Ludwig-Maximilians-Universitaet Muenchen 30.06.2015
Institute for Informatics
Prof. Dr. Volker Tresp
Gregor Jossé
Johannes Niedermayer

Machine Learning and Data Mining

Summer 2015

Exercise Sheet 11

Presentation of Solutions to the Exercise Sheet on the 08.07.2015

Exercise 11-1 Document Distance

Consider four documents from a document dataset, which has been mapped onto an lexicon of size $M = 100$ w.r.t. word frequency $x_{i,j} \in \{1,2,\dots\}$.

Let A denote the lexicon itself, i.e. $\forall j \in \{1,\dots,M\} : x_{A,j} = 1$. Let B be a document containing only the first word of A ($x_{B,1} = 1 \wedge \forall j \in \{2,\dots,M\} : x_{B,j} = 0$). Let C contain the first 50 words of A, and, finally, let D contain the 11th to 60th word twice.

a) Compute the pairwise distance of vectors A, B, C, D, w.r.t. the following distance measures:

$$
\begin{aligned}
dist_{\text{eucl}}(x,y) &= \Big(\sum_{i=1}^{M}(x_j - Y_j)^2\Big)^{1/2} \\
dist_{\text{simple}}(x,y) &= \frac{1}{M}\sum_{i=1}^{M}(1 - I(x_j = y_j)) \\
dist_{\text{simple00}}(x,y) &= \frac{1}{M-F}\sum_{i=1}^{M}(1 - I(x_j = y_j)) \\
dist_{\cos}(x,y) &= 1 - \frac{x^T y}{\|x\|\|y\|} \\
dist_{\text{pearson}}(x,y) &= 1 - \frac{\tilde{x}^T \tilde{y}}{\|\tilde{x}\|\|\tilde{y}\|}
\end{aligned}
$$

where $I(\text{condition})$ is the indicator function which is 1 iff the condition is fulfilled and 0 otherwise, F is the number of components in which both vectors are 0, and $\tilde{x} := x - mean(x)$.

b) How do the distances change if it is also known that the first fifty words are contained in 750 of the total $N = 1000$ documents in the set, while all other words only appear in 5 documents? *Remark:* You know the "term frequency", which measures the absolute frequency of words in a document. When there is additional information about the global frequency of a term (i.e., is it common or rare among all documents), it should also be taken into account. This is often done using the inverse document frequency:

$$\text{idf}_j = \log\left(\frac{N}{n_j}\right)$$

where N is the number of documents and n_j is the number of documents in which the word j occurs. The measures tf and idf are often combined by multiplication: $\text{tf idf} = \text{tf} \cdot \text{idf}$. What does this measure reflect? Use this measure for this exercise.

Possible Solution:

Es handelt sich hier bei den $x_{i,j}$ um Term Frequencies *tf* ($\text{tf}_{i,j}$: wie oft taucht Wort j in Dokument i auf?)
Die Dokument-Vektoren sehen also also folgendermaßen aus:

$$A = (x_1, \ldots, x_{100}) = (1, \ldots, 1)$$

$$B = (x_1, \ldots, x_{100}) \text{ , mit } x_i = \begin{cases} 1 & i = 1 \\ 0 & \text{sonst} \end{cases}$$

$$C = (x_1, ..., x_{100}) \text{ , mit } x_i = \begin{cases} 1 & i \in [1{,}50] \\ 0 & \text{sonst} \end{cases}$$

$$D = (x_1, ..., x_{100}) \text{ , mit } x_i = \begin{cases} 2 & i \in [11{,}60] \\ 0 & \text{sonst} \end{cases}$$

a) Distanzberechnung

Euklidische Distanz $dist_{eukl}(A, B) = \|A - B\| = \sqrt{\sum (a_i - b_i)^2}$:

$$dist_{eukl}(X, X) = 0$$

$$\begin{aligned} dist_{eukl}(A, B) &= \|A - B\| = \sqrt{\sum (a_i - b_i)^2} \\ &= \sqrt{(1-1)^2 + 99 \cdot (1-0)^2} = \sqrt{99} = 3\sqrt{11} \approx 9{,}95 \\ dist_{eukl}(A, C) &= \sqrt{50 \cdot (1-1)^2 + 50 \cdot (1-0)^2} = \sqrt{50} = 5\sqrt{2} \approx 7{,}71 \\ dist_{eukl}(A, D) &= \sqrt{50 \cdot (1-2)^2 + (10+40) \cdot (1-0)^2} = \sqrt{100} = 10 \\ dist_{eukl}(B, C) &= \sqrt{(1-1)^2 + 49 \cdot (1-0)^2 + 50 \cdot (0-0)^2} = \sqrt{49} = 7 \\ dist_{eukl}(B, D) &= \sqrt{(1-0)^2 + 49 \cdot (0-0)^2 + 50 \cdot (2-0)^2} = \sqrt{201} \approx 14{,}177 \\ dist_{eukl}(C, D) &= \sqrt{10 \cdot (1-0)^2 + 40 \cdot (2-1)^2 + 10 \cdot (2-0)^2 + 40 \cdot (0-0)^2} = 3\sqrt{10} = 9{,}49 \end{aligned}$$

Possible Solution:

$$\text{Indikatorfunktion } I(X) = \begin{cases} 1 & X = 1 \quad \text{bzw} \quad X = \text{true} \\ 0 & \text{sonst} \end{cases}$$

Simple Distanz: $dist_{simple}(A,B) = \frac{1}{M} \sum_{j=1}^{M} (1 - I(a_j = b_j))$

$$dist_{simple}(X,X) = \sum (1 - I(x_j = x_j)) = 0$$

$$dist_{simple}(A,B) = \frac{1}{100} \sum_{j=1}^{100} (1 - I(a_j = b_j)) = \frac{1}{100} \left[(1 - I(1=1)) + 99 \cdot (1 - I(1=0))\right] =$$

$$= \frac{1}{100} \left[1 - 1 + 99 \cdot (1-0)\right] = \frac{99}{100} = .99$$

$$dist_{simple}(A,C) = \frac{1}{100} \left[50 \cdot (1 - I(1=1)) + 50 \cdot (1 - I(1=0))\right] = \frac{50}{100} = .5$$

$$dist_{simple}(A,D) = \frac{1}{100} \left[50 \cdot (1 - I(1=0)) + 50 \cdot (1 - I(1=2))\right] = \frac{100}{100} = 1$$

$$dist_{simple}(B,C) = \frac{1}{100} \left[1 \cdot (1 - I(1=1)) + 49 \cdot (1 - I(0=1)) + 50 \cdot (1 - I(0=0))\right] = \frac{49}{100} = .49$$

$$dist_{simple}(B,D) = \frac{1}{100} \left[1 \cdot (1 - I(1=0)) + 49 \cdot (1 - I(0=0)) + 50 \cdot (1 - I(0=2))\right] = \frac{51}{100} = .51$$

$$dist_{simple}(C,D) = \frac{1}{100} \left[10 \cdot (1 - I(1=0)) + 40 \cdot (1 - I(1=2)) + 10 \cdot (1 - I(0=2) + \right.$$

$$\left. + 40 \cdot (1 - I(0=0))\right] = \frac{60}{100} = .6$$

Simple 00 Distanz: $dist_{simple00}(A,B) = \frac{1}{M-F} \sum_{j=1}^{M} (1 - I(a_j = b_j))$ mit $F :=$ Anzahl der Variablen in denen beide Vektoren übereinstimmend gleich 0 sind.

$$dist_{simple00}(X,X) = \sum (1 - I(x_j = x_j)) = 0$$

$$dist_{simple00}(A,B) = \frac{1}{100 - 0} \left[(1 - I(1=1)) + 99 \cdot (1 - I(1=0))\right] = dist_{simple}(A,B) = .99$$

gleiches gilt für die anderen $dist_{simple00}(A,X)$

$$dist_{simple00}(B,C) = \frac{1}{100 - 50} \left[1 \cdot (1 - I(1=1)) + 49 \cdot (1 - I(1=0)) + 50 \cdot (1 - I(0=0))\right] =$$

$$= \frac{49}{50} = .98$$

$$dist_{simple00}(B,D) = \frac{1}{100 - 49} \left[(1 - I(1=0)) + 49 \cdot (1 - I(0=0)) + 50 \cdot (1 - I(0=2))\right] =$$

$$= \frac{51}{51} = 1$$

$$dist_{simple00}(C,D) = \frac{1}{100 - 40} \left[10 \cdot (1 - I(1=0)) + 40 \cdot (1 - I(1=2)) + 10 \cdot (1 - I(0=2)) + \right.$$

$$\left. + 40 \cdot (1 - I(0=0))\right] = \frac{60}{60} = 1$$

Possible Solution:

Cosinusdistanz: $dist_{cos}(A,B) = 1 - \cos(A,B) = 1 - \frac{A^T B}{\|A\| \cdot \|B\|}$.

$$dist_{cos}(X,X) = 1 - \frac{X^T X}{\sqrt{X^T X}^2} = 0$$
$$dist_{cos}(A,B) = 1 - \frac{1}{\sqrt{100 \cdot 1^2} \cdot \sqrt{1^2}} = 1 - 1/10 = .9$$
$$dist_{cos}(A,C) = 1 - \frac{50}{\sqrt{100} \cdot \sqrt{50 \cdot 1^2}} = 1 - \frac{5}{5\sqrt{2}} = 1 - \frac{1}{1\sqrt{2}} = .29$$
$$dist_{cos}(A,D) = 1 - \frac{50 \cdot 2}{10 \cdot \sqrt{50 \cdot 2^2}} = 1 - \frac{100}{10 \cdot 5 \cdot \sqrt{2} \cdot 2} = 1 - \frac{1}{1\sqrt{2}} = .29$$

und zwar weil: $dist_{cos}(aX, bY) = dist_{cos}(X,Y)$ für positive Skalare a,b

$$dist_{cos}(B,C) = 1 - \frac{1}{\sqrt{1^2} \cdot 5\sqrt{2}} = 1 - \frac{100}{10 \cdot 5 \cdot \sqrt{2} \cdot 2} = 1 - \frac{1}{1\sqrt{2}} \approx .86$$
$$dist_{cos}(B,D) = 1 - \frac{0}{const} = 1$$
$$dist_{cos}(C,D) = 1 - \frac{40 \cdot 2}{5\sqrt{2} \cdot 10 \cdot \sqrt{2}} = 1 - \frac{80}{100} = .2$$

Pearsondistanz: $dist_{pearson}(A,B) = dist_{cos}(\tilde{A},\tilde{B}) = dist_{cos}(A - \text{mean}(A), B - \text{mean}(B)) = 1 - \frac{(A - \text{mean}(A))^T (B - \text{mean}(B))}{\|(A^T - \text{mean}(A))\| \cdot \|(B - \text{mean}(B))\|}$.

$$dist_{pearson}(X,X) = dist_{cos}(\tilde{X},\tilde{X}) = 0$$
$$dist_{pearson}(A,X) = \text{NA, weil } \left\|\tilde{A}\right\| = 0$$
$$dist_{pearson}(B,C) = 1 - \frac{1 \cdot .495 - 49 \cdot .005 + 50 \cdot .005}{\sqrt{.99^2 + 99 \cdot .01^2}\sqrt{50 \cdot .5^2 + 50 \cdot .5^2}} = 1 - \frac{.5}{.99995 \cdot 5} \approx .899$$
$$dist_{pearson}(B,D) = 1 - \cdots \approx 1.101$$
$$dist_{pearson}(C,D) = 1 - \cdots \approx .4$$

Noch zur Pearson-Distanz: Sind alle Elemente eines Vektors gleich, werden die Elemente des Vektors durch die Mittelwertbereinigung auf 0 gesetzt. Die Li¿½nge des Vektors ($\|x\|$) ist damit Null . $\Rightarrow$ Der Nenner der Kosinusdistanz wird dann ebenfalls 0 (Division by Zero). Semantisch wi¿½re das der Winkel eines Vektors zu einem Punkt (Ursprung).

Beispiel: $A = (2,2), \Rightarrow \bar{x} = 2$
$\Rightarrow \tilde{A} = (2 - \bar{x}, ...) = (0,0) \Rightarrow \|A\| = 0$

Sinn der i¿½bung: unterschiedliche Distanzmai¿½e ergeben unterschiedliche Distanzen, welche Distanz die „richtige“ ist, muss i¿½ber Experimente und/oder Spezialwissen entschieden werden.

Possible Solution:

Zusammenfassung:

Euklidische Distanz $\in [0,\infty]$

	A	B	C	D
A	0.000	9.950	7.071	10.000
B	9.950	0.000	7.000	14.177
C	7.071	7.000	0.000	9.487
D	10.000	14.177	9.487	0.000

Simple Distanz $\in [0,1]$

	A	B	C	D
A	0.000	0.990	0.500	1.000
B	0.990	0.000	0.490	0.510
C	0.500	0.490	0.000	0.600
D	1.000	0.510	0.600	0.000

Simple 00 Distanz $\in [0,1]$

	A	B	C	D
A	0.000	0.990	0.500	1.000
B	0.990	0.000	0.980	1.000
C	0.500	0.980	0.000	1.000
D	1.000	1.000	1.000	0.000

Kosinusdistanz $\in [0,2]$

	A	B	C	D
A	0.000	0.900	0.293	0.293
B	0.900	0.000	0.859	1.000
C	0.293	0.859	0.000	0.200
D	0.293	1.000	0.200	0.000

Pearson Distanz $\in [0, 2]$

	A	B	C	D
A	NaN	NaN	NaN	NaN
B	NaN	0.000	0.899	1.101
C	NaN	0.899	0.000	0.400
D	NaN	1.101	0.400	0.000

b) Wenn Informationen über die globalen Häufigkeiten der einzelnen Wörter bekannt sind, sollten diese mit den Dokumentenvektoren verrechnet werden: üblich ist die *inverse document frequency*:

$\text{idf}_j = \log\left(\frac{N}{n_j}\right)$.

N ist die Anzahl aller Dokumente, n_j die Anzahl aller Dokumente in denen Wort j vorkommt.

Hieraus wird dann die tfidf gebildet:

$\text{tfidf}_{i,j} = \text{tf}_{i,j} \cdot \text{idf}_j$.

In unserem Beispiel haben die ersten 50 Wörter eine idf_j von $\log\left(\frac{1000}{750}\right) = \log\left(\frac{4}{3}\right) \approx 0{,}288$, die anderen Wörter ($j \in \{51, \ldots, 100\}$) haben eine idf_j von $\log\left(\frac{1000}{5}\right) = \log(200) \approx 5{,}298$. Wir erhalten also neue Merkmalsvektoren:

$$A = (x_1, \ldots, x_{100}) = \text{ , mit } x_i = \begin{cases} \log\left(\frac{4}{3}\right) & i = \{1, \ldots, 50\} \\ \log(200) & \text{sonst} \end{cases}$$

$$B = (x_1, \ldots, x_{100}) \text{ , mit } x_i = \begin{cases} \log\left(\frac{4}{3}\right) & i = 1 \\ 0 & \text{sonst} \end{cases}$$

$$C = (x_1, ..., x_{100}) \text{ , mit } x_i = \begin{cases} \log\left(\frac{4}{3}\right) & i \in [1{,}50] \\ 0 & \text{sonst} \end{cases}$$

$$D = (x_1, ..., x_{100}) \text{ , mit } x_i = \begin{cases} 2 * \log\left(\frac{4}{3}\right) & i \in [11{,}50] \\ 2 * \log(200) & i \in [51{,}60] \\ 0 & \text{sonst} \end{cases}$$

Da also alle Merkmalsvektoren insgesamt heruntergewichtet wurden, ist zu erwarten, dass die Distanzen insgesamt abnehmen werden:

Possible Solution:

Distanzen:

Euklidische Distanz

	A	B	C	D
A	0.000	37.519	37.465	37.520
B	37.519	0.000	2.014	33.708
C	37.465	2.014	0.000	33.571
D	37.520	33.708	33.571	0.000

Simple Distanz

	A	B	C	D
A	0.000	0.990	0.500	1.000
B	0.990	0.000	0.490	0.510
C	0.500	0.490	0.000	0.600
D	1.000	0.510	0.600	0.000

Simple 00 Distanz

	A	B	C	D
A	0.000	0.990	0.500	1.000
B	0.990	0.000	0.980	1.000
C	0.500	0.980	0.000	1.000
D	1.000	1.000	1.000	0.000

Kosinusdistanz

	A	B	C	D
A	0.000	0.992	0.946	0.551
B	0.992	0.000	0.859	1.000
C	0.946	0.859	-0.000	0.903
D	0.551	1.000	0.903	-0.000

Pearson Distanz

	A	B	C	D
A	-0.000	1.101	2.000	0.734
B	1.101	0.000	0.899	1.042
C	2.000	0.899	0.000	1.266
D	0.734	1.042	1.266	0.000

Unterschiede:

Euklidische Distanz: durch die Gewichtung deutliche Unterschiede in der Höhe der Distanzen – auch die paarweisen Distanzen haben sich relativ zueinander verschoben. Beispiel AC, CD: ursprünglich ist AC $<$ CD; jetzt umgekehrt.

Simple Distanz / -00 Distanz: jeweils kein Einfluss – klar, es werden lediglich Indikatorvektoren vergleichen, und da die Gewichtung hier keinen Einfluss auf die Wertidentitäten hatte (muss nicht immer der Fall sein), gibt es keine Änderungen.

Kosinusdistanz: Starke Änderungen: da die Gewichtungen sich meist auf den Winkel der Vektoren auswirken – nur BC und BD sind nicht betroffen.

Pearsondistanz: Keine NaNs mehr, da A jetzt nicht Varianz 0 hat. Durch die Normalisierung hier auch ein Unterschied BD im Vergleich mit der Kosinusdistanz. (BC wird nicht beeinflusst, da die Gewichtungen sich nur auf Terme $\neq 0$ auswirken.)