CSIT5210 Data Mining and Knowledge Discovery (Fall Semester 2021)
Exercise 4
Subspace Clustering

Q1 Density-Based Subspace Clustering

Consider three dimensions (X, Y, Z). There are the following 10 three-dimensional data points.

(11, 13, 5), (12, 11, 21), (11, 17, 27), (13, 14, 38), (22, 37, 36),
(24, 31, 27), (25, 35, 21), (29, 34, 4), (35, 5, 4), (36, 6, 5).

Suppose each dimension ranges from 1 to 40.
Assume that the grid size of each dimension is 10. For example, dimension X has 4 grids or units, namely X1, X2, X3 and X4, where X1, X2, X3 and X4 correspond to [1, 10], [11, 20], [21, 30] and [31, 40], respectively.

Consider the density-based subspace clustering. Let the density threshold be 40%.

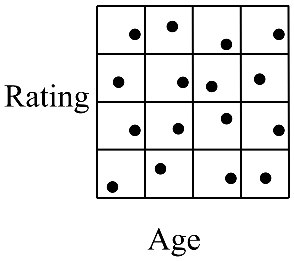
(a) Find all subspaces containing the dense units.
(b) Identify clusters in each subspace containing dense units

Q2 Entropy

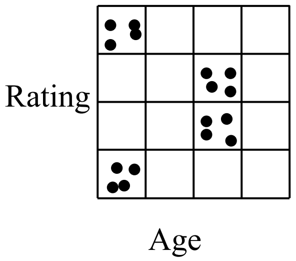
X \ Y	1	2
1	3/8	1/8
2	1/8	3/8

Consider the above table. What are H(X) and H(Y)?

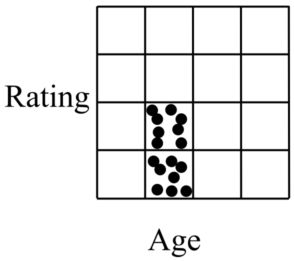
Q3 Density-Based Entropy



Case (a)



Case (b)



Case (c)

In each of the above three cases, what is the entropy H(Age, Rating)?

Q4 Conditional Entropy

X \ Y	1	2
1	1/4	0
2	1/4	1/2

(a) Calculate the conditional entropy of H(X|Y).

(b) Calculate H(X|Y) as

$$- \sum_{x\in A} \sum_{y\in B} p(x, y) \log p(x|y)$$

where A = {1, 2} and B = {1, 2}.

Q5 Entropy-Based Subspace Clustering

Consider three dimensions (X, Y, Z). There are the following 10 three-dimensional data points.

(11, 13, 5), (12, 11, 21), (11, 17, 27), (13, 14, 38), (22, 37, 36),
(24, 31, 27), (25, 35, 21), (29, 34, 4), (35, 5, 4), (36, 6, 5).

Suppose each dimension ranges from 1 to 40.
Assume that the grid size of each dimension is 10. For example, dimension X has 4 grids or units, namely X1, X2, X3 and X4, where X1, X2, X3 and X4 correspond to [1, 10], [11, 20], [21, 30] and [31, 40], respectively.

Consider the entropy-based subspace clustering. Let the entropy threshold be 2.0.
Find all subspaces containing good clusters.