

K-Nearest Neighbors in Empfehlungssystemen

- eine Literaturrecherche

Jan Arends

Hochschule Bonn-Rhein-Sieg

08. Juli 2021



Inhalt

- 1 Einleitung
 - k-Nearest Neighbors
 - Empfehlungssysteme

Inhalt

1 Einleitung

- k-Nearest Neighbors

- Empfehlungssysteme

2 Data Mining mit kNN

- Abstandsmessung

- Klassifizierung vs. Regression

- Einführung Beispiel

- User-based Recommendation

- Item-based Recommendation

- Item-based vs. User-based

Inhalt

- 1 Einleitung
 - k-Nearest Neighbors
 - Empfehlungssysteme
- 2 Data Mining mit kNN
 - Abstandsmessung
 - Klassifizierung vs. Regression
 - Einführung Beispiel
 - User-based Recommendation
 - Item-based Recommendation
 - Item-based vs. User-based
- 3 Ausblick

Inhalt

- 1 Einleitung
 - k-Nearest Neighbors
 - Empfehlungssysteme
- 2 Data Mining mit kNN
 - Abstandsmessung
 - Klassifizierung vs. Regression
 - Einführung Beispiel
 - User-based Recommendation
 - Item-based Recommendation
 - Item-based vs. User-based
- 3 Ausblick
- 4 Fazit

1 Einleitung

k-Nearest Neighbors

Empfehlungssysteme

2 Data Mining mit kNN

Abstandsmessung

Klassifizierung vs. Regression

Einführung Beispiel

User-based Recommendation

Item-based Recommendation

Item-based vs. User-based

3 Ausblick

4 Fazit

k-Nearest Neighbors

Basics

- Supervised Learning Algorithm
- Keine Lernphase sondern *direkte* Berechnung \Rightarrow Lazy Learner
- Datensatz im Arbeitsspeicher \Rightarrow ggf. teuer
- Kein Modell muss gewartet werden \Rightarrow Verfahren ist agil

k-Nearest Neighbors

Basics

- Supervised Learning Algorithm
- Keine Lernphase sondern *direkte* Berechnung \Rightarrow Lazy Learner
- Datensatz im Arbeitsspeicher \Rightarrow ggf. teuer
- Kein Modell muss gewartet werden \Rightarrow Verfahren ist agil

Verfahren

- 1 Wähle k
- 2 Berechne Distanzen zu allen anderen Datenpunkten
- 3 Bestimme k nächstgelegene Nachbarn
- 4 Bestimme anhand der Nachbarn die Klasse bzw. den Wert

Veranschaulichung Klassifizierungen

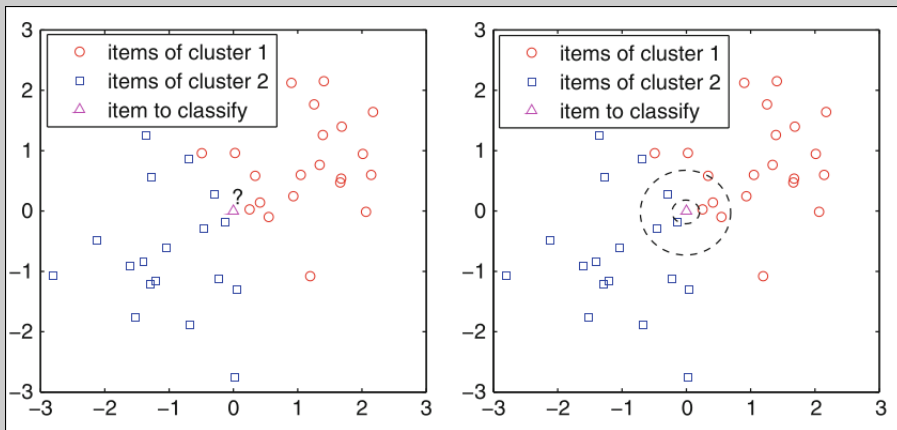


Figure: Veranschaulichung von kNN

1 Einleitung

k-Nearest Neighbors

Empfehlungssysteme

2 Data Mining mit kNN

Abstandsmessung

Klassifizierung vs. Regression

Einführung Beispiel

User-based Recommendation

Item-based Recommendation

Item-based vs. User-based

3 Ausblick

4 Fazit

Motivation

Seitens des Service Providers

- Anzahl der verkauften Artikel erhöhen
- Die Diversität von verkauften Artikeln erhöhen
- Kundenzufriedenheit & Benutzerfreundlichkeit erhöhen
- Benutzer besser verstehen

Motivation

Seitens des Service Providers

- Anzahl der verkauften Artikel erhöhen
- Die Diversität von verkauften Artikeln erhöhen
- Kundenzufriedenheit & Benutzerfreundlichkeit erhöhen
- Benutzer besser verstehen

Seitens des Service Users

- Gute oder bewährte Items finden
- Ganze Serien oder Gruppen von Items finden
- Das Stöbern (also ohne Ziel etwas zu kaufen oder konsumieren) erleichtern
- Profil ergänzen mit Sachen die der User mag oder nicht mag

Definition

The recommendation problem can be defined as estimating the response of a user for new items, based on historical information stored in the system, and suggesting to this user novel and original items for which the predicted response is high (filtering).

Definition

The recommendation problem can be defined as estimating the response of a user for new items, based on historical information stored in the system, and suggesting to this user novel and original items for which the predicted response is high (filtering).

User-Item Response

- Bewertungen, eng. *Ratings*
- Explizites Feedback, z.B. 1-5 Sterne, Likes
- "Leider" jedoch nicht immer vorhanden
- Impliziertes Feedback, z.B. Bestellhistorie, Access Pattern
- I.d.R. reichlich vorhanden

Ansätze (2 von 4)

content-based filtering

- Traditioneller Ansatz
- Empfehlungen von ähnlichen Items
- Fokus auf konkrete Produkteigenschaften
- Nachbarn sind Items

collaborative filtering

- State of the Art
- Fokus auf Bewertungen
- Nachbarn sind Nutzer (user-based)
- oder Items (item-based)

1 Einleitung

k-Nearest Neighbors

Empfehlungssysteme

2 Data Mining mit kNN

Abstandsmessung

Klassifizierung vs. Regression

Einführung Beispiel

User-based Recommendation

Item-based Recommendation

Item-based vs. User-based

3 Ausblick

4 Fazit

Euclidean Distance

Euclidean Distance

$$d(x, y) = \sqrt{\sum_{k=1}^n (x_k - y_k)^2} \quad (1)$$

Dabei repräsentiert n die Anzahl an Attributen (also der Dimensionen) und x_k und y_k jeweils die k -ten Attribute [2].

Cosine Distance und Cosine Similarity

- Wert zwischen -1 und 1
- $1 - \text{cosine-sim.} = \text{cosine-distance}$, $1 \Rightarrow x = y$

Cosine Similarity

$$\cos(x, y) = \frac{x \bullet y}{||x|| ||y||} = \cos(\theta) \quad (2)$$

Dabei repräsentiert \bullet das Skalarprodukt der Vektoren, $||x||$ die Norm des Vektors [2] und θ den Winkel der zwei Punkte.

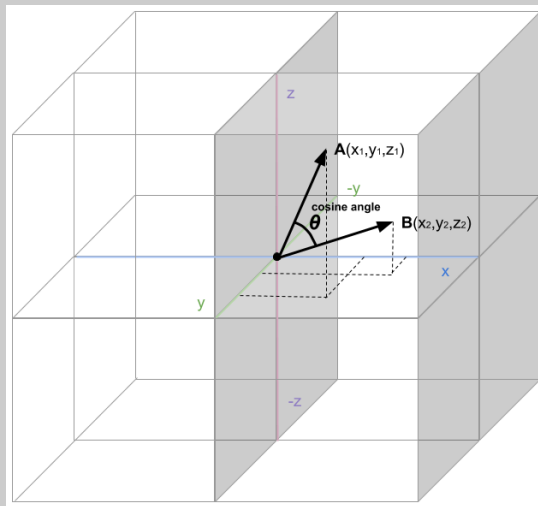


Figure: Geometrische Darstellung der Cosine Similarity [18]

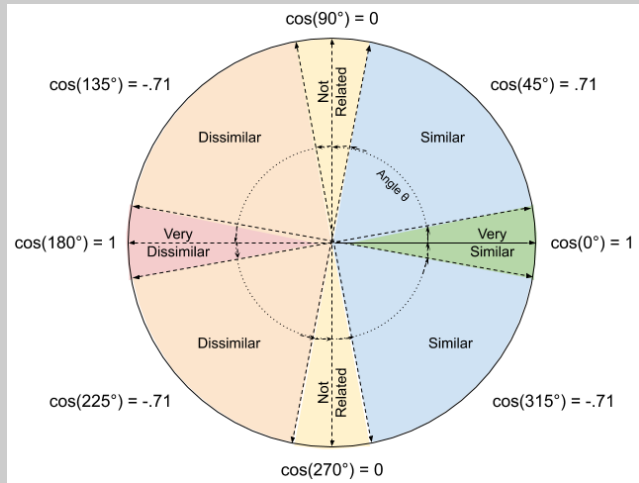


Figure: Interpretation des Ergebnisses der Cosine Similarity [18]

Korrelation

Gegeben sei die Kovarianz der Datenpunkte x und y und ihre Standardabweichung σ .

Pearson correlation coefficient

$$\text{Pearson}(x, y) = \frac{\sum(x, y)}{\sigma_x \times \sigma_y} \quad (3)$$

Matrixdarstellung

Egal wie man den Abstand misst, die Ergebnis werden typischerweise in Matrixform festgehalten.

Beispiel

	John	Lucy	Eric	Diane
John	1,000	-0,938	-0.839	0.659
Lucy	-0.938	1.000	0.922	-0.787
Eric	-0.839	0.922	1.000	-0.659
Diane	0.659	-0.787	-0.659	1,000

Table: Beispielhafte Ähnlichkeitswerte [3]

1 Einleitung

k-Nearest Neighbors

Empfehlungssysteme

2 Data Mining mit kNN

Abstandsmessung

Klassifizierung vs. Regression

Einführung Beispiel

User-based Recommendation

Item-based Recommendation

Item-based vs. User-based

3 Ausblick

4 Fazit

Klassifizierung vs. Regression

⇒ je nach Bewertungsskala:

Klassifizierung

- diskrete Werte
- beispielsweise "gut" oder "schlecht"

Regression

- kontinuierliche Zahl
- beispielsweise ein beliebiger Wert zwischen -10 und 10

1 Einleitung

k-Nearest Neighbors

Empfehlungssysteme

2 Data Mining mit kNN

Abstandsmessung

Klassifizierung vs. Regression

Einführung Beispiel

User-based Recommendation

Item-based Recommendation

Item-based vs. User-based

3 Ausblick

4 Fazit

Beispiel: Filmempfehlung

Bestimme Rating von Eric

	Matrix	Titanic	Die Hard	F. Gump	Wall-E
John	5	1		2	2
Lucy	1	5	2	5	5
Eric	2	?	3	5	4
Diane	4	3	5	3	

Table: Beispiel Datensatz

Wir wählen $k=2$.

1 Einleitung

k-Nearest Neighbors

Empfehlungssysteme

2 Data Mining mit kNN

Abstandsmessung

Klassifizierung vs. Regression

Einführung Beispiel

User-based Recommendation

Item-based Recommendation

Item-based vs. User-based

3 Ausblick

4 Fazit

Regression

Notationen

Seien

- r_{ui} die Bewertungen eines Nutzers u für ein Item i
- w_{uv} der Ähnlichkeitswert zwischen u und v
- $\mathcal{N}(u)$ k -nächstgelegenen Nachbarn von u
- $\mathcal{N}_i(u)$ k Nachbarn, die Item i bewertet haben

Regression

Notationen

Seien

- r_{ui} die Bewertungen eines Nutzers u für ein Item i
- w_{uv} der Ähnlichkeitswert zwischen u und v
- $\mathcal{N}(u)$ k -nächstgelegenen Nachbarn von u
- $\mathcal{N}_i(u)$ k Nachbarn, die Item i bewertet haben

Durchschnittliche Bewertung für i

$$\hat{r}_{ui} = \frac{1}{|\mathcal{N}_i(u)|} \sum_{v \in \mathcal{N}_i(u)} r_{vi} \quad (4)$$

Regression II

Besser:

Bewertung von i anhand der Gewichtungen: der Ähnlichkeitswerte

$$\hat{r}_{ui} = \frac{\sum_{v \in \mathcal{N}_i(u)} w_{uv} r_{vi}}{\sum_{v \in \mathcal{N}_i(u)} |w_{uv}|} \quad (5)$$

Beispiel Regression

	Matrix	Titanic	Die Hard	F. Gump	Wall-E
John	5	1		2	2
Lucy	1	5	2	5	5
Eric	2	?	3	5	4
Diane	4	3	5	3	

Table: Beispiel einer user-based recommendation

Beispiel Regression

	Matrix	Titanic	Die Hard	F. Gump	Wall-E
John	5	1		2	2
Lucy	1	5	2	5	5
Eric	2	?	3	5	4
Diane	4	3	5	3	

Table: Beispiel einer user-based recommendation

Angenommen Lucy hat einem Ähnlichkeitswert von 0,75 und Diane von 0,15.

Berechnung der Bewertung

$$\hat{r}_{\text{Eric, Titanic}} = \frac{0,75 * 5 + 0,15 * 3}{0,75 + 0,15} \simeq 4,67 \quad (6)$$

Klassifizierung

Abstimmung v_{ir} für die Bewertung r (diesmal ein konkreter Wert aus einer Menge δ), kann wie folgt durchgeführt werden.

Abstimmung

$$v_{ir} = \sum_{v \in N_i(u)} \delta(r_{vi} = r) w_{uv} \quad (7)$$

Dabei ist $\delta(r_{vi} = r) = 1$, falls $r_{vi} = r$ und ansonsten 0.

Beispiel Klassifizierung

Eine Stimme für 5 von Lucy und eine Stimme für 3 von Diane.

$$\begin{aligned}v_{\text{titanic},3} &= \delta(5 = 3) * 0,75 + \delta(3 = 3) * 0,15 \\ &= 0 * 0,75 + 1 * 0,15 = 0,15\end{aligned}$$

$$\begin{aligned}v_{\text{titanic},4} &= \delta(5 = 4) * 0,75 + \delta(3 = 4) * 0,15 \\ &= 0 * 0,75 + 0 * 0,15 = 0\end{aligned}$$

$$\begin{aligned}v_{\text{titanic},5} &= \delta(5 = 5) * 0,75 + \delta(3 = 5) * 0,15 \\ &= 1 * 0,75 + 0 * 0,15 = 0,75\end{aligned}$$

Da Lucy einen höheren Ähnlichkeitswert zu Eric hat als Diane, gewinnt Lucys Vote.

1 Einleitung

k-Nearest Neighbors

Empfehlungssysteme

2 Data Mining mit kNN

Abstandsmessung

Klassifizierung vs. Regression

Einführung Beispiel

User-based Recommendation

Item-based Recommendation

Item-based vs. User-based

3 Ausblick

4 Fazit

Regression

Seien $\mathcal{N}_u(i)$ die Nachbarn eines Items i , für welche ein Nutzer u Bewertungen abgegeben hat.

Vorhersage

$$\hat{r}_{ui} = \frac{\sum_{j \in \mathcal{N}_u(i)} w_{ij} r_{uj}}{\sum_{j \in \mathcal{N}_u(i)} |w_{ij}|} \quad (8)$$

Beispiel Regression

	Matrix	Titanic	Die Hard	F. Gump	Wall-E
John	5	1		2	2
Lucy	1	5	2	5	5
Eric	2	?	3	5	4
Diane	4	3	5	3	

Table: Table to test captions and labels

Beispiel Regression

	Matrix	Titanic	Die Hard	F. Gump	Wall-E
John	5	1		2	2
Lucy	1	5	2	5	5
Eric	2	?	3	5	4
Diane	4	3	5	3	

Table: Table to test captions and labels

Angenommen, die nächstgelegenen Nachbarn seien Forrest Gump mit einem Ähnlichkeitswert von 0,85 und Wall-E mit einem Ähnlichkeitswert von 0,75.

$$\hat{r}_{Eric, Titanic} = \frac{0,85 * 5 + 0,75 * 4}{0,85 + 0,75} \simeq 4,53 \quad (9)$$

Item-based vs. User-based

- Anzahl User größer als Anzahl Item \Rightarrow Item-based filtering (genauer und effizienter)
- Items eher statisch \Rightarrow Item-based für mehr Stabilität, anderenfalls content-based (z.B. Nachrichten)
- Für gute Anpassbarkeit \Rightarrow Item-based. Benutzer kann interaktiv am Empfehlungsprozess teilnehmen
- user-based filtering für durchwachsenere Empfehlungen

Klassifizierung

Analog zum klassifizieren bei content-based classification.

Abstimmung bei item-based classification

$$v_{ir} = \sum_{j \in N_u(i)} \delta(r_{uj} = r) w_{ij} \quad (10)$$

Klassifizierung Beispiel

Eine Stimme für 5 und eine Stimme für 4.

$$\begin{aligned}v_{\text{titanic},3} &= \delta(5 = 3) * 0,85 + \delta(4 = 3) * 0,75 \\ &= 0 * 0,85 + 0 * 0,75 = 0\end{aligned}$$

$$\begin{aligned}v_{\text{titanic},4} &= \delta(5 = 4) * 0,85 + \delta(4 = 4) * 0,75 \\ &= 0 * 0,85 + 1 * 0,75 = 0,75\end{aligned}$$

$$\begin{aligned}v_{\text{titanic},5} &= \delta(5 = 5) * 0,85 + \delta(3 = 5) * 0,75 \\ &= 1 * 0,85 + 0 * 0,75 = 0,85\end{aligned}$$

5 gewinnt mit 0,85.

1 Einleitung

k-Nearest Neighbors

Empfehlungssysteme

2 Data Mining mit kNN

Abstandsmessung

Klassifizierung vs. Regression

Einführung Beispiel

User-based Recommendation

Item-based Recommendation

Item-based vs. User-based

3 Ausblick

4 Fazit

Ausblick

- Verbesserungen, Erweiterungen und Varianten sind zu erwarten
- Mehr und mehr hybride Ansätze wurden veröffentlicht
- kNN wird weiterhin vertreten bleiben
- Ressourcenbedarf gilt es noch zu optimieren

1 Einleitung

k-Nearest Neighbors
Empfehlungssysteme

2 Data Mining mit kNN

Abstandsmessung
Klassifizierung vs. Regression
Einführung Beispiel
User-based Recommendation
Item-based Recommendation
Item-based vs. User-based

3 Ausblick

4 Fazit

Fazit

Stellenwert

- kNN hat einen enormen Stellenwert bei Empfehlungssystemen
- Nachbarschaftsbasierte Verfahren sind gängig bei content-based und collaborative filtering

Fazit

Stellenwert





- kNN hat einen enormen Stellenwert bei Empfehlungssystemen
- Nachbarschaftsbasierte Verfahren sind gängig bei content-based und collaborative filtering





Grenzen




- Probleme bei Skalierung
- Da alle Distanzen immer neu berechnet werden müssen, kann kNN sehr Ressourcen-intensiv sein
- Abstandsmessung ist ausschlaggebend für die Performance
- Aber: Durch die Wahl der richtige Methode können die Probleme adressiert werden




The End





Fragen?

-  Ricci F., Rokach L., Shapira B. (2015) Recommender Systems: Introduction and Challenges. In: Ricci F., Rokach L., Shapira B. (eds) Recommender Systems Handbook. Springer, Boston, MA. https://doi.org/10.1007/978-1-4899-7637-6_1
-  Amatriain X., Jaimes* A., Oliver N., Pujol J.M. (2011) Data Mining Methods for Recommender Systems. In: Ricci F., Rokach L., Shapira B., Kantor P. (eds) Recommender Systems Handbook. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-85820-3_2
-  Ning X., Desrosiers C., Karypis G. (2015) A Comprehensive Survey of Neighborhood-Based Recommendation Methods. In: Ricci F., Rokach L., Shapira B. (eds) Recommender Systems Handbook. Springer, Boston, MA. https://doi.org/10.1007/978-1-4899-7637-6_2
-  Koren Y., Bell R. (2015) Advances in Collaborative Filtering. In: Ricci F., Rokach L., Shapira B. (eds) Recommender Systems Handbook. Springer, Boston, MA. https://doi.org/10.1007/978-1-4899-7637-6_3

-  R. Burke, "Hybrid Web Recommender Systems", 2007, The Adaptive Web: Methods and Strategies of Web Personalization, pp. 377–408, Springer Berlin Heidelberg, doi: 10.1007/978-3-540-72079-9_12
-  Jannach D., Zanker M., Ge M., Gröning M. (2012) "Recommender Systems in Computer Science and Information Systems – A Landscape of Research", Huemer C., Lops P. (eds) E-Commerce and Web Technologies. EC-Web 2012. Lecture Notes in Business Information Processing, vol 123. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-32273-0_7.
-  G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions," in IEEE Transactions on Knowledge and Data Engineering, vol. 17, no. 6, pp. 734-749, June 2005, doi: 10.1109/TKDE.2005.99.
-  R. Burke, "Hybrid Recommender Systems: Survey and Experiments", 2002, User Modeling and User-Adapted Interaction, pp. 331-370, doi: 10.1023/A:1021240730564

-  F. Ricci, "Recommender Systems: Models and Techniques", 2017 Springer New York, Encyclopedia of Social Network Analysis and Mining, doi: 10.1007/978-1-4614-7163-9_88-1
-  V. W. Anelli, T. Di Noia, E. Di Sciascio, A. Ragone, and J. Trotta. 2019. "The importance of being dissimilar in recommendation". In Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing (SAC '19). Association for Computing Machinery, New York, NY, USA, 816–821. DOI:<https://doi.org/10.1145/3297280.3297360>.
-  J. Misztal-Radecka and B. Indurkha, 2020, "Getting to Know Your Neighbors (KYN). Explaining Item Similarity in Nearest Neighbors Collaborative Filtering Recommendations." In Adjunct Publication of the 28th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '20 Adjunct). Association for Computing Machinery, New York, NY, USA, 59–64. DOI:<https://doi.org/10.1145/3386392.3397599>.

-  A. Sagdic, C. Tekinbas, E. Arslan and T. Kucukyilmaz, "A Scalable K-Nearest Neighbor Algorithm for Recommendation System Problems," 2020 43rd International Convention on Information, Communication and Electronic Technology (MIPRO), 2020, pp. 186-191, doi: 10.23919/MIPRO48935.2020.9245195.
-  J. Sanz-Cruzado, P. Castells, and E. López. 2019. "A simple multi-armed nearest-neighbor bandit for interactive recommendation". In Proceedings of the 13th ACM Conference on Recommender Systems (RecSys '19). Association for Computing Machinery, New York, NY, USA, 358–362.
DOI:<https://doi.org/10.1145/3298689.3347040>.
-  M. Ludewig, I. Kamehkhosh, N. Landia, and D. Jannach. 2018. "Effective Nearest-Neighbor Music Recommendations". In Proceedings of the ACM Recommender Systems Challenge 2018 (RecSys Challenge '18). Association for Computing Machinery, New York, NY, USA, Article 3, 1–6.
DOI:<https://doi.org/10.1145/3267471.3267474>.

-  C. Yang, T. Liu, L. Liu and X. Chen, "A Nearest Neighbor Based Personal Rank Algorithm for Collaborator Recommendation," 2018 15th International Conference on Service Systems and Service Management (ICSSSM), 2018, pp. 1-5, doi: 10.1109/ICSSSM.2018.8465112.
-  F. Tempola, A. Arief and M. Muhammad, "Combination of case-based reasoning and nearest neighbour for recommendation of volcano status," 2017 2nd International conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE), 2017, pp. 348-352, doi: 10.1109/ICITISEE.2017.8285525.
-  B. Li, S. Wan, H. Xia and F. Qian, "The Research for Recommendation System Based on Improved KNN Algorithm," 2020 IEEE International Conference on Advances in Electrical Engineering and Computer Applications(AEECA), 2020, pp. 796-798, doi: 10.1109/AEECA49918.2020.9213566.
-  Don Cowan, Online-Resource, 2021, <https://www.ml-science.com/cosine-similarity>