DASH: ANS

Exam 2022

Version A

Solutions

Consider the D dataset below to answer questions along the exam:

	y1	y2	y3	y4	class	cluster
$\mathbf{x}_1$	-2	2	B	D	X	C1
$\mathbf{x}_2$	3	4	A	C	X	C2
$\mathbf{x}_3$	0	4	A	C	X	C1
$\mathbf{x}_4$	-2	2	A	D	Y	C1

I. Clustering [6.1v]

Given D and distance $d(\mathbf{x}_A, \mathbf{x}_B) = Manhattan(\mathbf{x}_A, \mathbf{x}_B | y_1, y_2) + Hamming(\mathbf{x}_A, \mathbf{x}_B | y_3, y_4)$

1. [0.5v] Complete the following pairwise distance matrix

	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$
$\mathbf{x}_1$	0	9	?	?
$\mathbf{x}_2$		0	3	8
$\mathbf{x}_3$			0	5
$\mathbf{x}_4$				0

$$d(\mathbf{x}_1, \mathbf{x}_3) = 6, \;\; d(\mathbf{x}_1, \mathbf{x}_4) = 1$$

2. [1v] Can the given clustering solution be obtained by an agglomerative algorithm under *single* link? Justify by presenting the final dendrogram.

No. Dendrogram: $\{\{\mathbf{x}_1, \mathbf{x}_4\}[1]\ \{\mathbf{x}_2, \mathbf{x}_3\}[3]\ \}[5]$

3. [1.2v] Let $\mathbf{x}_1$ and $\mathbf{x}_4$ be the initial centroids of k-means. Compute *one* iteration of the k-means, identifying the new centroids using *medoid* averaging criteria.

After iteration: $\{\mathbf{x}_1\}$, $\{\mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4\}$

Medoids: $\mathbf{x}_1$ and $\mathbf{x}_3$

4. [0.6v] Using $d(\mathbf{x}_A, \mathbf{x}_B)$, identify the silhouette of observation $\mathbf{x}_4$.

$$silhouette(\mathbf{x}_4) = 1 - \frac{3}{8} = \frac{5}{8}$$

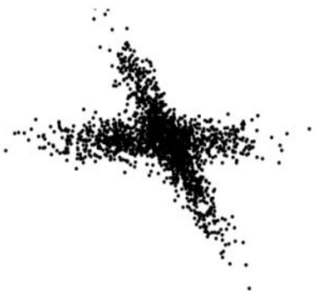
5. [0.8v] Consider *class* to be our ground truth, compute the purity of the clustering solution.

$$purity = \frac{1}{4}(1+2) = 0.75$$

6. [0.5v] Select the limitations of the k-Means algorithm (*i.e.* the true statements):

a) dependent on initialization/seeding ◀
b) sensitive to outliers under *mean* centroid criteria ◀
c) not suitable to discover clusters with irregular/non-convex shapes ◀
d) dependent on the specification of a proper linkage criterion

7. [0.5v] Given the following data plot (*right*), select the proper clustering stances to recover its clusters:



a) model-based clustering ◀
b) density-based clustering
c) soft clustering ◀
d) hard clustering
e) partition-based clustering

8. [1v] Classify the following statements as *True* or *False*:

a) Clustering becomes semi-supervised when pairs of observations are known to belong to the same cluster. True
b) Agglomerative clustering algorithms allow to manually select a desirable number of clusters once a dendrogram is inferred. True
c) Complete (maximum) link criterion tends to break large clusters and is biased towards globular clusters. True
d) A rand index that is close to zero suggests that the clustering algorithm was unable to guarantee high cluster dissimilarity. False

II. **Dimensionality reduction** [2.7v]

Consider that the application PCA over the numeric variables of D produced the following covariance matrix, eigenvectors and eigenvalues:

$$C = \begin{pmatrix} 5.58 & 2.33 \\ 2.33 & 1.33 \end{pmatrix}, \ \mathbf{v}_1 = ?, \ \mathbf{v}_2 = \begin{pmatrix} 0.9 \\ 0.4 \end{pmatrix}, \ \lambda_1 = 0.302, \ \lambda_2 = 6.614$$

9. [1v] What is the percentage of data variability explained by eigenvector $\mathbf{v}_2$?

$$\frac{\lambda_2}{\lambda_1+\lambda_2} = 95.6\%$$

10. [1.2v] Project the numeric values of D to the reduced space using $\mathbf{v}_2$.

	$c_2 = 0.9y_1 + 0.4y_2$
$\mathbf{x}_1$	-1
$\mathbf{x}_2$	4.3
$\mathbf{x}_3$	1.6
$\mathbf{x}_4$	-1

11. [0.5v] Identify the eigenvector $\mathbf{v}_1$.

Solving $C\mathbf{v}_1 = \lambda_1\mathbf{v}_1$ equations (and optional normalization) yields $\mathbf{v}_1 \approx \begin{pmatrix} -0.4 \\ 0.9 \end{pmatrix}$

III. **Pattern Mining** [6.95v]

12. [1.7v] Selecting y_3 and y_4, identify all the closed and maximal frequent itemsets with a relative support above 0.5.

closed: $A[sup = 3], AC[sup = 2], D[sup = 2]$

maximal: $AC[sup = 2], D[sup = 2]$

13. [0.8v] Given the association rule, $AC \Rightarrow X$, compute its support, confidence and lift.

$$sup(R) = sup(ACX) = 0.5, conf(R) = \frac{sup(ACX)}{sup(AC)} = 1, lift(R) = \frac{conf(R)}{sup(X)} = \frac{4}{3}$$

14. [1v] Consider that we have access to additional observations, leading to the following re-evaluation of rule

$$AC \Rightarrow X\ [\text{support} = 0.5, \text{Binomial}\ p\text{value} = 1E-3, \text{confidence} = 0.8, \text{lift} = 0.99]$$

Classify the following statements as *True* or *False*:

a) Assuming a significance level $\alpha = 0.1$, the given pattern is not statistically significant False

b) The given lift suggests an interesting/strong association rule False

c) The given lift suggests that the consequent, X, is highly frequent (support>0.5) True

d) If AC is a frequent itemset, a superset (e.g. ACX) is also frequent (monotonicity) False

15. [1.4v] Selecting y_1 and y_2, identify the largest constant bicluster and the largest order-preserving bicluster with δ=0 and no noise ($\varepsilon = 0$)

Constant (I={x1,x4},J={y1,y2}), Order-preserving (I={x1,x2,x3,x4},J={y1,y2})

16. [0.8v] Given the additive bicluster (I={x1,x2,x3,x4},J={y1,y2}) and δ=0, compute its quality.

Considering additive factors $\{\gamma_1=0, \gamma_2=2, \gamma_3=2, \gamma_4=0\}$, the quality is 7/8

17. [0.75v] Classify the following statements as *True* or *False*:

a) A biclustering solution with 2 biclusters with overlapping elements is always non-exhaustive on rows and columns. False

b) Given a biclustering search, a statistically significant bicluster that was not retrieved by this search is termed false positive. False

c) The coherence strength of a bicluster determines the deviations from expectations. True

18. [0.5v] Which of the following actions generally increase the average size of patterns in a solution (where size is the number of elements, i.e. support × pattern length):

a) increase tolerance to noise (i.e. decrease quality) ◀

b) choose closed pattern representations instead of all patterns ◀

c) given perfect quality, increase the cardinality of variables in discrete data

d) decrease coherence strength (higher deviations allowed) in real-valued data ◀

IV. **Outlier analysis** [1.25v]

19. Classify the following statements as *True* or *False*:

a) Given specific context variables, a contextual outlier observation is an observation that significantly deviates from other observations that share the same context. True

b) A collective outlier is an observation that deviates from neighbour observations. False

c) Observations in clusters with bad cohesion (sparse clusters) are outlier candidates. True

d) Given a data where a few observations are annotated with *normal/non-outlier* tag, these observations should be removed to better detect outliers. False

e) Density-based outlier analysis approaches can be used to identify local outliers. True

V. **Learning from Complex Data** [3v]

20. Classify the following statements as *True* or *False*:

a) The order of a multivariate time series corresponds to the number of time points. False

b) Pattern mining in time series can be either considered in the context of a single time series (e.g. motif discovery) or multiple time series (e.g. biclustering). True

c) When computing the distance between time series, Minkowski distances (e.g. Euclidean) cannot account for temporal misalignments. True

d) Statistics extracted with a sliding window along time series observations can be used to produce a multivariate dataset. True

e) As frequent itemsets are solely focused on co-occurrences, sequential patterns are solely focused on precedences. False

f) Given time series data, biclustering can be extended to accommodate time lags between observations. True

g) Nominal univariate events are also termed typed events. True

h) Complex patterns can generally be mapped into binary or numeric variables (one variable per pattern) for subsequent multivariate data analysis. True

i) The data of a system with stationary sensors producing signals at different locations can be described by a georeferenced time series structure. True

j) The spatial slicing principle suggests that is rather more important to learn global models than multiple local/regional models. False

k) Inductive logic can be used to capture associations between tables. True

END