Machine Learning
Exercises: kNN

Laura Kallmeyer

Summer 2016, Heinrich-Heine-Universität Düsseldorf

Exercise 1 *Consider the k nearest neighbor example from slide 20, with the following term frequency counts:*

Training:	*Class l*			*Class c*		*new docs:*	
terms	d_1	d_2	d_3	d_4	d_5	d_6	d_7
love	*10*	*8*	*7*	*0*	*1*	*5*	*1*
kiss	*5*	*6*	*4*	*1*	*0*	*6*	*0*
inspector	*2*	*0*	*0*	*12*	*8*	*2*	*12*
murderer	*0*	*1*	*0*	*20*	*56*	*0*	*4*

1. *Replace these counts with the corresponding $tf_{td}idf_t$ weights.*
2. *Then normalize the vectors of the $tf_{td}idf_t$ weights of d_1, d_4, d_6 and d_7 and calculate the Euclidian distances between each of the test documents d_6, d_7 and each of these training documents.*

Solution:

1. "love" and "kiss" both appear in 4 out ot 5 documents, "inspector" and "murderer" in 3 out ot 5. Consequently, for the first two, we multiply the count with $\log \frac{5}{4} = 0.1$ and for the latter two, we multiply with $\log \frac{5}{3} = 0.22$.

Training:	Class l			Class c		new docs:	
terms	d_1	d_2	d_3	d_4	d_5	d_6	d_7
love	1	0.8	0.7	0	0.1	0.5	0.1
kiss	0.5	0.6	0.4	0.1	0	0.6	0
inspector	0.44	0	0	2.64	1.76	0.22	2.64
murderer	0	0.22	0	4.4	12.32	0	0.88

2. normalized vectors for d_1 (division by 1.2), d_4 (division by 5.13), d_6 (division by 0.81) and d_7 (division by 2.78):

	d_1	d_4	d_6	d_7
love	0.83	0	0.62	0.04
kiss	0.42	0.02	0.74	0
inspector	0.37	0.51	0.27	0.95
murderer	0	0.86	0	0.32

Euclidian distances:

d_1 and d_6: $\sqrt{0.1638} = 0.4$
d_4 and d_6: $\sqrt{1.4101} = 1.19$

d_1 and d_7: $\sqrt{1.2393} = 1.11$
d_4 and d_7: $\sqrt{0.4872} = 0.7$

Exercise 2 *Now consider the weighted score on slide 27:*

$$score(c, d) = \sum_{d_t \in S_k(d)} I_c(d_t) \cos(\vec{v}(d_t), \vec{v}(d))$$

where $\vec{v}(d)$ is the vector of some document d.

Normalize this score so that we obtain a probability $P(c|d)$.

Solution:

$$P(c|d) = \frac{\sum_{d_t \in S_k(d)} I_c(d_t) \cos(\vec{v}(d_t), \vec{v}(d))}{\sum_{d_t \in S_k(d)} \cos(\vec{v}(d_t), \vec{v}(d))}$$

Exercise 3 *Assume that we have two classes, A and B and a new document d to be classified.*

The following training data is available:

d_i	*class*	$\cos(\vec{v}(d_i), \vec{v}(d))$
d_1	*A*	*1*
d_2	*B*	*0.95*
d_3	*B*	*0.94*
d_4	*A*	*0.45*
d_5	*A*	*0.4*
d_6	*B*	*0.39*

Let us assume that we use the cosine as a distance measure, i.e., the higher the cosine, the closer are two vectors.

Which class would be assigned to d with a k-nearest neighbor classifier using cosine if

1. $k = 3$ *and simple majority vote (score as in slide 23);*
2. $k = 5$ *and simple majority vote;*
3. $k = 3$ *and a weighted score as in slide 27;*
4. $k = 5$ *and a weighted score as in slide 27.*

Solution:

1. $k = 3$ and simple majority vote: $score(A, d) = 1$, $score(B, d) = 2$, therefore class B
2. $k = 5$ and simple majority vote: $score(A, d) = 3$, $score(B, d) = 2$, therefore class A
3. $k = 3$ and a weighted score as in slide 27: $score(A, d) = 1$, $score(B, d) = 0.95 + 0.94$, therefore class B
4. $k = 5$ and a weighted score as in slide 27: $score(A, d) = 1 + 0.45 + 0.4$, $score(B, d) = 0.95 + 0.94$, therefore class B