Machine Learning
Exercises: vector semantics

Laura Kallmeyer

Summer 2016, Heinrich-Heine-Universität Düsseldorf

Exercise 1 *Consider the following word-context matrix with the three words "orange", "banana" and "car" and the three context words "juice", "the" and "drive".*

	juice	*the*	*drive*
orange	*10*	*20*	*0*
banana	*8*	*20*	*0*
car	*1*	*20*	*10*

1. *Compute the MLEs using frequencies for the probabilities $P(w)$, $P(c)$ and $P(w,c)$ for each word w and each context word c.*
2. *Based on these, compute the PPMI values for the cells in the matrix.*
3. *Now compute the cosine similarity values of the PPMI vectors for "orange" and "banana" and for "orange" and "car".*

Solution:

1.

	juice	the	drive	$P(w)$
orange	$\frac{10}{89}$	$\frac{20}{89}$	0	$\frac{30}{89}$
banana	$\frac{8}{89}$	$\frac{20}{89}$	0	$\frac{28}{89}$
car	$\frac{1}{89}$	$\frac{20}{89}$	$\frac{10}{89}$	$\frac{31}{89}$
$P(c)$	$\frac{19}{89}$	$\frac{60}{89}$	$\frac{10}{89}$	

2. $PPMI(w,c) = \max(\log_2 \frac{P(w,c)}{P(w)P(c)}, 0)$

	juice	the	drive
orange	0.64	0	0
banana	0.42	$8.33 \cdot 10^{-2}$	0
car	0	0	1.52

orange, juice: $\log_2 \frac{\frac{10}{89}}{\frac{30}{89}\frac{19}{89}} = \log_2 \frac{89}{57}$

banana, juice: $\log_2 \frac{\frac{8}{89}}{\frac{28}{89}\frac{19}{89}} = \log_2 \frac{8 \cdot 89}{28 \cdot 19}$

car, juice: $\log_2 \frac{\frac{1}{89}}{\frac{31}{89}\frac{19}{89}} = \log_2 \frac{89}{31 \cdot 19} < 0$

...

orange, the: $-1.63 \cdot 10^{-2}$ banana, the: $8.33 \cdot 10^{-2}$ car, the: $-6.36 \cdot 10^{-2}$

car, drive: 1.52

3. $CosSim(\vec{v}, \vec{w}) = \frac{\vec{v} \cdot \vec{w}}{|\vec{v}| \ |\vec{w}|} = \frac{\sum_{i=1}^{n} v_i w_i}{\sqrt{\sum_{i=1}^{n} v_i^2} \sqrt{\sum_{i=1}^{n} w_i^2}}$

orange, banana: $\frac{0.64 \cdot 0.42}{0.64 \cdot \sqrt{0.42^2 + 0.0833^2}} = 0.98$

orange, car: 0