Basics of Clustering Analysis

Q1. What is clustering analysis?

[Answer] Clustering analyses data objects without consulting a known class label. The objects are clustered or grouped based on the principle of maximizing the intra-class similarity and minimizing the inter-class similarity. Each cluster that is formed can be viewed as a class of objects. Clustering analysis can also facilitate taxonomy formation, that is, the organization of observations into a hierarchy of classes that group similar events together.

Q2. What is the essential difference between classification and clustering?

[Answer] Classification is a supervised learning task, which needs training examples consisting of input data and their corresponding class label(s). As a result, the supervised learning process leads to a rule for decision making; i.e., given an unseen input, the classifier predicts its class label. In contrast, clustering is an unsupervised learning task where there are only input data available without knowing their class labels. As a result, the unsupervised learning process would discover the intrinsic structure underlying the input data.

Q3. Briefly outline how to compute the *dissimilarity* or *distance* between objects described by the following types of variables:
(a) Numerical variables
(b) Asymmetric binary variables
(c) Nonmetric vector objects

[Answer]

(a) For numerical variables, the generic metric is the Minkowski Distance.

$$d(\mathbf{x}_i, \mathbf{x}_j) = \left(| x_{i1} - x_{j1} |^p + \cdots + | x_{in} - x_{jn} |^p \right)^{\frac{1}{p}}$$

In particular, the Manhattan and the Euclidean distances corresponding to the special cases of the Minkowshi distance as $p=1, 2$ are two commonly used distances.

(b) For asymmetric binary variables, we first calculate the contingency table. In computing the distance between asymmetric binary variables, the number of negative matches, t, is considered unimportant and thus is ignored in the computation.

The Contingency table reflects the matching states between two objects i, j.

object i \ object j	1	0	sum
1	q	r	$q+r$
0	s	t	$s+t$
sum	$q+s$	$r+t$	p

Thus, the asymmetric binary variable distance is

$$d(i,j) = \frac{r+s}{q+r+s}.$$

(c) For Nonmetric vector objects, we often need to measure the distance between complex objects represented by vectors. It is often easier to abandon traditional metric distance computation and introduce a nonmetric similarity function. For example, the similarity between two vectors, $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, can be denoted as a cosine measure and further use the similarity measure to define a distance.

$$\cos(\mathbf{x}_i,\mathbf{x}_j) = \frac{x_{i1}x_{j1}+\cdots+x_{in}x_{jn}}{\sqrt{x_{i1}^2+\cdots+x_{in}^2}\sqrt{x_{j1}^2+\cdots+x_{jn}^2}}$$

$$d(\mathbf{x}_i,\mathbf{x}_j) = 1-\cos(\mathbf{x}_i,\mathbf{x}_j)$$

Q4. For three objects, A: (1, 0, 1, 1), B: (2, 1, 0, 2) and C: (2, 2, 2, 1), store them in a data matrix and use Manhattan, Euclidean and Cosine distances to generate distance matrices, respectively.

[Answer]

The data matrix should be $\begin{bmatrix} 1 & 0 & 1 & 1 \\ 2 & 1 & 0 & 2 \\ 2 & 2 & 2 & 1 \end{bmatrix}$.

Using the Manhattan distance, we have

$$d(A,B) = |1-2|+|0-1|+|1-0|+|1-2| = 1+1+1+1 = 4$$
$$d(A,C) = |1-2|+|0-2|+|1-2|+|1-1| = 1+2+1+0 = 4$$
$$d(B,C) = |2-2|+|1-2|+|0-2|+|2-1| = 0+1+2+1 = 4$$

The Manhattan distance matrix is $\begin{bmatrix} 0 & 4 & 4 \\ 4 & 0 & 4 \\ 4 & 4 & 0 \end{bmatrix}$.

Using the Euclidean distance, we have

$$d(A,B)=\sqrt{(1-2)^2+(0-1)^2+(1-0)^2+(1-2)^2}=\sqrt{1+1+1+1}=2$$
$$d(A,C)=\sqrt{(1-2)^2+(0-2)^2+(1-2)^2+(1-1)^2}=\sqrt{1+4+1+0}=\sqrt{6}$$
$$d(B,C)=\sqrt{(2-2)^2+(1-2)^2+(0-2)^2+(2-1)^2}=\sqrt{0+1+4+1}=\sqrt{6}$$

The Euclidean distance matrix is $\begin{bmatrix} 0 & 2 & \sqrt{6} \\ 2 & 0 & \sqrt{6} \\ \sqrt{6} & \sqrt{6} & 0 \end{bmatrix}$.

Using the Cosine distance, we have

$$d(A,B)=1-\frac{1\times 2+0\times 1+1\times 0+1\times 2}{\sqrt{1^2+0^2+1^2+1^2}\sqrt{2^2+1^2+0^2+2^2}}=1-\frac{4}{\sqrt{3}\sqrt{9}}=\frac{3\sqrt{3}-4}{3\sqrt{3}}$$
$$d(A,C)=1-\frac{1\times 2+0\times 2+1\times 2+1\times 1}{\sqrt{1^2+0^2+1^2+1^2}\sqrt{2^2+2^2+2^2+1^2}}=1-\frac{5}{\sqrt{3}\sqrt{13}}=\frac{\sqrt{39}-5}{\sqrt{39}}$$
$$d(B,C)=1-\frac{2\times 2+1\times 2+0\times 2+2\times 1}{\sqrt{2^2+1^2+0^2+2^2}\sqrt{2^2+2^2+2^2+1^2}}=1-\frac{8}{\sqrt{9}\sqrt{13}}=\frac{3\sqrt{13}-8}{3\sqrt{13}}$$

The Cosine distance matrix is $\begin{bmatrix} 0 & \frac{3\sqrt{3}-4}{3\sqrt{3}} & \frac{\sqrt{39}-5}{\sqrt{39}} \\ \frac{3\sqrt{3}-4}{3\sqrt{3}} & 0 & \frac{3\sqrt{13}-8}{3\sqrt{13}} \\ \frac{\sqrt{39}-5}{\sqrt{39}} & \frac{3\sqrt{13}-8}{3\sqrt{13}} & 0 \end{bmatrix}$.

Q5. Given two binary objects, A: (1 0, 1, 0) and B: (1, 1, 1, 0), calculate symmetric and asymmetric binary variable distances between them.

[Answer]
In order to calculate the distances, we need to have a contingency table for two objects A and B as follows (for the notation of this table, see the answer to Q3(c)):

2	0
1	1

Therefore, we can calculate distances with the above contingency table.

$$d_S(A,B)=\frac{0+1}{2+0+1+1}=\frac{1}{4},\quad d_{AS}(A,B)=\frac{0+1}{2+0+1}=\frac{1}{3}.$$