

# Introduction to Machine Learning, Fall 2014 - Exercise session II

Rodion “rodde” Efremov

November 6, 2015

## Problem 1 (3 points)

Consider a document-term matrix, where  $tf_{ij}$  is the number of times that the  $i^{th}$  word (term) appears in the  $j^{th}$  document, and let  $m$  be the total number of documents in the collection. Consider the variable transformation that is defined by

$$tf'_{ij} = tf_{ij} \log \frac{m}{df_i}, \quad (1)$$

where  $df_i$  is the number of documents in which the  $i^{th}$  term appears, which is known as the *document frequency* of the term. This transformation is known as the *inverse document frequency* transformation.

- (a) What is the effect of this transformation if a term occurs in only one document? In every document?
- (b) What is the overall effect and what might be the purpose of this transformation?
- (c) Can you think of other (non-document) data in which this transformation might be useful?

**(a)**

If the term occurs in only one document, we have  $df_i = 1$ , and, thus,  $tf'_{ij} = tf_{ij} \log m$ . If  $df_i = m$ , we have that  $tf'_{ij} = tf_{ij} \log \frac{m}{m} = 0$ .

**(b)**

The inverse document frequency aims to minimize the effect of over-emphasized terms such as “the”, “in”, and so on, and to emphasize the words that are not used very often, and, thus, allow better chance of differentiating between documents.

(c)

Suppose that the rows of the matrix represent particular market baskets of individual shoppers, and the columns represent particular items. Now, the "*inverse item frequency*" lessens the effect of the items that tend to be "popular", such as bread or milk, for instance.

## Problem 2 (3 points)

In this exercise we explore the relationships between the cosine and correlation similarity measures and Euclidean distance for data vectors in  $\mathbb{R}^n$ .

- (a) What is the range of values that are possible for the cosine measure?
- (b) If two objects have a cosine measure of 1, are they necessarily identical? Explain.
- (c) What is the relationship of the cosine measure to correlation, if any? (Hint: Look at statistical measures such as mean and standard deviation in cases where cosine and correlation are the same and different.)
- (d) Derive the mathematical relationship between cosine similarity and Euclidean distance when each data object has an  $L_2$  length (norm) of 1.
- (e) Derive the mathematical relationship between correlation and Euclidean distance when each data point has been standardized by subtracting its mean and dividing by its standard deviation.

(a)

The range of values is the range of values of cosine function, i.e.,  $[-1, 1]$ .

(b)

If the cosine measure of two objects  $\mathbf{x}$  and  $\mathbf{y}$  is 1, it implies that the two vectors have zero degree angle, i.e. they have the same orientation. Because the cosine measure is agnostic to the lengths of the vectors however, it implies that  $\mathbf{x} = a\mathbf{y}$  for some real  $a > 0$ .

(c)

Suppose we are given two data vectors  $X, Y$ , and their respective means are  $\bar{X}, \bar{Y}$ . Now we have that the correlation coefficient of the two is

$$r_{X,Y} = \frac{\sum_{k=1}^n (X_k - \bar{X})(Y_k - \bar{Y})}{\sqrt{\sum_{k=1}^n (X_k - \bar{X})^2} \sqrt{\sum_{k=1}^n (Y_k - \bar{Y})^2}}.$$

If  $A = (A_1, \dots, A_n) \in \mathbb{R}^n$  and  $\alpha \in \mathbb{R}$ , by  $A + \alpha$  we denote the vector  $(A_1 + \alpha, \dots, A_n + \alpha)$ . Now, let us substitute  $X' = X - \bar{X}$  and  $Y' = Y - \bar{Y}$ . We obtain

$$\begin{aligned}
r_{X,Y} &= \frac{\sum_{k=1}^n (X_k - \bar{X})(Y_k - \bar{Y})}{\sqrt{\sum_{k=1}^n (X_k - \bar{X})^2} \sqrt{\sum_{k=1}^n (Y_k - \bar{Y})^2}} \\
&= \frac{\sum_{k=1}^n X'_k Y'_k}{\sqrt{\sum_{k=1}^n X'^2_k} \sqrt{\sum_{k=1}^n Y'^2_k}} \\
&= \frac{X' \cdot Y'}{|X'| |Y'|} \\
&= \cos(X', Y') \\
&= \cos(X - \bar{X}, Y - \bar{Y}).
\end{aligned}$$

(d)

Suppose we are given two data objects  $X$  and  $Y$ , both having an  $L_2$  length of 1. Now we have that

$$\begin{aligned}
d(X, Y) &= \sqrt{\sum_{k=1}^n (X_k - Y_k)^2} \\
&= \sqrt{\sum_{k=1}^n (X_k^2 + Y_k^2 - 2X_k Y_k)} \\
&= \sqrt{|X|^2 + |Y|^2 - 2 \sum_{k=1}^n X_k Y_k} \\
&= \sqrt{|X|^2 + |Y|^2 - 2 \frac{X \cdot Y}{|X| |Y|}} \\
&= \sqrt{|X|^2 + |Y|^2 - 2 \cos \theta_{X,Y}} \\
&= \sqrt{2 - 2 \cos \theta_{X,Y}} \\
&= \sqrt{2(1 - \cos \theta_{X,Y})} \in [0, 2].
\end{aligned}$$

where  $\cos \theta_{X,Y}$  is the cosine measure of the vectors  $X$  and  $Y$ .

(e)

Suppose we are given two  $n$ -vectors  $X'$  and  $Y'$ . We compute the means of both of them, and denote the aforementioned as  $\bar{X}'$  and  $\bar{Y}'$ . Next, we compute standard deviations of both,  $\sigma_{X'}$  and  $\sigma_{Y'}$ , after which we have all the data needed to standardize  $X'$  and  $Y'$ : for each component, we subtract the respective mean and divide by the respective standard deviation, and thus, we obtain two new

vectors  $X$  and  $Y$  with both having means  $\bar{X} = \bar{Y} = 0$  and standard deviation  $\sigma_X = \sigma_Y = 1$ . Now we have that

$$\begin{aligned}
 d^2(X, Y) &= \sum_{k=1}^n (X_k - Y_k)^2 \\
 &= \sum_{k=1}^n (X_k^2 + Y_k^2 - 2X_k Y_k) \\
 &= |X|^2 + |Y|^2 - 2(n-1) \left( \frac{1}{n-1} \sum_{k=1}^n X_k Y_k \right) \\
 &= |X|^2 + |Y|^2 - 2(n-1) \left( \frac{\frac{1}{n-1} \sum_{k=1}^n (X_k - \bar{X})(Y_k - \bar{Y})}{\sigma_X \sigma_Y} \right) \\
 &= |X|^2 + |Y|^2 - 2(n-1)r_{X,Y},
 \end{aligned}$$

where  $r_{X,Y}$  denotes the correlation factor of the vectors  $X, Y$ .

### Problem 3 (3 points)

Proximity is typically defined between a pair of objects.

- (a) Give two ways in which you might define the 'proximity' among a set of (more than two) objects (i.e. a single measure of how similar an arbitrary number of items are all to one another)
- (b) How might you define the distance between two sets of points in Euclidian space?
- (c) How might you define the proximity between two sets of data objects? (Make no assumptions about the data objects, except that a proximity measure is defined between any pair of objects.)

(a)

If  $P(\mathbf{x}, \mathbf{y})$  is a "normal" function giving the proximity of data objects  $\mathbf{x}$  and  $\mathbf{y}$ , and  $A$  is an arbitrary set of data objects (which can be indexed like  $A_1, A_2, \dots, A_n$ ), I would try

$$P(A) = \left( \sum_{1 \leq i < j \leq n} P(A_i, A_j) \right) \binom{n}{2}^{-1},$$

or namely the average proximity over all pairs of data objects from  $A$ . For another one, I would choose a small  $\varepsilon > 0$ , and define  $P(A)$  as

$$\prod_{1 \leq i < j \leq n} \left( \max(P(A_i, A_j), \varepsilon) \right)^{\frac{1}{\binom{n}{2}}} \geq \varepsilon,$$

or namely the geometric mean over all pairs of data objects from  $A$ . The  $\varepsilon$  is chosen to be positive so that a single proximity value of 0 does not dominate the entire proximity of a set and bring it down to 0. (Above, it is assumed that  $P(\mathbf{x}, \mathbf{y}) \geq 0$  for all data objects  $\mathbf{x}, \mathbf{y}$ .)

(b)

I would do the way topologists do: if  $A$  and  $B$  are two (finite) sets of points in Euclidian space, then

$$P(A, B) = \min_{(\mathbf{x}, \mathbf{y}) \in A \times B} d(\mathbf{x}, \mathbf{y}).$$

(c)

Suppose we are given two sets of data objects,  $A$  and  $B$ . Now, the easiest way to define proximity between them is

$$P(A, B) = \sum_{\mathbf{x} \in A} \sum_{\mathbf{y} \in B} \frac{P(\mathbf{x}, \mathbf{y})}{|A||B|}.$$

## Problem 4 (15 points)

In this problem we will consider similarity measures for movies on the Movielens dataset.

(b)

The Jaccard coefficient for 'Three Colors: Red' and 'Three Colors: Blue' seems to be 0.598. The 5 closest movies to "Taxi Driver" seem to be

Movie	Jaccard coefficient
Alien (1979)	0.416666666667
Full Metal Jacket (1987)	0.416666666667
Apocalypse Now (1979)	0.404181184669
2001: A Space Odyssey (1968)	0.397058823529
Stand by Me (1986)	0.39406779661

I chose "Star Wars (1977)"; the closest five movies seem to be

Movie	Jaccard coefficient
GoodFellas (1990)	0.786885245902
Brazil (1985)	0.609951845907
GoldenEye (1995)	0.582568807339
Princess Bride, The (1987)	0.570247933884
Heavy Metal (1981)	0.565279770445

I do not know any of the above 5 movies. However they seem to be related if you consider their release years.

(c)

The correlation coefficient between ‘Toy Story’ and ‘GoldenEye’ seems to be 0.222. The correlation coefficient between ‘Three Colors: Red’ and ‘Three Colors: Blue’ seems to be 0.760.

What comes to the closest five movies to ‘Taxi Driver’, my implementation returned the following result:

Movie	Correlation coefficient
Mad Love (1995)	1.0
Diabolique (1996)	0.627455805138
Brother Minister: The Assassination of Malcolm X (1994)	0.612146092152
Muppet Treasure Island (1996)	0.450357203239
Stargate (1994)	0.392980655241

What comes to ‘Star Wars (1977)’, the five closest movies are

Movie	Correlation coefficient
Striptease (1996)	1.0
Long Kiss Goodnight, The (1996)	0.771743633141
Princess Bride, The (1987)	0.747981422379
GoodFellas (1990)	0.672555855888
Cold Comfort Farm (1995)	0.583860806367

(d)

Once again, I have not seen most of the movies involved in my answer, yet by judging from the names, it would seem that Jaccard coefficient is “better” than the correlation coefficient.