CSCI5523 Assignment 1 Answer

Daihui DOU ddou00005@umn.edu 5514178

October 1, 2018

1 Question1(From Text Book Chapter 2.)

1.1 Q3

(a) The boss is right. The satisfaction should be measured by (number of complaints)/(number of sold products)
(b) The original product satisfaction attribute type is meaningless because even if the best-selling products has the most complaints, it may still have a very low complaint ratio.

1.2 Q4

(a) The marketing director is in trouble. His approach does not work if, for example, a customer prefer 1 to 2, 2 to 3 and 3 to 1.
(b) A possible solution is to have customers compare only 1 and 2, then 2 and 3. In general, it's difficult to create an ordinal measurement scale based on pairwise comparisons because customers' inconsistent.
(c) The product's average can't be directly computed from rankings because it's a ordinal scale instead of interval or ratio. Another possible approach could be computing its median value.

1.3 Q7

Daily temperature is likely to show more temporal auto-correlation because the temperature between consecutive days tends to be more similar than rainfall, considering that rainfall is much more capricious.

1.4 Q9

Observational science have issues to deal with the quality of data since the observed data may contain noise and error, which is similar to what we do in data mining. By contrast, experimental science has less problem with data quality.

1.5 Q15

The (a) sampling scheme make sure each group in sample data has the same ratio as that in original data, while the (b) sampling scheme does not.

1.6 Q16

(a)If a term occurs in one document then the idf will be maximum, which is tf*log(m). If a term occurs in every document, idf will be zero.
(b) The transformation reflects each term's real weight,i.e., a term appears in many documents has low weight to distinguish them, while that appears in few documents have high weight.

1.7 Q18

(a) Hamming distance = 3
Jaccard similarity = 2/(10-5) = 0.4
(b)Hamming distance is more similar to the Simple Matching Coefficient, since SMC = Hamming/(number of digits). Jaccard similarity is more similar to the cosine measure because neither of them count 0-0 matches.

(c)Jaccard similarity is more appropriate because it measures the same gene.
(d)I would use Hamming distance because human beings share $> 99.9\%$ of the same genes so we should measure the difference using Hamming distance.

1.8 Q27

1(a). $\forall \mathbf{x}, \mathbf{y}, 0 \leq \cos(\mathbf{x}, \mathbf{y}) \leq 1$, so $d(\mathbf{x}, \mathbf{y}) \geq 0$
1(b) $d(\mathbf{x}, \mathbf{y}) = \arccos(1) = 0$
2. $d(\mathbf{x}, \mathbf{y}) = \arccos(\cos(\mathbf{x}, \mathbf{y})) = \arccos(\cos(\mathbf{y}, \mathbf{x})) = d(\mathbf{y}, \mathbf{x})$
3. If $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{z}$ are in the same plane, then obviously $d(\mathbf{x}, \mathbf{z}) < d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z})$. It they are not in the same plane, let $\mathbf{y}'$ be the projection of $\mathbf{y}$ in $\mathbf{x} - \mathbf{z}$ plane, then $d(\mathbf{x}, \mathbf{z}) < d(\mathbf{x}, \mathbf{y}') + d(\mathbf{y}', \mathbf{z}) < d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z}))$.

2 Question 2

Construct a $N \times N$ matrix M for clustering where M_{ij} represents the similarity value between object i and object j.

3 Question 3

Construct a matrix to store data where each row is a protein object. The text description of the protein is stored in the first column beginning with '>'. The amino acid code of the protein sequence is stored in the following columns.

4 Question 4

The original $n \times m$ image matrix A could be decomposed to

$$A = U\Sigma V^T$$

where U is an $n \times n$ unitary matrix, Σ is a diagonal $n \times m$ matrix with non-negative real numbers on the diagonal, and V^T is an $m \times m$ unitary matrix. When performing dimension reduction using k dimensions, we reduce the dimension of U, Σ and V by choosing the first k eigenvalues of U, Σ and V. Thus the result matrix

$$A_k = U_k \Sigma_k V_k^T$$

where U is an $n \times k$ matrix, Σ is a $k \times k$ matrix and V is an $k \times m$ matrix.

5 Question 5

5.1 (a)

Processing data: data/mnist_test.csv
Data Size: (10000, 785)
Result using Euclidean Distance:9558
Result using Cosine Similarity:9614
Result using Jaccard Similarity:9601
Time consumed:983.6086549758911s

5.2 (b)

Processing data: data/5d_ U_ mnist_test.csv
Data Size: (10000, 6)
Result using Euclidean Distance:6800
Result using Cosine Similarity:6668
Result using Jaccard Similarity:6813
Time consumed:454.2277569770813s

Processing data: data/5d_U*Sigma_mnist_test.csv
Data Size: (10000, 6)
Result using Euclidean Distance:6847
Result using Cosine Similarity:6687
Result using Jaccard Similarity:6846
Time consumed:432.4796097278595

Processing data: data/10d_U_mnist_test.csv
Data Size: (10000, 11)
Result using Euclidean Distance:8852
Result using Cosine Similarity:8900
Result using Jaccard Similarity:8940
Time consumed:439.0083842277527s

Processing data: data/10d_U*Sigma_mnist_test.csv
Data Size: (10000, 11)
Result using Euclidean Distance:8904
Result using Cosine Similarity:9020
Result using Jaccard Similarity:8900
Time consumed:447.56175541877747s

Processing data: data/20d_U_mnist_test.csv
Data Size: (10000, 21)
Result using Euclidean Distance:9496
Result using Cosine Similarity:9534
Result using Jaccard Similarity:9532
Time consumed:507.8388545513153s

Processing data: data/20d_U*Sigma_mnist_test.csv
Data Size: (10000, 21)
Result using Euclidean Distance:9562
Result using Cosine Similarity:9592
Result using Jaccard Similarity:9554
Time consumed:498.6824266910553s

Processing data: data/40d_U_mnist_test.csv
Data Size: (10000, 41)
Result using Euclidean Distance:9507
Result using Cosine Similarity:9554
Result using Jaccard Similarity:9561
Time consumed:478.1144211292267s

Processing data: data/40d_U*Sigma_mnist_test.csv
Data Size: (10000, 41)
Result using Euclidean Distance:9630
Result using Cosine Similarity:9690
Result using Jaccard Similarity:9647
Time consumed:482.30477929115295s

5.3 (c)

Processing data: data/7*7_mnist_test.csv...
Data Size: (10000, 50)
Result using Euclidean Distance:9378
Result using Cosine Similarity:9439
Result using Jaccard Similarity:9401
Time consumed:513.892637014389s

The code for question 5 is complied in Question5_code.zip with readme, it is also available on my github repository: `https://github.com/ddouup/CSCI5523_hw1`