Introduction to machine learning Exercise set 2 report

Elias Annila 014328901

Pen and paper:

1. a)

If the term occurs in every document so that $m/df_i = 1$, tf_i will always be *0.* If the term occurs in only one document $tf_{ij}' = tf_{ij} \log m$. If $m>1$ $tf_{ij}' > 0$ and the larger the m or tf_{ij} is, larger tf_{ij}' will be.

b)

Overall tf_{ij}' is always positive number and it gets larger the proportion of the documents containing the i^{th} term decreases, or the amount of terms found in the j^{th} document increases. The transformation then tells how common i^{th} term is in j^{th} document adjusted by how common term it is overall. It might be useful for deciding if i^{th} term is an important term in the j^{th} document or not.

c)

Instead of terms and documents the same logic could be extended to finding defining features in any object containing some features. For example if a program was given images of handwritten digits classified according to the digit that the image presents and the task was to find what features (locations of black/white pixels) define each character the transformation could be used to find features that are common in images representing one digit, but not in the . For example if all the digits were underlined the transformation would give only a small significance to that feature as a feature defining any digit.

2. a)

Possible values for cosine range from -1 to 1.

b)

Objects having cosine similarity of 1 are not necessarily identical. Cosine similarity measures the angle between the two vectors, not their magnitude, so the vectors may be of different length although they have a cosine similarity of 1. For example if x=(1,2,3) and y (1,2,3) their cosine similarity is 1 as they are identical. However scaling x by 2 so that x' = (2, 4, 6) will still result in cosine similarity of 1 between x' and y.

c)

If values of variables x and y have been shifted so that mean of x and y are 0 correlation between x and y is equivalent to cosine similarity between x and y. The values need to be shifted as correlation allows adding a constant to the variables without effect on correlation and cosine similarity does not. Say x = (1,2,3) and y (1,2,3) both correlation and cosine similarity between x and y are 1, but if x' = (2,3,4) cos(x,y) < 1 and correlation is still 1. Dividing datasets x and y by their standard deviations does not have an effect on similarity between cosine similarity and correlation as dividing is equivalent to multiplying and as mentioned earlier cosine similarity and correlation are both invariant to multiplying.

d)

Cosine similarity between vectors p and q is $\frac{p\cdot q}{\|p\|\|q\|}=\frac{\sum_{i=1}^{n} p_i q_i}{\sqrt{\sum_{i=1}^{n} p_1^2}\sqrt{\sum_{i=1}^{n} q_1^2}}$

as ||p||=||q|| = 1, cosine similarity between p and q is $p\cdot q=\sum_{i=1}^{n} p_i q_i$

Euclidean distance between points p and q in n dimensional space is

$$\sqrt{\sum_{i=1}^{n}(p_i-q_i)^2}=\sqrt{\sum_{i=1}^{n}(p_i^2-2p_i q_i+q_i^2)}=\sqrt{\sum_{i=1}^{n} p_i^2+\sum_{i=1}^{n}(-2p_i q_i)+\sum_{i=1}^{n} q_i^2}$$

we can see that cosine similarity includes some same terms and as ||p||=||q|| = 1:

$$distance(p,q)=\sqrt{\|p\|-2\sum_{i=1}^{n}(p_i q_i)+\|q\|}=\sqrt{1-2\,p\cdot q+1}=\sqrt{2-2\cos(p,q)}=\sqrt{2(1-\cos(p,q))}$$

e)

$$correlation(p,q)=\frac{covariance(p,q)}{std(p)\,std(q)}$$

as the data has been normalized the standard deviation of p and q are 1 and their means are 0 so:

$$correlation(p,q)=covariance(p,q)=\frac{1}{n}\sum_{i=1}^{n}((p-mean(p))(q-mean(q)))=\frac{1}{n}\sum_{i=1}^{n}(p_i q_i)$$

as above $distance(p,q)=\sqrt{\sum_{i=1}^{n} p_i^2-2\sum_{i=1}^{n}(p_i q_i)+\sum_{i=1}^{n} q_i^2}$ as p and q are normalized $\sum_{i=1}^{n} p_i^2=n$

so $distance(p,q)=\sqrt{2n-2\sum_{i=1}^{n}(p_i q_i)}=\sqrt{2n-2n\,correlation(p,q)}=\sqrt{2n(1-correlation(p,q))}$

3. a)

To define a proximity of two sets I could choose from both an object with smallest average proximity to other objects in the set and then use proximity of these two representative objects as the proximity of the two sets.

Alternatively I could compute the smallest proximity from each point of one set to any point on the other set and use the largest of these proximities as the proximity between the two sets.

b)

Distance between two sets using latter of the above mentioned methods:

distance between points p and q: $distance(p,q)=\sqrt{\sum_{i=1}^{n}(p_i-q_i)^2}$

distance from a point p to a set s would be minimum distance from a point to any point in the set:

$$distance(p,s)=min(distance(p,s_1),\ldots,distance(p,s_n))$$

distance between two sets a and b would be maximum distance between any point of a to set b:

$$distance(a,b)=max(distance(a_{1,}b),\ldots,distance(a_n,b))$$

So Euclidean distance between two sets of m objects in n dimensional space would be:

$$distance(a,b)=max(min(distance(a_{1,}b_1),\ldots,distance(a_1,b_m)),\ldots,min(distance(a_m,b_1),\ldots,distance(a_m,b_m)))$$

$$distance(a,b)=max(min(\sqrt{\sum_{i=1}^{n}(a_{1i}-b_{1i})^2},\ldots,\sqrt{\sum_{i=1}^{n}(a_{1i}-b_{mi})^2},\ldots,min(\sqrt{\sum_{i=1}^{n}(a_{mi}-b_{1i})^2},\ldots,\sqrt{\sum_{i=1}^{n}(a_{mi}-b_{mi})^2})))$$

c)

proximity between two sets a and b:

$$prox(a,b)=max(min(prox(a_{1,}b_1),\ldots,prox(a_1,b_m)),\ldots,min(prox(a_m,b_1),\ldots,prox(a_m,b_m)))$$

Programming:

b)

- Jaccard coefficent for Three Colors: Red and Three Colors: Blue: 0.598
- 5 movies with highest Jaccard coefficients to Taxi Driver excluding Taxi Driver itself:
    - Godfather: Part II, The (1974) Jaccard: 0.417
    - Clockwork Orange, A (1971) Jaccard: 0.40
    - Citizen Kane (1941) Jaccard: 0.397
    - Chinatown (1974) Jaccard: 0.394
    - Psycho (1960) Jaccard: 0.385
- I searched for top 5 movies with highest Jaccard coefficients for movie Aladdin (id: 95) and the results seem sensible enough, as among top five scorers there are 3 other Disney animations. Top 5:
    - Lion King, The (1994) Jaccard: 0.614
    - Beauty and the Beast (1991) Jaccard: 0.577
    - Jurassic Park (1993) Jaccard: 0.514
    - Snow White and the Seven Dwarfs (1937) Jaccard: 0.492
    - E.T. the Extra-Terrestrial (1982) Jaccard: 0.491

c)

- Correlationcoefficient of ratings between Toy Story and Golden Eye: 0.222

- Correlationcoefficient of ratings between Tree Colors: Red and Tree Colors: Blue: 0.760
- For calculating correlation it seemed to me that only requirement would be that standard deviation of both movie's scores had to be non zero, so this would require at least two users having seen both movies. Still it turned out that in all the high correlation movies amount of people having seen both movies was quite low. For example for Taxi Driver the top 5 list was:
  - Lost in Space (1998) Correlation: 0.951 Users having rated both movies: 7
  - Twin Town (1997) Correlation: 0.945 Users having rated both movies: 3
  - Bye Bye, Love (1995) Correlation: 0.905 Users having rated both movies: 4
  - Traveller (1997) Correlation: 0.899 Users having rated both movies: 4
  - Unzipped (1995) Correlation: 0.881 Users having rated both movies: 7

  If higher number of users was requested the list turned out to be significantly different. For example with requirement for 50 users having rated both movies the list was:

  - Chinatown (1974) Correlation: 0.477 Users having rated both movies: 93
  - Psycho (1960) Correlation: 0.459 Users having rated both movies: 117
  - L.A. Confidential (1997) Correlation: 0.445 Users having rated both movies: 67
  - Glory (1989) Correlation: 0.442 Users having rated both movies: 74
  - Dr. Strangelove or: How I Learned to Stop Worrying and Love the Bomb (1963) Correlation: 0.423 Users having rated both movies: 102

  At a glance the later list seems to correspond much better to both what Jaccard coefficent based method suggested as well as to how similar movies seemed to be based on quickly taking a look at them in IMDB. Still further increasing the amount of people who had to have seen the movie resulted in lists getting saturated with very popular movies. Also increasing amount of people who had to see the movie will also mean that no recommendation can be given on less popular movies. For these reasons it would seem sensible to adjust the requirement based on popularity of the movie. For example one fifth of the people who had seen the movie seemed to result in somehow sensible lists.
- For movie Aladdin the list obtained in above mentined manner was as follows:
  - Fox and the Hound, The (1981) Correlation: 0.591 Users having rated both movies: 56
  - Aristocats, The (1970) Correlation: 0.56 Users having rated both movies: 48
  - Lion King, The (1994) Correlation: 0.514 Users having rated both movies: 167
  - Robin Hood: Prince of Thieves (1991) Correlation: 0.486 Users having rated both movies: 59
  - Cinderella (1950) Correlation: 0.481 Users having rated both movies: 98

This seems to bring up similar movies quite well. Certainly a lot better than allowing low numbers of people having seen the movie which resulted in following list:

- Palmetto (1998) Correlation: 0.982 Users having rated both movies: 3
- Gumby: The Movie (1995) Correlation: 0.971 Users having rated both movies: 3
- Jason's Lyric (1994) Correlation: 0.966 Users having rated both movies: 6
- Panther (1995) Correlation: 0.962 Users having rated both movies: 4
- Bushwhacked (1995) Correlation: 0.926 Users having rated both movies: 6

d)

- Both methods seemed to bring up somewhat similar movies, and which is better, seems to depend on the movie. For example in case of Aladdin, which is a popular cartoon, correlation method seemed to result in other popular cartoons, whereas Jaccard coefficient resulted in list containing only more popular movies, both cartoons and others. On the other hand for movies with small amount of people having seen them Jaccard coefficient seems to also work quite well. For example in case of Three Colors: Red Jaccard coefficient included on the top 5 list both Three Colors: White and Three Colors: Blue whereas correlation coefficient left out Three Colors: White. As a bottom line Jaccard coefficient seems to be unsuitable for very popular movies, and Correlation coefficient unsuitable for movies with very few people having seen them.