# Metriken und Ähnlichkeitsmaße. Evaluation von embeddings

Katja Markert

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

May 6, 2019

# Bisher und heute

1. Bisher: Darstellung eines Wortes als Vektor der Assoziationsmaße zu anderen Wörtern: sparse embeddings
2. Bisher: Vektorräume und Normen
3. Jetzt: Abstände und Ähnlichkeiten zwischen Vektoren
4. Evaluation von Embeddings via menschlicher Ähnlichkeitsannotationen

# Übersicht

# Übersicht

# Induktion einer Metrik durch eine Norm

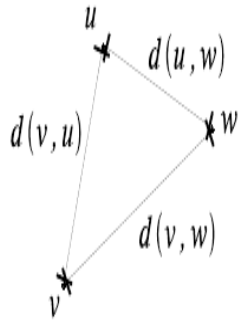
Gegeben sei ein normierter reeller Vektorraum $(V, \|\cdot\|)$. Dann kann man auf $V$ den **Abstand** zweier Vektoren $\vec{v}, \vec{w}$ wie folgt definieren:

$$d(\vec{v}, \vec{w}) := \|\vec{v} - \vec{w}\|$$

# Eigenschaften

Wir erfüllen die Axiome einer Metrik für alle $\vec{v}, \vec{w}, \vec{u} \in V$

1. Aufgrund der Definitheit der Norm, gilt $d(\vec{v}, \vec{w}) = 0$, genau dann wenn $\vec{v} = \vec{w}$
2. Symmetrie:
   $d(\vec{v}, \vec{w}) = \|\vec{v} - \vec{w}\| = \|(-1) \cdot (\vec{w} - \vec{v})\| = |-1| \cdot \|\vec{w} - \vec{v}\| = d(\vec{w}, \vec{v})$
3. Dreiecksungleichung: $d(\vec{v}, \vec{w}) \leq d(\vec{v}, \vec{u}) + d(\vec{u}, \vec{w})$




Dies gilt unabhängig von der induzierenden Norm!

# Weitere Eigenschaften

- Da Norm immer nicht-negativ, ist der Abstand zweier Vektoren immer nicht-negativ
- Die Vektorlänge ist damit der Abstand des Vektors vom Ursprung:

$$\|\vec{v}\| = \|\vec{v} - \vec{0}\| = d(\vec{v}, \vec{0})$$

# Euklidische Metrik

Die von der euklidischen Norm induzierte euklidische Metrik auf dem $\mathbb{R}^n$ ist also:

$$d_2(\vec{v}, \vec{w}) = \sqrt{\sum_{i=1}^{n} (v_i - w_i)^2}$$

Dies entspricht der geometrischen Interpretation des "Luftlinienabstands" im $\mathbb{R}^2$ (oder $\mathbb{R}^3$):

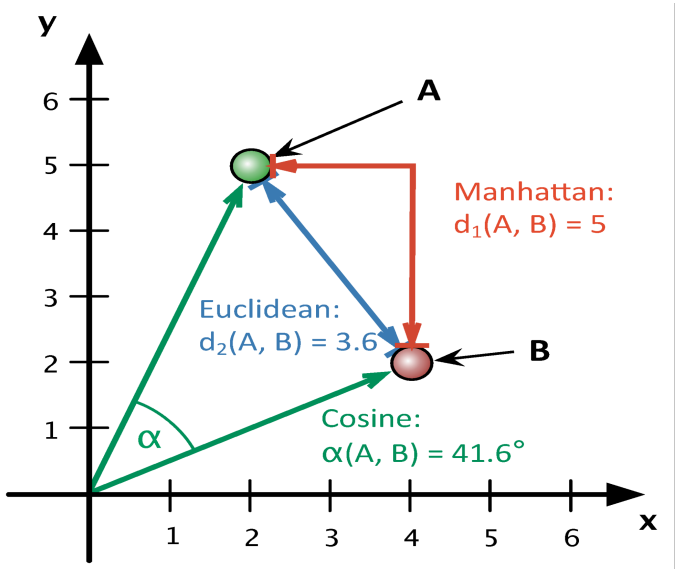



Bild von `http://dh2016.adho.org/static/data/290.html`

# Manhattan-Metrik

Die von der Summennorm induzierte Manhattan-metrik auf dem $\mathbb{R}^n$ ist also:

$$d_1(\vec{v}, \vec{w}) = \sum_{i=1}^{n} |v_i - w_i|$$

Wir laufen Umwege um einen “Block” herum:

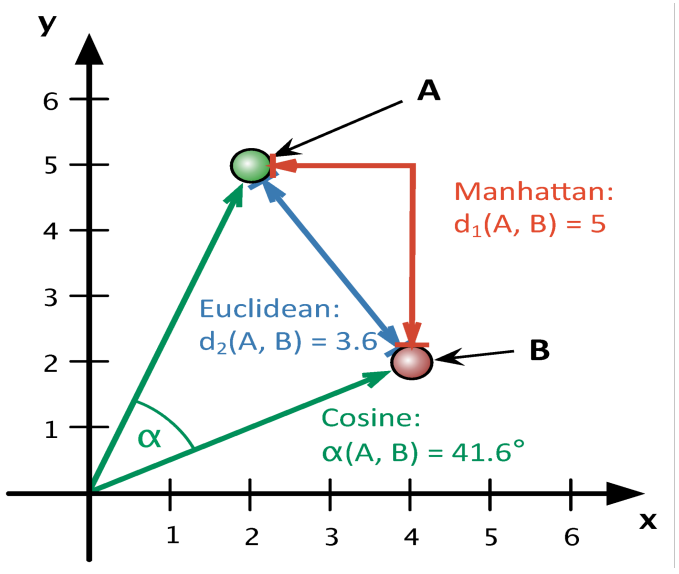



Bild von `http://dh2016.adho.org/static/data/290.html`

# Unser NLP-Beispiel

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

$d_2(cat, carnivore) = \sqrt{(59-21)^2 + (5-1)^2 + (304-21)^2} = 285$
(gerundet)

$d_2(cat, feline) = \sqrt{(59-2)^2 + (5-0)^2 + (304-5)^2} = 304$
(gerundet)

$d_2(cat, airport) = \sqrt{(59-4)^2 + (5-12)^2 + (304-2)^2} = 307$
(gerundet)

# Unser NLP-Beispiel

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

| Paar | $d_2$ (gerundet) |
|---|---|
| cat, carnivore | 285 |
| cat, feline | 304 |
| cat, airport | 307 |

Ist dies, was wir wollen? Wo liegt das Problem?

# Andere Norm?

Summennorm/Manhattanmetrik:

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

| Paar | $d_1$ |
|---|---|
| cat, carnivore | $\lvert 59-21\rvert + \lvert 5-1\rvert + \lvert 304-21\rvert = 325$ |
| cat, feline | 356 |
| cat, airport | 364 |

Dies scheint keine Lösung zu sein...

# Probleme

- Abhängigkeit von Vektorlänge = Worthäufigkeit
- Distanz deswegen auch nicht nach oben beschränkt
- Distanz anstatt Ähnlichkeit $\rightarrow$ Umwandlung in Ähnlichkeit z.B. mit $sim(v, w) = 1 - d(v, w)$ $\rightarrow$ Negative Ähnlichkeiten

Besser: Direkte Ähnlichkeitsmaße, die nicht längenabhängig sind.

- Eine Möglichkeit: Normiere Vektoren zuerst (siehe Übungsaufgabe)
- Zweite Möglichkeit: Cosine Similarity

# Übersicht

# Skalarprodukt/Dot Product: Definition und Beispiel

Sei $V$ der $\mathbb{R}^n$. Dann ist das **Skalarprodukt** zweier Vektoren definiert als eine Abbildung $\cdot : V \times V \to \mathbb{R}$ mit $\vec{v} \cdot \vec{u} := \sum_{i=1}^{n} v_i \cdot u_i$

Notation: oft auch geschrieben als $\langle \vec{v}, \vec{u} \rangle$.

Beispiel im $\mathbb{R}^3$:

$$(1,-2,1)\cdot(3,4,-1) = 1\cdot 3 + (-2)\cdot 4 + 1\cdot(-1) = 3 + (-8) + (-1) = -6$$

Bitte Skalarprodukt nicht mit Skalarmultiplikation verwechseln!

Warum definiert man das Skalarprodukt so? Weil man damit dann schön rechnen kann (siehe n”achste Folie)

# Eigenschaften des Skalarproduktes

- Das Skalarprodukt ist **symmetrisch**, also $\vec{v} \cdot \vec{w} = \vec{w} \cdot \vec{v}$ für alle Vektoren $\vec{v}, \vec{w}$
- **Gemischtes Assoziativgesetz**:
  $(a \cdot \vec{v}) \cdot \vec{w} = a \cdot (\vec{v} \cdot \vec{w}) = \vec{v} \cdot (a \cdot \vec{w})$ für alle Vektoren $\vec{v}, \vec{w}$ und alle Skalare $a \in \mathbb{R}$
- **Distributivgesetze** für alle Vektoren $\vec{u}, \vec{v}, \vec{w}$:

$$\vec{u} \cdot (\vec{v} + \vec{w}) = \vec{u} \cdot \vec{v} + \vec{u} \cdot \vec{w}$$

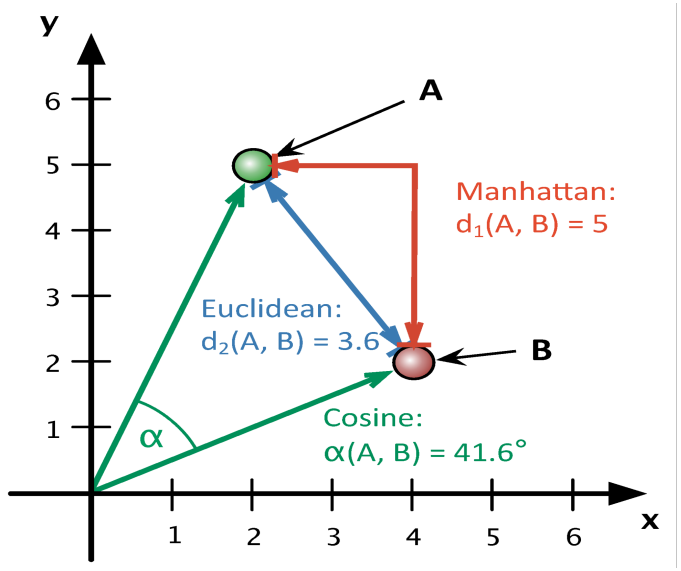
$$(\vec{u} + \vec{v}) \cdot \vec{w} = \vec{u} \cdot \vec{w} + \vec{v} \cdot \vec{w}$$

- Skalarprodukt des Vektors mit sich selbst ist Vektorlänge quadriert:

$$\vec{v} \cdot \vec{v} = \sum_{i=1}^{n} v_i^2 = \|\vec{v}\|_2^2$$

- Skalarprodukt ist keine Metrik: kann negativ sein...

# Geometrische Interpretation des Skalarprodukts




## Kosinussatz

Für zwei Vektoren $\vec{a}, \vec{b}$ gilt:

$$\|\vec{a} - \vec{b}\|_2^2 = \|\vec{a}\|_2^2 + \|\vec{b}\|_2^2 - 2\|\vec{a}\|_2\|\vec{b}\|_2 \cos\alpha$$

wobei $\alpha$ der Winkel zwischen $\vec{a}$ und $\vec{b}$ ist.

Für den Beweis verweise ich auf Schulbücher und Wikipedia-Eintrag für Skalarprodukt...



Bild von `http://dh2016.adho.org/static/data/290.html`

# Geometrische Interpretation des Skalarprodukts

Es gilt aber auch

$$\|\vec{a}-\vec{b}\|_2^2 = (\vec{a}-\vec{b})\cdot(\vec{a}-\vec{b}) = \vec{a}\cdot\vec{a}-2\cdot\vec{a}\cdot\vec{b}+\vec{b}\cdot\vec{b} = \|\vec{a}\|_2^2+\|\vec{b}\|_2^2-2\vec{a}\cdot\vec{b}$$

und daraus folgt zusammen mit dem Kosinussatz

$$\vec{a}\cdot\vec{b} = \|\vec{a}\|_2\|\vec{b}\|_2\cos\alpha$$

bzw (solange alle Vektoren nicht Null)

$$\cos\alpha = \frac{\vec{a}\vec{b}}{\|\vec{a}\|_2\|\vec{b}\|_2}$$

# Cosine Similarity: Eigenschaften

Wir erinnern uns:

$$\cos\alpha = \frac{\vec{a}\vec{b}}{\|\vec{a}\|_2\|\vec{b}\|_2}$$

Wenn wir diesen Kosinus nun als Ähnlichkeitsmaß $sim_{cos}(\vec{a},\vec{b})$ benutzen, hat dies einige schöne Eigenschaften, obwohl keine Metrik:

- Symmetrie
- Sind $\vec{a},\vec{b}$ parallel zueinander, dann ist $\alpha = 0$ und damit $sim_{cos} = 1$, unabhängig von der Vektorlänge
- Es gilt, dass $sim_{cos} = 0$ genau dann wenn $\vec{a},\vec{b}$ **orthogonal** zueinander ($\alpha = 90$ *Grad*). Allgemein: zwei Vektoren sind orthogonal, wenn ihr Skalarprodukt = 0 ist!
- Haben die Vektoren nur positive Einträge (Frequenzen, PPMI), dann ist $sim_{cos}$ zwischen Null und Eins.

# Unser NLP-Beispiel

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

$$cos_{sim}(cat, carnivore) = \frac{59 \cdot 21 + 5 \cdot 1 + 304 \cdot 21}{\sqrt{59^2 + 5^2 + 304^2} \cdot \sqrt{21^2 + 1^2 + 21^2}} = \frac{7628}{\sqrt{95922}\sqrt{883}} = 0.828$$

$$cos_{sim}(cat, feline) = \frac{59 \cdot 2 + 5 \cdot 0 + 304 \cdot 5}{\sqrt{59^2 + 5^2 + 304^2} \cdot \sqrt{2^2 + 0^2 + 5^2}} = 0.98$$

$$cos_{sim}(cat, airport) = \frac{59 \cdot 4 + 5 \cdot 12 + 304 \cdot 2}{\sqrt{59^2 + 5^2 + 304^2} \cdot \sqrt{4^2 + 12^2 + 2^2}} = 0.227$$

# Unser NLP-Beispiel

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

| Paar | $cos_{sim}$ |
|---|---|
| cat, carnivore | 0.828 |
| cat, feline | 0.98 |
| cat, airport | 0.227 |

Hurrah!

# Zusammenfassung

1. Metriken können aus Normen abgeleitet werden
2. Direkte Verwendung der euklidischen Metrik ist kein besonders gutes Ähnlichkeitsmaß für Embeddings
3. Besser: Cosine Similarity. Worte sind ähnlich, wenn ihre Embeddingsvektoren in die gleiche Richtung zeigen, unabhängig von Wortlänge
4. Weiterer Vorteil von Cosine Similarity: zwsichen Null und Eins, wenn Embeddings-einträge positiv sind
5. Es gibt noch weitere Ähnlichkeitsmaße, wie die relative entropy (Kullback-Leibner-divergence), wenn die jeder Ebeddingsvektor eine Wahrscheinlichkeitsverteilung bilden

# Übersicht

# Grundidee

## Problem

Wir haben word embeddings gebildet, sprich Wörter in einen Vektorraum eingebettet. Wie wissen wir nun, ob die embeddings *gut* sind?

Idee: Berechne Wortähnlichkeiten und vergleiche mit menschlichen Wortähnlichkeitsnormen wie WordSim353. `http://www.cs.technion.ac.il/~gabr/resources/data/wordsim353/`

| Wort1 | Wort2 | Human Rating |
|---|---|---|
| tiger | cat | 7.35 |
| tiger | tiger | 10.00 |
| drug | abuse | 6.85 |
| bread | butter | 6.19 |
| cup | coffee | 6.58 |
| cup | object | 3.69 |
| king | cabbage | 0.23 |
| king | queen | 8.58 |
| king | rook | 5.92 |

# Wir müssen also vergleichen

Mit erfundenen menschlichen Ähnlichkeiten:

| Paar | $d_2$ | $cos_{sim}$ | Mensch |
|---|---|---|---|
| cat, carnivore | 285 | 0.828 | 7 |
| cat, feline | 304 | 0.98 | 9 |
| cat, airport | 307 | 0.227 | 1 |

Welche Performanz ist besser?

Wir brauchen allgemeine Methoden, um zwei Variablen/Messreihen zu vergleichen $\rightarrow$ Korrelationen

# Übersicht

# Motivation

## Statistiken für einzelne Variablen

Mittelwerte und Varianz sind **single variable statistics**. Wir wollen nun Korrelationen zwischen Variablen messen.

## Dependenzen zwischen zwei Messreihen

- Sind Studierende, die gut in Mathematik sind, auch gut in Informatik? (Notenmessreihen)
- Korrelieren automatische Ähnlichkeitswerte zwischen Wörtern mit menschliche?

# The Preston Curve

`http://www.gapminder.org/world`

## Scatterplot

Zeigt die Beziehung zwischen zwei numerischen Variablen, deren Werte auf der gleichen Population gemessen wurden. Die *explanatory variable* befindet sich auf der *x*-Achse und die *response variable* auf der Y-Achse.

## Positive und Negative Assoziation

Positive Assoziation: Kurve nach oben. Negative Assoziation: Kurve nach unten. Nicht alle Kurven sind Geraden...

# Pearson Korrelation: Definition

**Pearson Korrelation**

Die Pearson Korrelation $r$ beschreibt die Richtung und Stärke einer Assoziation in Form einer Geraden (zwischen zwei numerischen Variablen $X$ und $Y$).

$$r(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{X})(y_i - \bar{Y})}{(n-1)\sigma_X \sigma_Y}$$

wobei $\bar{X}, \bar{Y}$ Mittelwerte und $\sigma_X, \sigma_Y$ die Standardabweichungen der Variablen sind

# Pearson Korrelation: unser NLP Beispiel

| Paar | $d_2$ | $cos_{sim}$ | M(ensch) |
|---|---|---|---|
| cat, carnivore | 285 | 0.828 | 7 |
| cat, feline | 304 | 0.98 | 9 |
| cat, airport | 307 | 0.227 | 1 |

| Maß | Mittelwert | Standardabweichung |
|---|---|---|
| $d_2$ | 298.66 | 11.93 |
| $cos_{sim}$ | 0.678 | 0.398 |
| M(ensch) | 5.66 | 4.16 |

$$r(M, d_2) = \frac{\sum_{i=1}^{3}(M_i - \bar{M})(d2_i - \bar{d2})}{(3-1)\sigma_M \sigma_{d2}} = \frac{-39.33}{2 \cdot 11.93 \cdot 4.16} = -0.3959$$

$$r(M, cos_{sim}) = 0.9987$$

# Pearson Korrelation: Eigenschaften

- Positiv, wenn Assoziation positiv. Negativ sonst.
- Immer zwischen 1 und -1.
- Symmetrisch
- Sollte nur für Geraden benutzt werden: nimmt lineare Beziehung an
- Wenige Ausreißer ruinieren das Ergebnis...

# Beware of Pearson correlations: Anscombe Quartet

**Das Anscombe-Quartett**

| I | | II | | III | | IV | |
|---|---|---|---|---|---|---|---|
| x | y | x | y | x | y | x | y |
| 10,0 | 8,04 | 10,0 | 9,14 | 10,0 | 7,46 | 8,0 | 6,58 |
| 8,0 | 6,95 | 8,0 | 8,14 | 8,0 | 6,77 | 8,0 | 5,76 |
| 13,0 | 7,58 | 13,0 | 8,74 | 13,0 | 12,74 | 8,0 | 7,71 |
| 9,0 | 8,81 | 9,0 | 8,77 | 9,0 | 7,11 | 8,0 | 8,84 |
| 11,0 | 8,33 | 11,0 | 9,26 | 11,0 | 7,81 | 8,0 | 8,47 |
| 14,0 | 9,96 | 14,0 | 8,10 | 14,0 | 8,84 | 8,0 | 7,04 |
| 6,0 | 7,24 | 6,0 | 6,13 | 6,0 | 6,08 | 8,0 | 5,25 |
| 4,0 | 4,26 | 4,0 | 3,10 | 4,0 | 5,39 | 19,0 | 12,50 |
| 12,0 | 10,84 | 12,0 | 9,13 | 12,0 | 8,15 | 8,0 | 5,56 |
| 7,0 | 4,82 | 7,0 | 7,26 | 7,0 | 6,42 | 8,0 | 7,91 |
| 5,0 | 5,68 | 5,0 | 4,74 | 5,0 | 5,73 | 8,0 | 6,89 |

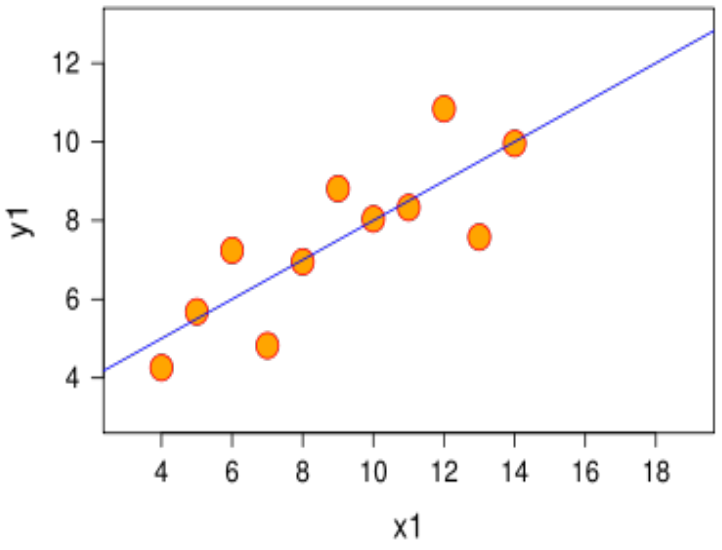
# Anscombe Quartett

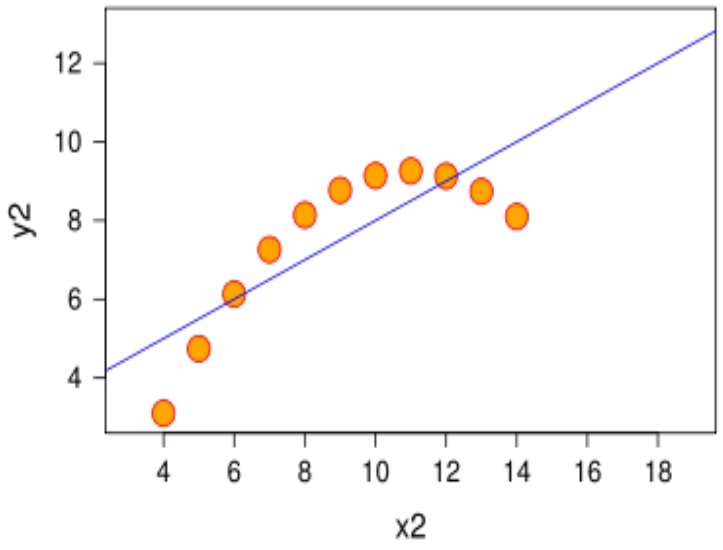
Es gilt:

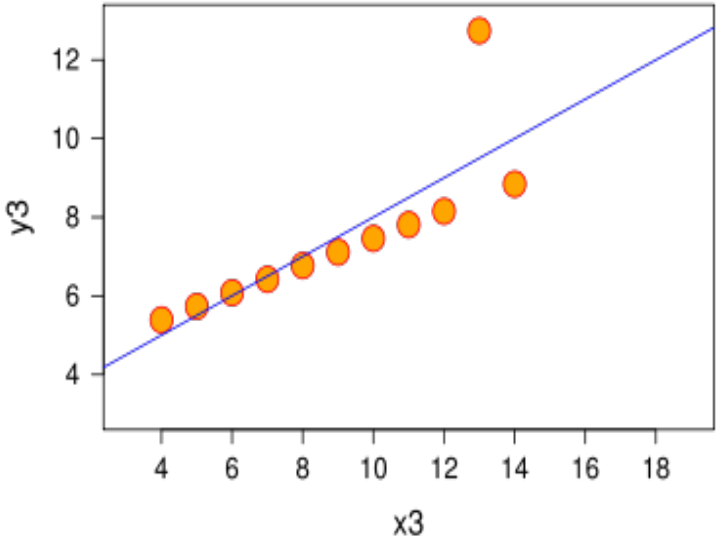
| | |
|---|---|
| Mittelwert aller X | 9 |
| Standardabweichung aller X | $\sqrt{11}$ |
| Mittelwert aller Y | 7.50 |
| Standardabweichung aller Y | $\sqrt{4.122}$ |

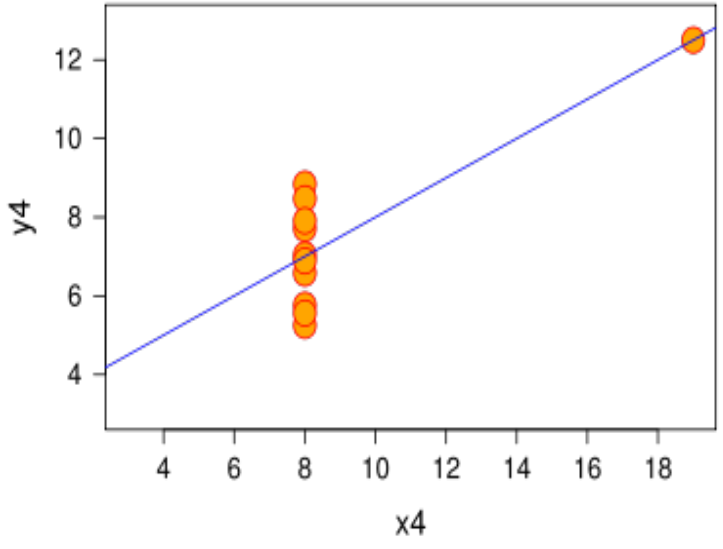
Damit Pearson-Korrelation zwischen $X_i$ und $Y_i = 0.816$ für alle $i$ von 1 bis 4.

# Anscombe Quartet



# Spearman Rank Correlation: Idee

**Idee**

Mich interessieren die genauen Werte nicht, sondern nur das Ranking. Damit sind numerische Outlier nicht mehr so wichtig. Die Beziehung muss nicht mehr linear sein.

Konvertiere die Variablen in Rankings und berechne dann auf den Rankings Pearson correlation. (Vorsicht: hier wird nach Ähnlichkeit geranked,also müssen wir bei Distanz das Ranking invertieren.)

| Paar | $d_2$ | $d_2$ Rank | $cos_{sim}$ | $cos_{sim}$ Rank | Mensch | Mensch Ran |
|---|---|---|---|---|---|---|
| cat, carnivore | 285 | 1 | 0.828 | 2 | 7 | 2 |
| cat, feline | 304 | 2 | 0.98 | 1 | 9 | 1 |
| cat, airport | 307 | 3 | 0.227 | 3 | 1 | 3 |

# Spearman Rank Correlation: Beispiel

Nur noch Ranks interessieren uns!

| Paar | $d_2$ Rank | $cos_{sim}$ Rank | Mensch Rank |
|---|---|---|---|
| cat, carnivore | 1 | 2 | 2 |
| cat, feline | 2 | 1 | 1 |
| cat, airport | 3 | 3 | 3 |

| Maß | Mittelwert | Standardabweichung |
|---|---|---|
| $d_2$ Rank | 2 | 1 |
| $cos_{sim}$ Rank | 2 | 1 |
| Mensch Rank | 2 | 1 |

Spearman Korrelation ρ zwischen $d_2$ und Mensch (= Pearson zwischen $d_2$ Ranks und Mensch-Ranks)

$$\rho(Mensch, d_2) = \frac{1}{2 \cdot 1 \cdot 1} = 0.5$$

Spearman Korrelation zwischen $cos_{sim}$ und Mensch:

$\rho(Mensch, cos_{sim}) = 1$

# Spearman Rank Correlation

- Auch zwischen 1 und -1, auch symmetrisch
- Braucht keine lineare Relation (Gerade)
- Untersucht nur Richtigkeit des Rankings
- Falls ein Wert mehrfach auftaucht, müssen wir bei der Rankingsgenerierung *fractional rankings* verwenden. Beispiel:

| X | Rank |
|---|---|
| 70 | 1 |
| 60 | 2.5 |
| 60 | 2.5 |
| 50 | 4 |

# Zusammenfassung

- Eine Möglichkeit der Embeddingsevaluation: Vergleiche Wortähnlichkeiten mit menschlichen Ähnlichkeitsnormen (Wortpaare ohne Kontext).
- Berechne Pearson Korrelations(koeffizient)
- oder besser Spearman Rank Korrelations(koeffizient)

# Literatur

- Gerd Fischer: Lineare Algebra. Eine Einführung für Studienanfänger
- D.G. Rees: Essential Statistics (2001)
- David Moore: Statistics: Concepts and Controversies (2001)
- `http://www.statsoft.com/textbook`, insbesondere `http://www.statsoft.com/Textbook/Basic-Statistics`
- Nette online Korrelationsberechnung mit Scatterplots `https://www.answerminer.com/calculators/correlation-test/`
- Übungsblatt 1, Aufgabe 2