DASH

Prototype Exam

Practical exercises

Group I. Calculus [14.6v]

Considering the following dataset, where $y_1 \in [0,3]$, y_2 is ordinal and z is a nominal target.

	y_1	y_2	z
x_1	1.2	C	X
x_2	0.2	B	X
x_3	3	A	Y
x_4	0.5	B	Y
x_5	0.3	A	Y

1. [1.5v] Considering y_2 numerical encoding, $\{\text{A}: 0, \text{B}: 1, \text{C}: 3\}$, Manhattan distance, and fully unsupervised setting. Draw the dendrogram under complete linkage.

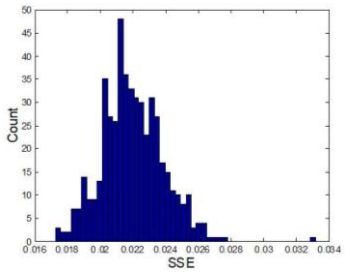
2. [1v] Are there multivariate outliers in accordance with DBSCAN (p=3, ε=3)? Which?

3. Assuming a solution with maximal purity against the output variable z.

a) [1v] Identify the medoid of the larger cluster
b) [1.5v] Identify the silhouette of the smaller cluster

4. [1v] Consider the following analysis of sum squared errors (SSE) gathered from a thousand of randomized datasets using k-means. Identify the correct statement:

i. A SSE in [0.02,0.023] is statistically significant
ii. A SSE above 0.34 is statistically significant
iii. A SSE below 0.017 is statistically significant
iv. None of above



5. Under the same numerical encoding, consider the biclustering task on $\{y_1, y_2\}$.

a) [1v] Identify the largest perfect constant assuming coherence strength δ=1.5
b) [1.1v] Compute the quality of order-preserving bicluster (I=$\{x_1, x_2, x_4, x_5\}$,J=$\{y_1, y_2\}$)

6. [1v] Binarize y_1 after standard scaling using equal-range discretization.

7. Assuming the binarization yields y_1=(Q,Q,R,Q,Q) and $\{y_1, y_2\}$ variables:

c) [1.5v] Identify the set of closed co-occurrence patterns satisfying $\min_{sup} = 0.4$
d) [2.4v] Compute the statistical significance of pattern QB and the lift of rule Q⇒B

8. [1.6v] Consider the covariance matrix C obtained from the given dataset and the corresponding eigenvectors. Which of the eigenvectors should be considered to reduce the dimensionality?

$$C = \begin{pmatrix} 1.353 & -0.225 \\ -0.225 & 1.5 \end{pmatrix}, \quad v_1 = \begin{pmatrix} -0.8 \\ -0.59 \end{pmatrix}, \quad v_2 = \begin{pmatrix} 0.59 \\ -0.8 \end{pmatrix}$$

Group II. True-and-false [5.4v] (+0.45v for correct, -0.2v for incorrect)

1. An outlier can be inconsistent with the remaining data or just its neighbourhood.
2. In supervised outlier analysis, assessing the sensitivity of a classifier is often preferred over its accuracy.
3. Contextual outliers only deviate on a compact subset of variables.
4. Hamming distance is adequate to handle ordinal variables with high cardinality.
5. k-means does not adequately identify spherical clusters.
6. Purity is biased when the number of found clusters approaches the total number of observations.
7. Spearman correlation is preferred over Pearson correlation if the order of values is more relevant than their absolute value.
8. Given a m-dimensional dataset, PCA can reconstruct any data point using $m - 1$ components of PCA with zero reconstruction error.
9. The lift measure of an association rule $\mathrm{A} \Rightarrow \mathrm{B}$ does not change if we add a new transaction that does not either contain A or B.
10. Sequential pattern mining can be applied to find frequent co-occurrences in symbolic multivariate time series data.
11. A false negative tricluster is a statistically significant subspace that was not retrieved.
12. The search for triclusters with low variance allows the discovery of additive subspaces.

END