Data mining

1. Question 1

1.1. Jaccard distance

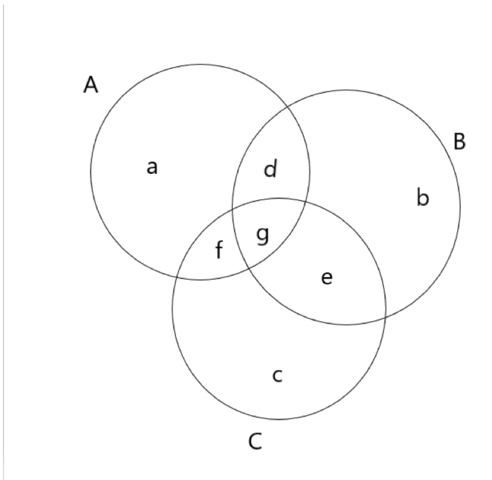

图 1: *Set figure*

$$d(A,B) = 1 - \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cup B| - |A \cap B|}{|A \cup B|}$$

Non-negativity:

$d(A,B) >= 0$ since $|A \cup B| \geq |A \cap B|$

Identity of indiscernible:

When $d(A,B) = 0$,then $|A \cup B| = |A \cap B|$,so A = B.And vice versa.

Symmetry:

$d(A,B) = d(B,A)$.This is obvious.

Triangle inequality:

We want to prove $d(A,B) + d(B,C) \geq d(C,A)$,to prove this,based on the figure,we need to prove:

$$\frac{a+b+e+f}{a+b+d+e+f+g} + \frac{b+c+d+f}{b+c+d+e+f+g}$$

$$\leq \frac{a+c+d+e}{a+c+d+e+f+g}$$

put it in WolfarmAlpha tool,the equation can be simplified to

$$\begin{aligned}(&a^2b+a^2c+a^2d+a^2f+ab^2+2abc+2abd+2abe+4abf+2abg+ac^2+\\ &2acd+2ace+4acf+2acg+ad^2+ade+4adf+adg+\\ &3aef+3af^2+3afg+b^2c+b^2d+b^2e+2b^2f+2b^2g+bc^2+\\ &2bcd+2bce+4bcf+2bcg+bd^2+2bde+5bdf+3bdg+\\ &be^2+5bef+3beg+4bf^2+6bfg+2bg^2+c^2e+c^2f+cde+\\ &3cdf+ce^2+4cef+ceg+3cf^2+3cfg+2d^2f+4def+\\ &4df^2+4dfg+2e^2f+4ef^2+4efg+2f^3+4f^2g+2fg^2)/\\ &((a+b+d+e+f+g)(a+c+d+e+f+g)(b+c+d+e+f+g)) \geq 0\end{aligned}$$

图 2: *The simplicity*

And a,b,c,d,e,f,g are all non-negative,so the result always holds.The conclusion is right.

1.2. Cosine distance

Cosine distance is not a metric.

Let$A = (-1,0), B = (1,1), C = (0,1)$.Then

$$dist(A,C) = -\frac{1}{\sqrt{2}}, dist(B,C) = \frac{1}{\sqrt{2}}, dist(C,A) = 0$$

$$dist(C,A) + dist(A,B) = -\frac{1}{\sqrt{2}} < \frac{1}{\sqrt{2}} = dist(B,C)$$

The triangle inequality does not holds.

1.3. Edit distance

Non-negativity:

$d(A,B) \geq 0$ this is obvious due to the definition.

Identity of indiscernible:

When $d(A,B) = 0$,then A and B will be the same string,so A = B.And vice versa.

Symmetry:

$d(A,B) = d(B,A)$.This is obvious.

Triangle inequality:

We want to prove $d(A,B) + d(B,C) \geq d(A,C)$.Suppose not,then $d(A,B) + d(B,C) < d(A,C)$,this means if we want to change A to C,we can first change A to B and then change B to C.However $d(A,C)$ gives the smallest edit distance,but A to B to C gives another smaller edit distance.This contradicts to the definition of the edit distance.So the triangle inequality must holds.

So,edit distance is a metric.

1.4. Hamming distance

Hamming Distance 表示两个等长字符串在对应位置上不同字符的数目。此处以 d(x,y) 指代。

hamming distance 是 metric。证明:

Non-negativity:

When d(x,y) reaches its min,it means x and y are the same string,this time d(x,y)=0

Identity of indiscernible: If d(x,y)=0,then for any i,$x_i = y_i$,so x=y

If x=y,then by the definition,d(x,y)=0

Symmetry: By the definition,d(x,y)=d(y,x)

Triangle inequality: First suppose the length of x,y,z is n.Suppose a and c are different in 0 to n_0,and same in n_0+1 to n.Then $d(x,z) = n_0$.

For y,the letter in 0 to n_0 will at least be different from one of x_i or z_i.So

$$d(x[0,n_0], y[0,n_0]) + d(y[0,n_0], z[0,n_0]) \geq 0$$

and let $n_1 = n_0 + 1$,then

$$d(x[n_1,n], y[n_1,n]) + d(y[n_1,n], z[n_1,n]) \geq 0$$

So combine them,

$$d(x,y) + d(y,z) \geq d(x,z)$$

So Hamming distance is a metric.

2. Question 2

Set the segment on $[0,L]$,let X and Y be the two point on the segment.Then it is easily to know that X and Y are independent.And they holds uniform distribution.So the joint density function is

$$f(x,y) = \begin{cases} \frac{1}{L^2} & 0 \leq x \leq L, 0 \leq y \leq L \\ 0 & otherwise \end{cases}$$

The distance is $T = |X - Y|$

$$E(T) = E(|X-Y|) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} |x-y| f(x,y) dxdy$$

$$= \int_0^L \int_0^L |x-y| \frac{1}{L^2} dxdy = 2\int_0^L dx \int_0^x (x-y)\frac{1}{L^2} dy$$

$$= \int_0^L \frac{x^2}{L^2} dx = \frac{1}{3}L$$

So the average distance is $\frac{L}{3}$.

3. Question 3

(This refers to a princeton solution however I really do not fully understand) We only need to prove for any k-rank matrix C,there is

$$\|A - B\|_F \leq \|A - C\|_F$$

Consider the C which can make $\|A - C\|_F$reach min,we set it to be C_m,and its vector space's dim will at most be k,and every line of C_m represents the projection corresponding to the line of A.(Or you can replace C_m's line with A's line,this do not change V and not increase the rank of C_m,and this can decrease $\|A - C\|_F$)

And the line of C_m is the projection corresponding to the line of A,so $\|A - C_m\|_F$ is the sum square distance for A to V's line,and the line of B is the line of A to A's k-st singular vector and they form the space,so the B satisfies B to A's k dim subspace has the least square sum,so for any k-rank metrix C:

(I mean this is disorder.But I do not fully understand it)

$$\|A-B\|_F \leq \|A-C_m\|_F \leq \|A-C\|_F$$

Here is my reference:
https://www.cs.princeton.edu/courses/archive/spring12/cos598C/svdchapter.pdf

4. Question 4

D represent the Jaccard similarity.k represents $|S \cap T|$

$$E(D)=\sum_{k=0}^{m} P(D=d(k))d(k)$$

$$=\sum_{k=0}^{m} \frac{C_n^m C_m^k C_{n-m}^{m-k}}{C_n^m C_n^m}\left(\frac{k}{2m-k}\right)$$

$$=\sum_{k=0}^{m} \frac{C_m^k C_{n-m}^{m-k}}{C_n^m} \frac{k}{2m-k}$$