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.

	x_1	x_2	x_3	x_4	x_5
x_1	0	3	4.8	2.7	3.9
x_2		0	3.8	0.3	1.1
x_3			0	3.5	2.7
x_4				0	1.2
x_5					0

Dendrogram: {{{{x2,x4}[0.3], x5}[1.2], x3}[3.8], x1}[4.8]

2. [1v] Are there multivariate outliers in accordance with DBSCAN (p=3, ε=3)? Which? $\{\text{x1}, \text{x3}\}$

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

a) [1v] Identify the medoid of the larger cluster

Larger cluster $\{x_3, x_4, x_5\}$
Average distances q to other observations in the cluster: $q(x_3) = 3.1, q(x_4) = 2.35, q(x_5) = 1.95$

b) [1.5v] Identify the silhouette of the smaller cluster

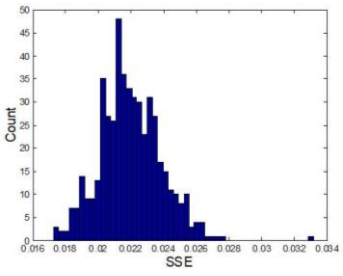
$$silhouette(x_1) = 1 - \frac{3}{3.8} = 0.21, \quad silhouette(x_2) = \frac{\frac{5.2}{3}}{3} - 1 = -0.42,$$

Note that when $a(x) > b(x)$ then a better proxy to the silhouette of an observation is $\frac{b(x)}{a(x)} - 1$

$$silhoute(c_1) = \frac{0.21 - 0.42}{2} = -0.1$$

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 with coherence strength $\delta = 0.5$

Constant: B=(I={x2,x4},J={y1,y2})

b) [1.1v] Compute the quality of order-preserving bicluster (I={x1,x2,x4,x5},J={y1,y2})

Quality $= 7/8$ (mode orders are y1 $\leq$ y2)

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

Standardly scaled y1|X = (0.14, -0.72, 1.69, -0.46, -0.64)
Two bin equal-range of N(0,1) is]-∞,0[and]0,∞[
As a result, binarized y1 = (U, L, U, L, L)

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

a) [1.5v] Identify the set of closed co-occurrences satisfying $\min_{sup} = 0.4$

All patterns: Q, A, B, QB
Closed patterns: Q, A, QB

b) [2.4v] Compute the statistical significance of pattern QB and the lift of rule Q⇒B

$$p_{null}(QD) = \frac{4}{5} \times \frac{2}{5} = \frac{8}{25} = 0.32$$

$$P(X \geq 2 \,|X \sim Binomial(p = 0.32, N = 5) = 0.51$$

$$\text{lift} = sup(QB)/(sup(Q) \times sup(B)) = (2/5)/(4/5 \times 2/5) = 1.25$$

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?

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

Given $Cv_1 = \lambda_1 v_1$, then $\lambda_1 \approx 1.189$
Given $Cv_2 = \lambda_2 v_2$, then $\lambda_2 \approx 1.66$
The second eigenvector is used to reduce the space has it explains more variability

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. True
2. In supervised outlier analysis, assessing the sensitivity of a classifier is often preferred over its accuracy. True
3. Contextual outliers only deviate on a compact subset of variables. False
4. Hamming distance is adequate to handle ordinal variables with high cardinality. False
5. k-means does not adequately identify spherical clusters. False
6. Purity is biased when the number of found clusters approaches the total number of observations. True
7. Spearman correlation is preferred over Pearson correlation if the order of values is more relevant than their absolute value. True
8. Given a m-dimensional dataset, PCA can reconstruct any data point using $m - 1$ components of PCA with zero reconstruction error. False

9. The lift measure of an association rule $A \Rightarrow B$ does not change if we add a new transaction that does not either contain A or B. False
10. Sequential pattern mining can be applied to find frequent co-occurrences in symbolic multivariate time series data. True
11. A false negative tricluster is a statistically significant subspace that was not retrieved. True
12. The search for triclusters with low variance allows the discovery of additive subspaces. False

END