8.4 The following table contains the attributes *name*, *gender*, *trait-1*, *trait-2*, *trait-3*, and *trait-4*, where *name* is an object-id, *gender* is a symmetric attribute, and the remaining *trait* attributes are asymmetric, describing personal traits of individuals who desire a penpal. Suppose that a service exists that attempts to find pairs of compatible penpals.

name	*gender*	*trait-1*	*trait-2*	*trait-3*	*trait-4*
Kevan	M	N	P	P	N
Caroline	F	N	P	P	N
Erik	M	P	N	N	P
⋮	⋮	⋮	⋮	⋮	⋮

For asymmetric attribute values, let the value *P* be set to 1 and the value *N* be set to 0.

Suppose that the distance between objects (potential penpals) is computed based only on the asymmetric variables.

(a) Show the *contingency matrix* for each pair given Kevan, Caroline, and Erik.

(b) Compute the *simple matching coefficient* for each pair.

(c) Compute the *Jaccard coefficient* for each pair.

(d) Who do you suggest would make the best pair of penpals? Which pair of individuals would be the least compatible?

(e) Suppose that we are to include the symmetric variable *gender* in our analysis. Based on the Jaccard coefficient, who would be the most compatible pair, and why?

8.5 What is clustering? Briefly describe the following approaches to clustering methods: *partitioning* methods, *hierarchical* methods, *density-based* methods, *grid-based* methods, and *model-based* methods. Give examples in each case.

8.6 Suppose that the data mining task is to cluster the following eight points (with (x, y) representing location) into three clusters.

$$A_1(2, 10), A_2(2, 5), A_3(8, 4), B_1(5, 8), B_2(7, 5), B_3(6, 4), C_1(1, 2), C_2(4, 9).$$

The distance function is Euclidean distance. Suppose initially we assign A_1, B_1, and C_1 as the center of each cluster, respectively. Use the *k-means* algorithm to show *only*

(a) the three cluster centers after the first round execution, and

(b) the final three clusters.

8.7 Use a diagram to illustrate how, for a constant *MinPts* value, *density-based clusters* with respect to a higher density (i.e., a lower value for ϵ, the neighborhood radius) are completely contained in density-connected sets obtained with respect to a lower density.