Computer-Linguistische Anwendungen

CLA | B.Sc. | LMU

Matrix Faktorisierung

15''

Matrix Faktorisierung: Überblick

- Wir möchten die Wort-Dokument Matrix in ein Produkt von Matrizen zerlegen
- Die Zerlegung die wir verwenden ist: Singular Value Decomposition (SVD, Singulärwertzerlegung)

  SVD: $C = U \Sigma V^T$ (wobei C = Wort-Dokument Matrix ist) - wir schauen uns dies genauer an
- Nach der Zerlegung verwenden wir SVD um die neue Wort-Dokument Matrix C' zu berechnen
- Die Ähnlichkeits-Werte in C' sind besser im Vergleich zu C
- Die Verwendung von SVD in diesem Kontext (Information Retrieval) bezeichnet man auch als latent semantic indexing (LSI)

Σ: Sigma

Matrix Faktorisierung: Überblick

Intuition:

Wenn wir eine Matrix haben, ist es oft hilfreich, sie in ihre Bestandteile zu zerlegen und auf eine sinnvolle Weise anzuordnen. Die Singulärwertzerlegung ist eine Methode, um dies zu tun.

Die Singulärwertzerlegung (SVD) zerlegt eine Matrix in drei Teile: U, Σ und V^T. U und V^T sind dabei orthogonale Matrizen, die aus den Eigenvektoren der Matrix CC^T bzw. C^TC bestehen. Die Matrix Σ ist eine Diagonalmatrix, deren Einträge die Singulärwerte der Matrix C sind.

Matrix Faktorisierung: Überblick

Intuition:

Um die Matrizen U, Σ und V^T zu berechnen, müssen wir ein paar Schritte durchführen:

1. Zuerst müssen wir die Matrix $C*C^T$ berechnen und ihre Eigenwerte und Eigenvektoren finden. Dies gibt uns die Matrix U.
2. Als nächstes müssen wir die Matrix C^T*C berechnen und ihre Eigenwerte und Eigenvektoren finden. Dies gibt uns die Matrix V.
3. Wir ordnen die Eigenwerte, die wir in Schritt 1 und 2 gefunden haben, in einer Diagonalmatrix an. Dies gibt uns die Matrix Σ.

Matrix Faktorisierung: Überblick

Intuition Eigenwerte und Eigenvektoren:

$C*C^T$ ist ein Produkt aus der Matrix C und ihrer Transponierten. Indem wir ihre Eigenwerte und Eigenvektoren finden, können wir eine neue Basis der Daten erstellen, in der sie einfacher zu analysieren sind.

Wenn wir $C*C^T$ berechnen, multiplizieren wir jede Zeile von C mit jeder Spalte von C^T. Das Ergebnis ist eine neue symmetrische Matrix, die uns sagt, wie ähnlich sich die einzelnen Zeilen von C sind.

Wenn wir die Eigenvektoren und Eigenwerte dieser Matrix finden, können wir eine neue Basis der Daten erstellen, in der sie einfacher zu analysieren sind.

Matrix Faktorisierung: Überblick

Intuition Eigenwerte und Eigenvektoren:

Wenn wir C mit einem Vektor multiplizieren, ändert sich der Vektor normalerweise in eine neue Richtung. Aber es gibt bestimmte Vektoren, die nach der Multiplikation mit C in der gleichen Richtung bleiben, nur um einen bestimmten Faktor gestreckt oder gestaucht werden. Diese Vektoren werden als **Eigenvektoren** von C bezeichnet.

Der Faktor, um den der Vektor gestreckt oder gestaucht wird, wird als Eigenwert bezeichnet. Die Eigenwerte und Eigenvektoren von C können uns viel über die Struktur und Eigenschaften der Daten in C verraten.

Matrix Faktorisierung: Überblick

SVD: $C = U \Sigma V^T$

C: Die Matrix C ist dabei eine Darstellung eines Datensatzes als Matrix, wobei die Zeilen die einzelnen Beobachtungen darstellen und die Spalten die Merkmale der Beobachtungen abbilden.

```
C = [[大小1, 重量1],
     [大小2, 重量2],
     [大小3, 重量3],
      ...,
     [大小N, 重量N]]
```

在这里，每一行都是一个单独的观测（一个动物），每一列代表一个特征（大小或重量）。

Matrix Faktorisierung: Überblick

SVD: $C = U \Sigma V^T$

U中的每个向量代表着数据中的一个独立方向

我们以一种新的方式来展示数据

U: Die Matrix U wird benötigt, um die Daten auf eine neue Basis zu projizieren, die aus den Hauptkomponenten der Daten besteht. Die Hauptkomponenten sind dabei die Merkmale, die am meisten zur Varianz der Daten beitragen. Indem wir die Daten auf diese neue Basis projizieren, können wir die Dimensionalität der Daten reduzieren und wichtige Merkmale extrahieren.

数据的主要成分

减少数据维度
提取重要特征

→ Die Matrix U enthält Informationen über die lineare Unabhängigkeit der **Reihen** von C

Matrix Faktorisierung: Überblick

SVD: $C = U \Sigma V^T$

V^T: Die Matrix V^T stellt eine neue Basis für die Spalten von C dar, die sogenannten rechts-singulären Vektoren. Auch diese Basis ist so konstruiert, dass sie die wichtigsten Strukturen in den Daten enthält. Die Singulärwerte von V^T geben uns Auskunft darüber, wie wichtig jede dieser neuen Basisvektoren für die Rekonstruktion der ursprünglichen Daten ist.

→ Die Matrix V^T enthält Informationen über die lineare Unabhängigkeit der **Spalten** von C

Matrix Faktorisierung: Überblick

SVD: $C = U \Sigma V^T$

Σ: Die Matrix Σ in der Singulärwertzerlegung (SVD) von C enthält Informationen über die **Bedeutung der Basisvektoren von U und V^T**. Die Diagonalmatrix A enthält die Singulärwerte von C, die anzeigen, wie wichtig jeder der neuen Basisvektoren von U und V^T für die Rekonstruktion der ursprünglichen Daten in C ist. Die Singulärwerte in A sind in absteigender Reihenfolge angeordnet, was bedeutet, dass die ersten Singulärwerte die größte Bedeutung für die Rekonstruktion der Daten haben.

Beispiel der Matrix $C = U \Sigma V^T$

C	d_1	d_2	d_3	d_4	d_5	d_6
ship	1	0	1	0	0	0
boat	0	1	0	0	0	0
ocean	1	1	0	0	0	0
wood	1	0	0	1	1	0
tree	0	0	0	1	0	1

Document 1: The **ship** in the **ocean** delivers **wood**.

Document 2: The old man the **boat** near the **ocean.**

Wir verwenden eine nicht-gewichtete Matrix um das Beispiel zu vereinfachen (nur 1'en)

Beispiel der Matrix C = U Σ V^T

U	1	2	3	4	5
ship	−0.44	−0.30	0.57	0.58	0.25
boat	−0.13	−0.33	−0.59	0.00	0.73
ocean	−0.48	−0.51	−0.37	0.00	−0.61
wood	−0.70	0.35	0.15	−0.58	0.16
tree	−0.26	0.65	−0.41	0.58	−0.09

Dies ist die orthonormale Matrix. Als orthonormal werden in der Mathematik Vektoren bezeichnet, die zueinander orthogonal sind **und** alle die Norm (anschaulich: Länge) eins besitzen.

- Jeder Reihenvektor hat Einheitslänge
- Jeder Reihenvektor ist orthogonal zu einem allen anderen Reihenvektoren

- Eine Reihe pro Wort, eine **Zeile** pro *min(M, N)* wobei *M* die Anzahl der Wörter und *N* die Anzahl der Dokumente ist.
- Man stelle sich die Dimensionen als “semantische” Dimensionen vor, welche Themen wie Politik, Sport, Wirtschaft abbilden. Z.B. Zeile 2 unterscheidet LAND und WASSER
- Jede Zahl u_{ij} in der Matrix indiziert, wie stark ein Wort *i* im Thema der semantischen Dimension *j* *repräsentiert* ist.

Beispiel der Matrix C = U Σ V^T

U	1	2	3	4	5
ship	−0.44	−0.30	0.57	0.58	0.25
boat	−0.13	−0.33	−0.59	0.00	0.73
ocean	−0.48	−0.51	−0.37	0.00	−0.61
wood	−0.70	0.35	0.15	−0.58	0.16
tree	−0.26	0.65	−0.41	0.58	−0.09

- Wie finden wir diese Werte?
  - Matrix C*CT berechnen und ihre Eigenwerte und Eigenvektoren finden
  - Wie finden wir (Eigenwerte und) Eigenvektoren?

Wir wollen einen bestimmten Vektor finden, der nach der Multiplikation mit C in der gleichen Richtung bleibt, nur um einen bestimmten Faktor gestreckt oder gestaucht wird.

Wir starten mit einem zufälligen Vektor, multiplizieren ihn mit C und normalisieren ihn dann, um sicherzustellen, dass seine Länge gleich bleibt. Der resultierende Vektor ist dann eine bessere Schätzung des gesuchten Eigenvektors.

Dieser Prozess wird so lange wiederholt, bis der Vektor nicht mehr wesentlich geändert wird. Der resultierende Vektor ist dann der Eigenvektor von C mit dem entsprechenden Eigenwert. → dafür benutzen wir effizient-geschriebenen Code

Beispiel der Matrix C = U Σ V^T

V^T	d_1	d_2	d_3	d_4	d_5	d_6
1	−0.75	−0.28	−0.20	−0.45	−0.33	−0.12
2	−0.29	−0.53	−0.19	0.63	0.22	0.41
3	0.28	−0.75	0.45	−0.20	0.12	−0.33
4	0.00	0.00	0.58	0.00	−0.58	0.58
5	−0.53	0.29	0.63	0.19	0.41	−0.22

Dies ist auch eine orthonormale Matrix:

- Jeder Reihenvektor hat Einheitslänge
- Jeder Reihenvektor ist orthogonal zu einem allen anderen Reihenvektoren

- Eine Reihe pro Wort, eine **Reihe** pro *min(M, N)* wobei *M* die Anzahl der Wörter und *N* die Anzahl der Dokumente ist.

- Jede Zahl u_{ij} in der Matrix indiziert, wie stark ein Wort *i* im Thema der semantischen Dimension *j* *repräsentiert* ist.

Beispiel der Matrix C = U Σ V^T

Σ	1	2	3	4	5
1	2.16	0.00	0.00	0.00	0.00
2	0.00	1.59	0.00	0.00	0.00
3	0.00	0.00	1.28	0.00	0.00
4	0.00	0.00	0.00	1.00	0.00
5	0.00	0.00	0.00	0.00	0.39

Diese Matrix ist die quadratische, diagonal Matrix mit den Dimensionen mn(M, N) ✖ min(M, N).

Sie besteht aus den **Singulärwerten** von C.

Diese Werte messen die Wichtigkeit der semantischen Dimension, welche wir nutzen, um "unwichtige" Dimensionen auszuschließen.

Zerlegung von CC^T oder C^TC in die Matrizen V, Σ und V^T. Hierbei ist Σ eine Diagonalmatrix, die die Eigenwerte von CC^T oder C^TC enthält.

Die Eigenwerte sind alle positiv und können daher als Quadratwurzeln dargestellt werden. Diese Quadratwurzeln werden dann als Singulärwerte in die Diagonalmatrix Σ der SVD von C eingefügt.

Beispiel der Matrizen C, U, Σ, V^T

C	d_1	d_2	d_3	d_4	d_5	d_6
ship	1	0	1	0	0	0
boat	0	1	0	0	0	0
ocean	1	1	0	0	0	0
wood	1	0	0	1	1	0
tree	0	0	0	1	0	1

$=$

U	1	2	3	4	5
ship	−0.44	−0.30	0.57	0.58	0.25
boat	−0.13	−0.33	−0.59	0.00	0.73
ocean	−0.48	−0.51	−0.37	0.00	−0.61
wood	−0.70	0.35	0.15	−0.58	0.16
tree	−0.26	0.65	−0.41	0.58	−0.09

$\times$

Σ	1	2	3	4	5
1	2.16	0.00	0.00	0.00	0.00
2	0.00	1.59	0.00	0.00	0.00
3	0.00	0.00	1.28	0.00	0.00
4	0.00	0.00	0.00	1.00	0.00
5	0.00	0.00	0.00	0.00	0.39

$\times$

V^T	d_1	d_2	d_3	d_4	d_5	d_6
1	−0.75	−0.28	−0.20	−0.45	−0.33	−0.12
2	−0.29	−0.53	−0.19	0.63	0.22	0.41
3	0.28	−0.75	0.45	−0.20	0.12	−0.33
4	0.00	0.00	0.58	0.00	−0.58	0.58
5	−0.53	0.29	0.63	0.19	0.41	−0.22

Zusammenfassung: SVD

- Wir haben die Wort-Dokument Matrix C in das Produkt der drei Matrizen U, Σ, V^T zerlegt
- Die Wort Matrix U - besteht aus einem Reihen-Vektor für jedes Wort
- Die Dokumenten Matrix V^T besteht aus einem Spalten-Vektor für jedes Dokument
- Die Singulärwert Matrix Σ besteht aus einer diagonalen Matrix mit Singulärwerten, welche die Wichtigkeit jeder Dimension reflektieren
- Warum tun wir dies?