.84

Neue Ähnlichkeitsberechnungen

	$h1$	$h2$
ship	-0.93	1.04
boat	-0.658	-0.49
ocean	-1.38	-0.298
motor	-1.38	-0.298
wood	-0.227	0.84

$$cos_{sim}(ship, boat) = \frac{(-0.93) \cdot (-0.658) + 1.04 \cdot (-0.49)}{\sqrt{(0.93^2 + 1.04^2)} \cdot \sqrt{(0.658^2 + 0.49^2)}} = 0.09$$

$$cos_{sim}(ship, ocean) = \frac{(-0.93) \cdot (-1.38) + 1.04 \cdot (-0.29)}{\sqrt{(0.93^2 + 1.04^2)} \cdot \sqrt{(1.38^2 + 0.29^2)}} = 0.49$$

$$cos_{sim}(boat, ocean) = 0.9$$

Variationen

- Benutze nur $U_{m \times k}$ zur Ähnlichkeitsberechnung
- Benutze $U_{m \times k} \cdot \Sigma^{p}_{k \times k}$ zur Ähnlichkeitsberechnung $\rightarrow$ Parameter p sollte getuned werden. Kann große Unterschiede machen!
- Oft $p = 0, p = 0.5, p = 1$

Performanz der Reduced-SVD-Embeddings

Levy, O.; Goldberg, Y. *Neural word embedding as implicit matrix factorization.* In: Advances in neural information processing systems. 2014. S. 2177-2185.

- (Englische) Wikipedia, 1.5 Milliarden Tokens
- Fenster 2 beiderseits
- Vokabular 189, 533 für Terme sowie Kontexte
- PPMI Matrizen
- SVD sowie ohne Zerlegung (PPMI) sowie neurale (SGNS)

Algorithmus	Spearman Rank zu WordSim 353
SVD	0.652
SGNS	0.633
PPMI	0.605

Latent Semantic Analysis und Word Embeddings

Levy2014:

"We analyze skip-gram with negative-sampling (SGNS) [...] and show that it is implicitly factorizing a word-context matrix, whose cells are the pointwise mutual information (PMI) of the respective word and context pairs, shifted by a global constant. . . . When dense low-dimensional vectors are preferred, exact factorization with SVD can achieve solutions that are at least as good as SGNS'ss solutions for word similarity tasks. On analogy questions SGNS remains superior to SVD. We conjecture that this stems from the weighted nature of SGNS's factorization."

Zusammenfassung

- Reduced SVD resultiert in einer optimalen Approximation der Originalmatrix mit niedrigerem Rang. Optimal = optimal bzgl der Frobeniusnorm (bzw mean squared error)
- Dadurch erhalten wir eine mathematisch fundierte, dichtere Matrix, die Rauschen (weniger wichtige Dimensionen) eliminiert
- Wir können mittels $U_{m\times k}\Sigma_{k\times k}$ weiterhin Wortähnlichkeiten berechnen
- Es können Resultate erreicht werden, die bei *word similarity* mindestens ebensogut sind wie mit neuralen word embeddings
- Nachteile: Kosten der SVD Berechnung für $m \times n$ matrix: $O(n \cdot m^2)$ (when $m < n$) $\rightarrow$ Nicht gut für sehr grosse Matrizen
- Nachteile: Geht nicht gut mit unbekannten Wörtern um
- Nachteile: schlechter bei Wortanalogieaufgaben

Literatur

- Guter erweiterter Hintergrund: `http://theory.stanford.edu/~tim/s15/l/l7.pdf` bis `http://theory.stanford.edu/~tim/s15/l/l9.pdf`
- SVD Tutorial (ohne vollständigen Hintergrund) : Kirk Baker (2005): Singular Value Decomposition Tutorial `https://datajobs.com/data-science-repo/SVD-Tutorial-[Kirk-Baker].pdf`. Leider ist hier im Konstruktionsverfahren I die Vorzeichenproblematik fälschlicherweise nicht behandelt worden...
- Levy, O.; Goldberg, Y. *Neural word embedding as implicit matrix factorization.* In: Advances in neural information processing systems. 2014. S. 2177-2185.
- Übungsblatt II