582631 Introduction to Machine Learning, Autumn 2015
Sample solutions for exercise set 2 (based on first week of lectures)

Prepared by Johannes Verwijnen and Amin Sorkhei. Please contact the authors with any fixes or suggestions.

Pen-and-paper problems

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

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

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?

If a term appears in only one document, $df_i = 1$, and thus $\log \frac{m}{df_i} = \log m$ and $tf'_{ij} = tf_{ij} \log m$. If a term appears in every document, $df_i = m$, and thuslog $\frac{m}{df_i} = \log 1 = 0$ and $tf'_{ij} = 0$.

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

The inverse document frequency transformation has the effect of returning a relatively high value for rare terms and a low value for frequent terms, reflecting their information value. This can be handy when weighing the importance of search terms found in documents (ie. giving less weight to very common words, such as 'and' or 'the') when ranking search results.

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

It can be used (for example) to filter out items that are very common from recommendation systems or extracting useful features in classifications tasks (image classification as an example).

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 R^n.

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

The range of the cosine measure is identical to the range of the cosine function, $[-1, 1]$. (Lecture slide 61)

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

The cosine measure only takes into account the angle between vectors, not the magnitude. Therefore vectors $\mathbf{x} = \{1, 1\}$ and $\mathbf{y} = \{2, 2\}$ have $\cos(\mathbf{x}, \mathbf{y}) = 1$ although they are not identical.

(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.)

Pearson's correlation coefficient is by definition the cosine of the angle between two vectors for centered data. Thus when the mean of both vectors is zero, they are equivalent.

This can be easily seen by

$$\begin{aligned}
\cos(\mathbf{x},\mathbf{y}) &= \frac{\mathbf{x}\cdot\mathbf{y}}{\|\mathbf{x}\|_2\,\|\mathbf{y}\|_2} \qquad \textit{Lecture slide 61}\\
&= \frac{\sum_{i=1}^{d} x_i y_i}{\sqrt{\sum_{i=1}^{d} x_i^2}\sqrt{\sum_{i=1}^{d} y_i^2}}\\
&= \frac{\sum_{i=1}^{d} (x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{d} (x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{d} (y_i-\bar{y})^2}} \qquad \textit{if } \bar{x}=\bar{y}=0\\
&= r(\mathbf{x},\mathbf{y}) \qquad \textit{Lecture slide 61}
\end{aligned}$$

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

$$\begin{aligned}
d(\mathbf{x},\mathbf{y}) &= \|\mathbf{x}-\mathbf{y}\|_2 \qquad \textit{Lecture slide 55}\\
&= \sqrt{\sum_{i=1}^{d} (x_i-y_i)^2}\\
&= \sqrt{\sum_{i=1}^{d} (x_i^2 - 2x_iy_i + y_i^2)}\\
&= \sqrt{\sum_{i=1}^{d} x_i^2 - 2\sum_{i=1}^{d} x_iy_i + \sum_{i=1}^{d} y_i^2}\\
&= \sqrt{\|\mathbf{x}\|_2^2 - 2\sum_{i=1}^{d} x_iy_i + \|\mathbf{y}\|_2^2} \qquad \text{as } \|\mathbf{z}\|_2 = \sqrt{\sum_{i=1}^{d} z_i^2}\\
&= \sqrt{2 - 2\sum_{i=1}^{d} x_iy_i} \qquad \text{as } \|\mathbf{x}\|_2 = \|\mathbf{y}\|_2 = 1\\
&= \sqrt{2(1-\cos(\mathbf{x},\mathbf{y}))} \qquad \text{as } \|\mathbf{x}\|_2 = \|\mathbf{y}\|_2 = 1 \rightarrow \cos(\mathbf{x},\mathbf{y}) = \mathbf{x}\cdot\mathbf{y} = \sum_{i=1}^{d} x_iy_i
\end{aligned}$$

(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.

There are two definitions of Pearson's correlation coefficient available, the slide set's definition (slide 61)

$$r(\mathbf{x},\mathbf{y}) = \frac{\sum_{i=1}^{d} (x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{d} (x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{d} (y_i-\bar{y})^2}}, \tag{2}$$

for standardized data, $\tilde{\mathbf{z}} = \frac{\mathbf{z}-\bar{z}}{\sigma_z}$, simplifies to

$$r(\tilde{\mathbf{x}},\tilde{\mathbf{y}}) = \frac{\sum_{i=1}^{d} \tilde{x}_i\tilde{y}_i}{\sqrt{\sum_{i=1}^{d} \tilde{x}_i^2}\sqrt{\sum_{i=1}^{d} \tilde{y}_i^2}} = \frac{\sum_{i=1}^{d} \tilde{x}_i\tilde{y}_i}{d} \tag{3}$$

as $\bar{x} = \bar{y} = 0$ and $\sum_{i=1}^{d} z_i^2 = d$ if $\bar{z} = 0, \sigma_z = 1$. Now we can start midway from part d)

$$\begin{aligned} d(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) &= \sqrt{\sum_{i=1}^{d} \tilde{x}_i^2 - 2\sum_{i=1}^{d} \tilde{x}_i \tilde{y}_i + \sum_{i=1}^{d} \tilde{y}_i^2} \\ &= \sqrt{d - 2\sum_{i=1}^{d} \tilde{x}_i \tilde{y}_i + d} \qquad \text{given } \sigma_z = 1, \bar{z} = 0 \rightarrow \sum_{i=1}^{d} z_i^2 = d \\ &= \sqrt{2d\left(1 - r\left(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}\right)\right)} \qquad \text{by (3)} \end{aligned}$$

Using the book's definition (p.45)

$$r(\mathbf{x}, \mathbf{y}) = \frac{\frac{1}{d}\sum_{i=1}^{d} x_i y_i - \bar{x}\bar{y}}{\sqrt{\sigma_x^2 \sigma_y^2}}, \tag{4}$$

we see that with the standardized data we also get

$$r(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) = \frac{\sum_{i=1}^{d} \tilde{x}_i \tilde{y}_i}{d}, \tag{5}$$

as above.

Most problems came from mixing the two definitions, but forgetting that the covariance/standard deviation -definition included division by the dimensionality of the vector, d, whereas the slide set definition does not.

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)

You can use pairwise proximity measures of the set and select either the minimum similarity, or the maximum or mean dissimilarity to create a proximity measure for the whole set.

Another way would be to calculate the centroid (euclidean centrepoint) of the data and use the sum or average of distances to that as the proximity measure.

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

You can use the minimum or maximum distance between members of different sets, or the distance between the centroids of the two sets.

(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.)

Once again you could use the minimum pairwise similarity or maximum pairwise dissimilarity between members of different sets or a (dis)similarity measure between the centroids of the two sets.

Programming problem

Problem 4 (15 points)

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

(b) We will now construct a similarity measure over the movies. For simplicity, let us first consider a simple measure that does not use the explicit (numerical) ratings given by the users, nor the time stamps of the ratings, but only whether or not a given movie was rated by a given user. Create a function that, given two different movie IDs as input, outputs the Jaccard coefficient: the number of users who rated both movies divided by the number of users who rated at least one of the movies. For example, for the movies

'Toy Story' and 'GoldenEye' the coefficient should be 0.217. 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 movies with highest Jaccard coefficient to that movie? Do they make sense?

The Jaccard coefficient between 'Three Colors: Red' and 'Three Colors: Blue' is 0.598 and the top 5 most related movies to 'Taxi Driver' are:

1. GoodFellas
2. The Godfather: Part II
3. A Clockwork Orange
4. Citizen Kane
5. Chinatown

Let's find the top 5 most related movies regarding 'The Shining' using Jaccard Coefficient:

1. Jaws
2. Psycho
3. The Silence of the Lambs
4. Cape Fear
5. Aliens

(c) Now let's try a similarity measure that uses the explicit ratings. Create a second function that, given two different movie IDs as input, outputs the correlation coefficient of the ratings given to those two movies by all users which have rated both movies. (Note, the function may need to return 0 when the number of users who have rated both is so low that one cannot compute a correlation coefficient.) 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.

The Correlation coefficient between 'Three Colors: Red' and 'Three Colors: Blue' is 0.760 and it is 0.222 between 'Toy Story' and 'Golden Eye'. Top 5 most related movies to 'Taxi Driver' are:

1. The Funeral
2. Casper
3. Sgt. Bilko
4. Hate (Haine, La)
5. Night Falls on Manhattan

Let's find top 5 most correlated movies regarding 'The Shining' using Correlation Coefficient:

1. Mrs. Parker and the Vicious Circle
2. Boxing Helena
3. Paths of Glory
4. Anastasia
5. Mute Witness

In this example, the threshold has been set to 10 meaning that if less than 10 have rated both movies, 0 is returned as the correlation coefficient. Lots of students failed to notice the fact that in order to compare the rating vectors, they need to be sorted by user IDs. In other words, rating vectors only can be compared element wise if and only if the rating is issued by the same user. More precisely $x[0]$ and $y[0]$ can only be compared, if they are issued by the same user, thus one need to sort both of rating vectors based on user IDs in order to obviate errors.

(d) Provide some brief thoughts on which similarity measure seems to work 'better', in the sense that the computed similarity matches your intuitive sense of similarity. Why do you think this is? Explain.

While Jaccard Coefficient totally ignores the ratings, it works better. In addition to this, Jaccard Coefficients works pretty well as ti finds movies with the same genre. One problem with Jaccard Coefficient is that it favors popular movies without considering the ratings. Thus one wise choice would be finding similar movies based on Jaccard coefficient and then rank them using Correlation Coefficient, in order to consider rankings as well.