Please carefully read and follow the general instructions regarding exercises. Failing to meet the requirements might lead to penalties. https://elearn.uef.fi/mod/page/view.php?id=293750

If you suspect that something is wrong with some exercise question, please contact the lecturer.

If you face persistent issues while working on an exercise, do ask for help, e.g. during a course meeting or by contacting the lecturer via email.

Consider the dataset consisting of 16 data points shown in Figure 1.

id	v_B	v_D
(1)	0.282	0.562
(2)	0.295	0.593
(3)	0.323	0.467
(4)	0.377	0.655
(5)	0.418	0.626
(6)	0.106	0.539
(7)	0.119	0.426
(8)	0.198	0.301
(9)	0.196	0.503
(10)	0.331	0.586
(11)	0.053	0.820
(12)	0.099	0.874
(13)	0.119	0.884
(14)	0.113	0.793
(15)	0.137	0.866
(16)	0.165	0.850

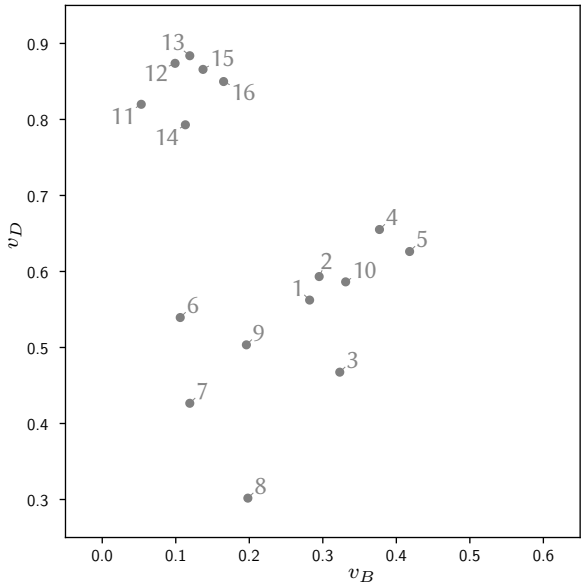

Figure 1: Dataset, as a list of data points (left) and as a plot (right)

The corresponding matrix of ℓ_2 pairwise distances is shown in Table 1.

Table 1: Matrix of pairwise distances between the data points

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
(1)	0	0.034	0.103	0.133	0.150	0.177	0.212	0.274	0.104	0.055	0.345	0.362	0.361	0.286	0.337	0.311
(2)	0.034	0	0.129	0.103	0.127	0.197	0.243	0.308	0.134	0.037	0.332	0.343	0.340	0.270	0.315	0.288
(3)	0.103	0.129	0	0.196	0.185	0.229	0.208	0.208	0.132	0.119	0.444	0.465	0.464	0.388	0.440	0.414
(4)	0.133	0.103	0.196	0	0.050	0.295	0.345	0.397	0.236	0.083	0.364	0.354	0.345	0.298	0.320	0.288
(5)	0.150	0.127	0.185	0.050	0	0.324	0.360	0.392	0.254	0.096	0.413	0.404	0.395	0.348	0.370	0.338
(6)	0.177	0.197	0.229	0.295	0.324	0	0.114	0.255	0.097	0.230	0.286	0.335	0.345	0.254	0.328	0.317
(7)	0.212	0.243	0.208	0.345	0.360	0.114	0	0.148	0.109	0.266	0.399	0.448	0.458	0.367	0.440	0.426
(8)	0.274	0.308	0.208	0.397	0.392	0.255	0.148	0	0.202	0.315	0.539	0.581	0.588	0.499	0.568	0.550
(9)	0.104	0.134	0.132	0.236	0.254	0.097	0.109	0.202	0	0.158	0.348	0.383	0.389	0.302	0.368	0.348
(10)	0.055	0.037	0.119	0.083	0.096	0.230	0.266	0.315	0.158	0	0.363	0.370	0.366	0.301	0.341	0.312
(11)	0.345	0.332	0.444	0.364	0.413	0.286	0.399	0.539	0.348	0.363	0	0.071	0.092	0.066	0.096	0.116
(12)	0.362	0.343	0.465	0.354	0.404	0.335	0.448	0.581	0.383	0.370	0.071	0	0.022	0.082	0.039	0.070
(13)	0.361	0.340	0.464	0.345	0.395	0.345	0.458	0.588	0.389	0.366	0.092	0.022	0	0.091	0.025	0.057
(14)	0.286	0.270	0.388	0.298	0.348	0.254	0.367	0.499	0.302	0.301	0.066	0.082	0.091	0	0.077	0.077
(15)	0.337	0.315	0.440	0.320	0.370	0.328	0.440	0.568	0.368	0.341	0.096	0.039	0.025	0.077	0	0.032
(16)	0.311	0.288	0.414	0.288	0.338	0.317	0.426	0.550	0.348	0.312	0.116	0.070	0.057	0.077	0.032	0

Problem 1 (Internal validation criteria).

a) Looking back at the clusterings obtained by applying the k-means algorithm and the agglomerative algorithm with complete linkage, compute the sum of square distances to centroids, the intra-cluster vs. inter-cluster distance ratio and the silhouette coefficient for the two clusterings.

b) Comment on the obtained values.

Problem 2 (Comparing clusterings).

The data points represent measurements of physical properties of dry beans.[1] The beans that constitute this small data set are of three different species. We might consider the grouping of beans by species as a reference and compare the obtained clusterings against it.

The partitioning of beans according to their species is shown in Figure 2, with *dermason*, *cali*, and *horoz* species in blue, red and yellow, respectively.

id	species
(1)	*cali*
(2)	*cali*
(3)	*cali*
(4)	*cali*
(5)	*cali*
(6)	*horoz*
(7)	*horoz*
(8)	*horoz*
(9)	*horoz*
(10)	*horoz*
(11)	*dermason*
(12)	*dermason*
(13)	*dermason*
(14)	*dermason*
(15)	*dermason*
(16)	*dermason*

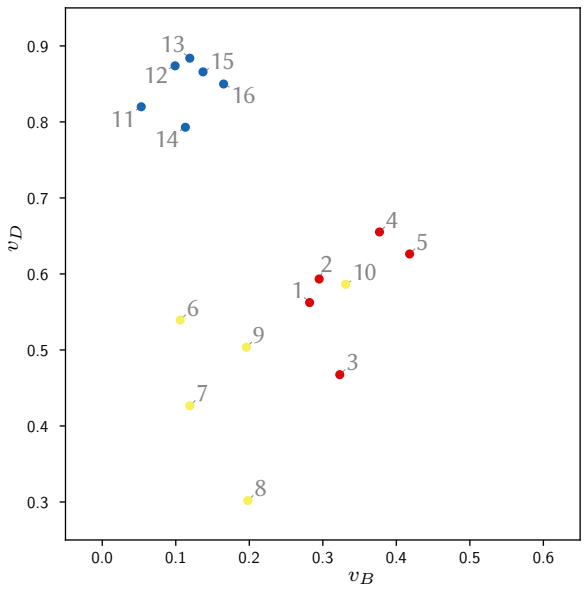

Figure 2: Partitioning of beans according to their species, as a list of cluster labels (left) and as a plot (right)

a) Compare the clusterings obtained by applying the k-means algorithm and the agglomerative algorithm with complete linkage to this reference. Write down the respective contingency matrices and compute the cluster purity, Gini index and entropy.

That is, using the notation from the lecture, you should compute Purity$(\mathcal{C}, \mathcal{C}_R)$, Gini$(\mathcal{C}, \mathcal{C}_R)$, and Entropy$(\mathcal{C}, \mathcal{C}_R)$, where $\mathcal{C}$ and $\mathcal{C}_R$ respectively represent the evaluated clustering and the reference clustering.

b) Comment on the obtained values.

[1] See https://doi.org/10.1016/j.compag.2020.105507 for details

```
### Data matrix
## id, vB, vD
1, 0.282, 0.562
2, 0.295, 0.593
3, 0.323, 0.467
4, 0.377, 0.655
5, 0.418, 0.626
6, 0.106, 0.539
7, 0.119, 0.426
8, 0.198, 0.301
9, 0.196, 0.503
10, 0.331, 0.586
11, 0.053, 0.820
12, 0.099, 0.874
13, 0.119, 0.884
14, 0.113, 0.793
15, 0.137, 0.866
16, 0.165, 0.850
```