University of Helsinki, Department of Computer Science

Assignment 2 of Introduction to Machine Learning

Gao Yuan

November 6, 2013

1 Problem 1

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

$$tf'_{ij} = tf_{ij} log \frac{m}{df_i}$$

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.

Answer:

(a) What is the effect of this transformation if a term occurs in only one document? In every document?

If a term only occurs in one documentation, the df_i would be 1 then $log \frac{m}{df_i}$ would be comparatively large. If it appears in every documentation, then $log \frac{m}{df_i}$ would be 0, which results tf'_{ij} to be zero. To sum up the tf'_{ij} of a rare term is high, whereas the tf'_{ij} of a frequent term is likely to be low.

(b) What is the overall effect and what might be the purpose of this transformation?

Considering the following scenario, when searching for some document, you provide main idea constructed with a sentence.Then this transformation will give relative significance of each word contained in the sentence. This will help the algorithm to find more accurate result.

(c) Can you think of other (non-document) data in which this transformation might be useful?

It might be applied in simultaneous localization and mapping(SLAM) in robotics, especially in visual SLAM. If we consider each component in the environment as one term and the environments as documents, using this technique, we can filter the environments stored in the large database and get a reasonable probability of the where we are. For example, the common objects like chairs and desks will be filtered and the rare component will be a key that identifies the right environment.

2 Problem 2

In this exercise we explore the relationships between the cosine and correlation similarity measures and Euclidean distance for data vectors in R^n.
Answer:

(a) What is the range of values that are possible for the cosine measure?

A cosine measure is defined as

$$CosineSimilarity(A, B) = cos(\theta) \tag{2.1}$$

$$= \frac{A \cdot B}{\|A\| \cdot \|B\|} \tag{2.2}$$

$$= \frac{\sum_{i=1}^{n} a_i \times b_i}{\sqrt{\sum_{i=1}^{n} (a_i)^2} \times \sqrt{\sum_{i=1}^{n} (b_i)^2}} \tag{2.3}$$

As a consequence, theoretically its range is from -1 to 1.

(b) If two objects have a cosine measure of 1, are they necessarily identical? Explain.

Not necessarily, we could say that it is identical cross measured features. But we could not say two objects are identical in any cases.

(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.)

The Pearson correlation coefficient has been defined as follows:

$$CorrelationCoefficient(X,Y) = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2}\sqrt{\sum (y_i - \bar{y})^2}} \tag{2.4}$$

$$= \frac{\langle X - \bar{x},\ X - \bar{y} \rangle}{||X - \bar{x}||\ ||X - \bar{y}||} \tag{2.5}$$

We can observe that

$$CorrelationCoefficient(X,Y) = CosineSimilarity(X - \bar{x}, X - \bar{x}) \tag{2.6}$$

where $\bar{x}$ and $\bar{y}$ are the means of vectors X and Y.

(d) Derive the mathematical relationship between cosine similarity and Euclidean distance when each data object has an L_2 length (norm) of 1.

By applying the conditions that $\|x\| = 1$ and $\|y\| = 1$, equation (2,2) can imply that

$$CosineSimilarity(X,Y) = X \cdot Y \tag{2.7}$$

$$= \sum_{i=1}^{n} x_i y_i \tag{2.8}$$

On the other hand under same conditions, the euclidean distanceformula would be

$$D(X,Y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{2.9}$$

$$= \sqrt{\sum_{i=1}^{n} (x_i^2 - 2x_i y_i + y_i^2)} \tag{2.10}$$

$$= \sqrt{\sum_{i=1}^{n} (x_i^2 - 2x_i y_i + y_i^2)} \tag{2.11}$$

$$= \sqrt{\sum_{i=1}^{n} x_i^2 - 2\sum_{i=1}^{n} x_i y_i + \sum_{i=1}^{n} y_i^2} \tag{2.12}$$

$$= \sqrt{\|X\| - 2\sum_{i=1}^{n} x_i y_i + \|Y\|} \tag{2.13}$$

$$= \sqrt{2 - 2\sum_{i=1}^{n} x_i y_i} \tag{2.14}$$

$$= \sqrt{2 - 2 \times CosineSimilarity(X,Y)} \tag{2.15}$$

(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.

Under the condition of being standardized, the equation (2.5) would imply

$$D(X,Y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{2.16}$$

$$= \sqrt{\sum_{i=1}^{n}(x_i^2 - 2x_iy_i + y_i^2)} \tag{2.17}$$

$$= \sqrt{\sum_{i=1}^{n}(x_i^2 - 2x_iy_i + y_i^2)} \tag{2.18}$$

$$= \sqrt{\sum_{i=1}^{n} x_i^2 - 2\sum_{i=1}^{n} x_iy_i + \sum_{i=1}^{n} y_i^2} \tag{2.19}$$

$$= \sqrt{2n - 2nCorrelationCoefficient(X,Y)} \tag{2.20}$$

3 Problem 3

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)

1. The first way would be calculating the statistical properties of each set of objects e.i mean and variance, and use these properties for measuring if the data is in $R^{n\times n}$ space.
2. The second way would be defining proximity of a set of objects to be the maximum similarity between any data of to-be-compared two objects.

(b) How might you define the distance between two sets of points in Euclidean space?

We could compute the distance between centres of two clusters of sets. The centre is a point that has minimum sum of distances to all points in this set.

(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.)

We just calculate the average proximity between any pair of data in one set and in another set.

4 Problem 4

(a) Download the Movielens data from the course web page. In addition to the data, the le also containssome functions for easily loading the data into Matlab/Octave/R and some example code that you canuse if you wish. See the README les for details.

Answer:
done !

(b)What is the Jaccard coefficient between Three Colors: Red' and 'Three Colors: Blue'? What are the 5 movies with highest Jaccard coefficient to Taxi Driver'? Select a movie of your own choosing (which you are familiar with), what are the 5 movieswith highest Jaccard coefficient to that movie? Do they make sense?

Answer:
1) The jaccard coeficient of Three Colors: Red' and 'Three Colors: Blue' is: 0.59783.
2) the 5 movies with highest Jaccard coefficient to Taxi Driver' are Chinatown (1974), Citizen Kane (1941), Clockwork Orange, A (1971), Godfather: Part II, The (1974), GoodFellas (1990) with coefficients 0.3941, 0.3971, 0.4042, 0.4167 and 0.4167 respectively.
3) the 5 movieswith highest Jaccard coe?cient to Star Wars (1977) are Fargo (1996), Empire Strikes Back, The (1980),Toy Story (1995), Raiders of the Lost Ark (1981) and Return of the Jedi (1983) with coefficient 0.5653, 0.5702, 0.5826, 0.6100 and 0.7869.
4) They do make sense, the geeks like geeky movies.

(c) What is now the similarity between 'Toy Story' and 'GoldenEye'? How about 'Three Colors: Red' and 'Three Colors: Blue'? What are the 5 movies with highest similarity to 'Taxi Driver'? Again, select a movie of your own choosing and list the 5 movies with highest similarity.

Answer:
1)The correlation coeficient of 'Toy Story' and 'GoldenEye' is: 0.2218
2)The correlation coeficient of Three Colors: Red' and 'Three Colors: Blue' is: 0.7597.
3)The 5 movies with highest similarity to 'Taxi Driver' Nixon (1995) are Shadowlands (1993), Get on the Bus (1996), Cable Guy, The (1996) and Othello (1995) with coefficient 0.4894, 0.4895, 0.5198, 0.5287 and 0.5350 respectively.
4)The 5 movies with highest similarity to Star Wars (1977) are Raiders of the Lost Ark (1981), Ghost in the Shell (Kokaku kidotai) (1995), Meet John Doe (1941), Return of

the Jedi (1983) and Empire Strikes Back (1980) with coefficient 0.5361, 0.5996, 0.6333, 0.6726 and 0.7480.

(d)Why do you think this is? Explain.

Answer:
My intuition is the correlation coefficient performs better. My first thought is that the correlation coefficient provides a reasonable interval(around 0) for the situation of randomness. I think the jaccard coefficient tells us whether some movies are popular or belong to same category. However the correlation coefficient also provides information about whether some movies are worth watching or not.