Worte als Vektoren, Vektorräume, Normen

Katja Markert

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

May 6, 2019

Bisher und heute

1. Bisher: Assoziationmaße für Kollokationen: Kookkurrenz, Signifikanztests, (P)PMI
2. Heute: Darstellung eines Wortes als Vektor der Assoziationsmaße zu anderen Wörtern: sparse embeddings
3. Heute: Wiederholung von Standardmathe (Vektorräume, Vektoroperationen, Normen und Metriken)

Übersicht

Übersicht

1. Worte als Vektoren
2. Kurze Wiederholung Vektoren und Vektoroperationen
3. Normen
4. Zusammenfassung

Bisher: Assoziationsmaß als Funktion *Assoc* von der Menge aller Bigramme zu $\mathbb{R}$

Assoc	$assoc(w_1, w_2)$
$assoc_{bin}$	1, wenn mindestens einmal $w_1 w_2$ in Korpus, 0 sonst
$assoc_{freq}$	$f(w_1, w_2)$
$assoc_{cond}$	$p(w_2 \| w_1) = f(w_1, w_2)/f(w_1)$
$assoc_{pmi}$	$\log \frac{p(w_1, w_2)}{p(w_1) \cdot p(w_2)}$
$assoc_{ttest}$	$\frac{\bar{x} - \mu}{\sqrt{\frac{\sigma^2}{N}}}$
$assoc_{\chi^2}$	$\sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$

Implizit: Die benutzten Frequenzen und Wahrscheinlichkeiten beziehen sich auf die Relation rel_{bigram}. Beispiele:

- Binär: das Bigram $w_1 w_2$ erscheint oder nicht
- Häufigkeit $f(w_1, w_2)$ ist Häufigkeit des Bigrams $w_1 w_2$
- ...

Von Kollokationen zu Embeddings I

Bisher: Assoziationsmaß als Funktion *Assoc* von der Menge aller Bigramme zu $\mathbb{R}$

Assoc	$assoc(w_1, w_2)$
$assoc_{bin}$	1, wenn mindestens einmal $w_1 w_2$ in Korpus, 0 sonst
$assoc_{freq}$	$f(w_1, w_2)$
$assoc_{cond}$	$p(w_2 \mid w_1) = f(w_1, w_2)/f(w_1)$
$assoc_{pmi}$	$\log \frac{p(w_1, w_2)}{p(w_1) \cdot p(w_2)}$
$assoc_{ttest}$	$\frac{\bar{x} - \mu}{\sqrt{\frac{\sigma^2}{N}}}$
$assoc_{\chi^2}$	$\sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$

Implizit: Die benutzten Frequenzen und Wahrscheinlichkeiten beziehen sich auf die Relation rel_{bigram}. Beispiele:

- Binär: das Bigram $w_1 w_2$ erscheint oder nicht
- Häufigkeit $f(w_1, w_2)$ ist Häufigkeit des Bigrams $w_1 w_2$
- ...

Von Kollokationen zu Embeddings II

Man kann dies auf andere Relationen erweitern. Beispiel: rel_{win5}

- Binär: w_1 und w_2 erscheinen in einem Fenster von 5 Tokens zusammen oder nicht
- Häufigkeit $f(w_1, w_2)$ ist Häufigkeit des gemeinsamen Vorkommens in einem 5-Token Fenster
- ...

Assoc	$assoc(w_1, w_2)$
$assoc_{bin}$	1, wenn mindestens einmal w_1, w_2 in Korpus, 0 sonst
$assoc_{freq}$	$f(w_1, w_2)$
$assoc_{cond}$	$p(w_2 \mid w_1) = f(w_1, w_2)/f(w_1)$
$assoc_{pmi}$	$\log \frac{p(w_1, w_2)}{p(w_1) \cdot p(w_2)}$
$assoc_{ttest}$	$\frac{\bar{x} - \mu}{\sqrt{\frac{\sigma^2}{N}}}$
$assoc_{\chi^2}$	$\sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$

rel_{bigram} ist eine geordnete Version von rel_{win1}

Von Kollokationen zu Embeddings III

Wir wählen eine Relation *rel* und ein Assoziationsmaß *assoc*. Wir können nun ein Wort w auch als Vektor der Assoziationsmaße zu anderen Wörtern $w_1, \ldots w_n$ eines Vokabulars *Voc* darstellen.

$$\vec{w} = (assoc(w, w_1), assoc(w, w_2), assoc(w, w_3) \ldots, assoc(w, w_n))$$

Hiermit haben wir ein embedding von w, da wir w nun einen Vektor aus $\mathbb{R}^n$ zugewiesen haben!

Von Kollokationen zu Embeddings IV

Korpus BNC, Zielwort w, $V = \{species, computer, animal\}$, rel_{win10} (window 5 to the right and 5 to left of w), $A_{freq} = assoc_{freq}$

$\vec{cat} = (A_{freq}(cat, species), A_{freq}(cat, computer), A_{freq}(cat, animal)) = (59, 5, 304)$

$\vec{carnivore} = (A_{freq}(carnivore, species), A_{freq}(carnivore, computer), A_{freq}(carnivore, animal) \ldots (21, 1, 21)$

$\vec{airport} = (A_{freq}(airport, species), A_{freq}(airport, computer), A_{freq}(airport, animal)) = (4, 12, 2)$

Von Kollokationen zu Embeddings V

Solange wir für alle Vektordarstellungen das gleiche Vokabular, das gleiche Assoziationsmaß sowie die gleiche Relation nehmen, sind die embeddings verschiedener Wörter vergleichbar (da in den gleichen Vektorraum eingebettet).

Matrixschreibweise (mit Vokabular in Spalten):

	species	*computer*	*animal*
cat	59	5	304
carnivore	21	1	21
feline	2	0	5
airport	4	12	2

Hierbei immer implizit: Assoziationsmaß sowie Relation!

Beobachtungen

- Vektoren und Matrizen: Gesamte lineare Algebra steht zur Verfügung
- Insbesondere: Wir können die Abstände bzw Ähnlichkeiten zwischen zwei Wörtern als die Distanz der beiden im Vektorraum messen!

Beobachtungen

	species	*computer*	*animal*
cat	59	5	304
carnivore	21	1	21
feline	2	0	5
airport	4	12	2

- Man sieht die ähnlicheren Wortpaare sind in den gleichen Vokabulardimensionen "stark" bzw schwach
- Wenn man unbedacht vorgeht, dann könnte die Ähnlichkeit fälschlicherweise von der Worthäufigkeit (= Vektorlänge) abhängen!
- Die Matrix ist im allgemeinen hochdimensional und "sparse" (siehe Matrix in Übungsblatt 1)

Beobachtungen und Fragen

- Wie wähle ich das Vokabular? Oft die häufigsten *n* Unigramme ohne Funktionswörter. Warum?
- Wie wähle ich die Relation? (Siehe ECL)
- Was für ein Assozationsmaß wähle ich? (Diskutiert in vorherigen Vorlesungen)

Beobachtungen und Fragen

Die gleiche Matrix mit $assoc_{bin}$:

	species	*computer*	*animal*
cat	1	1	1
carnivore	1	1	1
feline	1	0	1
airport	1	1	1

- Wie bekomme ich einfach die Matrix in (P)PMI? (ECL, Wiederholung nächstes Mal)
- Wie berechne ich am besten die Vektorenähnlichkeit: jetzt!

Übersicht

(Reeller) Vektorraum: Definition

Es sei V eine Menge, $\oplus : V \times V \to V$ eine innere zweistellige Verknüpfung (Vektoraddition) und $\odot : \mathbb{R} \times V \to V$ eine äußere zweistellige Verknüpfung, genannt Skalarmultiplikation. Man nennt $(V, \oplus, \odot)$ einen **reellen Vektorraum**, wenn gilt für alle $\alpha, \beta \in \mathbb{R}$ und $u, v, w \in V$

1. $u \oplus (v \oplus w) = (u \oplus v) \oplus w$(Assoziativgesetz)
2. Existenz eines neutralen Elements (Nullvektors) $0_V \in V$ mit $v \oplus 0_V = 0_V \oplus v = v$
3. Existenz eines zu $v \in V$ inversen Elements $-v \in V$ mit $v \oplus (-v) = (-v) \oplus v = 0_V$
4. $v \oplus u = u \oplus v$ (Kommutativgesetz)

sowie

1. $\alpha \odot (u \oplus v) = (\alpha \odot u) \oplus (\alpha \odot v)$
2. $(\alpha + \beta) \odot v = (\alpha \odot v) \oplus (\beta \odot v)$
3. $(\alpha \cdot \beta) \odot v = \alpha \odot (\beta \odot v)$
4. $1 \odot v = v$

Notationen

Wir werden von nun an $\oplus$ einfach als $+$ schreiben, obwohl es sich hier um eine Vektoraddition handelt. Ebenso $\odot$ als $\cdot$.

Je nach Ambiguität schreiben wir Vektoren auch mit einem Pfeil $\vec{v}$.

Beispiel: $\mathbb{R}^n$

Der altbekannte $\mathbb{R}^n := \{(v_1, \ldots, v_n) | v_1, \ldots v_n \in \mathbb{R}\}$ als Menge aller n-Tupel über $\mathbb{R}$ bildet einen reellen Vektorraum, wenn man definiert:

1. Vektoraddition zwischen $u = (u_1, u_2, \ldots, u_n), v = (v_1, v_2, \ldots, v_n) \in \mathbb{R}^n$: $u + v = (u_1, u_2, \ldots, u_n) + (v_1, v_2, \ldots, v_n) := (u_1 + v_1, u_2 + v_2, \ldots, u_n + v_n)$
2. Skalarmultiplikation zwischen $\alpha \in \mathbb{R}$ und $v = (v_1, v_2, \ldots, v_n) \in \mathbb{R}^n$: $\alpha \cdot v = \alpha \cdot (v_1, v_2, \ldots, v_n) := (\alpha v_1, \alpha v_2, \ldots, \alpha v_n)$

Er wird auch **Koordinatenraum** gennant (mit elementweiser Addition und Skalarmultiplikation).

1. Neutrales Element (Nullvektor) = $(0, \ldots, 0)$
2. Inverses Element zu $v = (v_1, \ldots v_n)$ ist $v = (-v_1, \ldots, -v_n)$
3. Beispiel in der euklidischen Ebene $\mathbb{R}^2$

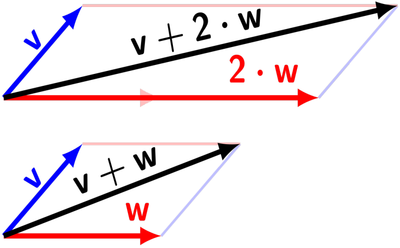


Bild von Martin Thoma, CC BY 3.0,
https://commons.wikimedia.org/w/index.php?curid=21388478

- **Komponenten** eines Vektores: *i*-te Komponente: $v_1, \ldots \mathbf{v_i}, \ldots v_n$
- Anzahl der Komponenten = **Vektordimensionalität**
- Man kann Vektoren auch vertikal schreiben:

$$v = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix} \tag{1}$$

- Andere reelle Vektorräume: Matrizenräume (nächste Woche), reelle Polynome ...

Übersicht

Norm Definition

Idee

Ordne einem Element eines Vektorraums eine Größe zu.

Eine **Norm** ist definiert als eine Abbildung:

$$\|\cdot\|: V \to \mathbb{R}_0^+, v \mapsto \|v\|$$

mit den folgenden Eigenschaften für alle $v, w \in V$ und $\alpha \in \mathbb{R}$:

1. Definitheit: $\|v\| = 0 \Leftrightarrow v = 0$
2. absolute Homogenität: $\|\alpha \cdot v\| = |\alpha| \cdot \|v\|$
3. Subadditivität oder Dreiecksungleichung: $\|v + w\| \leq \|v\| + \|w\|$

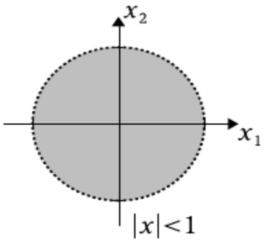
1. $\|-v\| = |-1| \cdot \|v\| = \|v\|$
2. Aus Dreiecksungleichung:
   $\|\vec{0}\| = \|v + (-v)\| \leq \|v\| + \|-v\| = 2\|v\|$ und damit: Normen sind immer nicht-negativ

Verschiedene Normen auf dem $\mathbb{R}^n$

Die **euklidische Norm** oder L_2-Norm:

$$\|v\|_2 := \sqrt{\sum_{i=1}^{n} v_i^2}$$

- Wir überlegen uns, warum dies die Normaxiome erfüllt (Tafel)
- Beispiel im $\mathbb{R}^3$: $\|(1, -2, 1)\|_2 = \sqrt{1^2 + 2^2 + 1^2} = \sqrt{6}$
- Offene Einheitskugel: *Menge aller v mit* $\|v\|_2 < 1$



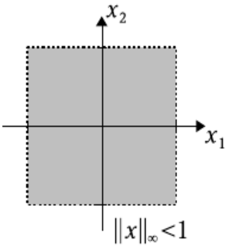
(b)

Verschiedene Normen auf dem $\mathbb{R}^n$

Maximumsnorm oder Unendlich-Norm oder Tschebyschev-Norm

$$\|v\|_\infty := max_{i=1,\dots n}|v_i|$$

- Wir überlegen uns, warum dies die Normaxiome erfüllt (Tafel)
- Beispiel im $\mathbb{R}^3$: $\|(1,-2,1)\|_\infty = 2$
- Offene Einheitskugel: Menge aller v mit $\|v\|_\infty < 1$

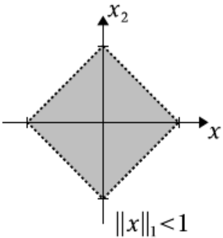


Verschiedene Normen auf dem $\mathbb{R}^n$

Summennorm oder L_1-Norm:

$$\|v\|_1 = \sum_{i=1}^{n} |v_i|$$

- Wir überlegen uns, warum dies die Normaxiome erfüllt (Tafel)
- Beispiel im $\mathbb{R}^3$: $\|(1, -2, 1)\|_1 = 4$
- Offene Einheitskugel: Menge aller v mit $\|v\|_1 < 1$



(a)

Benutzung von Normen

- Normierung von Vektoren, um Längeneffekte zu vermeiden (jetzt)
- Induzierung von Metriken (jetzt)
- Normen in objective functions beim Maschinellen Lernen, als Fehlerfunktion (Vorlesung Statistische Methoden)
- Normen in objective functions als Regularisierung, um “Overfitting” zu Trainingsdaten zu minimieren (Vorlesung “Statistische Methoden”)

Vektornormierungen, Einheitsvektoren

Gegeben sei ein normierter Vektorraum und v ein Vektor, der nicht der Nullvektor ist. Wir nennen den Vektor $\frac{v}{\|v\|}$ einen **Einheitsvektor**, denn es gilt:

$$\left\|\frac{v}{\|v\|}\right\| = \left|\frac{1}{\|v\|}\right| \cdot \|v\| = 1$$

Beispiel im $\mathbb{R}^3$ Vektor $\vec{v} = (1, -2, 1)$

1. Mit der euklidischen Norm gilt $\|(1, -2, 1)\|_2 = \sqrt{6}$ und damit der von $\vec{v}$ abgeleitete Einheitsvektor $\vec{v}_e = \frac{\vec{v}}{\|\vec{v}\|_2} = (\frac{1}{\sqrt{6}}, \frac{-2}{\sqrt{6}}, \frac{1}{\sqrt{6}})$ und damit
$\|\vec{v}_e\|_2 = \sqrt{(\frac{1}{\sqrt{6}})^2 + (\frac{-2}{\sqrt{6}})^2 + (\frac{1}{\sqrt{6}})^2} = \sqrt{1/6 + 4/6 + 1/6} = 1$
2. Mit der Maximumsnorm gilt $\|(1, -2, 1)\|_\infty = 2$ und damit ist der von $\vec{v}$ abgeleitete Einheitsvektor $\vec{v}_e =$
$\frac{\vec{v}}{\|\vec{v}\|_\infty} = (\frac{1}{2}, \frac{-2}{2}, \frac{1}{2}) = (0.5, -1, 0.5)$. Damit ist $\|\vec{v}_e\|_\infty = 1$

Übersicht

Zusammenfassung

- Wir können Worte als Vektoren von Assoziationsmaßen zu anderen Wörtern auffassen $\longrightarrow$ Embeddings
- Lineare Algebra: Vektorräume und Vektoroperationen
- Normen in Vektorräumen erlauben uns, Vektoren eine Größe zuzuordnen
- Wir können Vektoren normieren

Literatur

- Jurafsky und Martin, Edition II: Kapitel 20
- Gerd Fischer: Lineare Algebra. Eine Einführung für Studienanfänger
- Serie von Videos von 3Blue1Brown. Startet hier: `https://www.youtube.com/watch?v=fNk_zzaMoSs&list=PLZHQObOWTQDPD3MizzM2xVFitg`
- Übungsblatt 1: Aufgabe 2