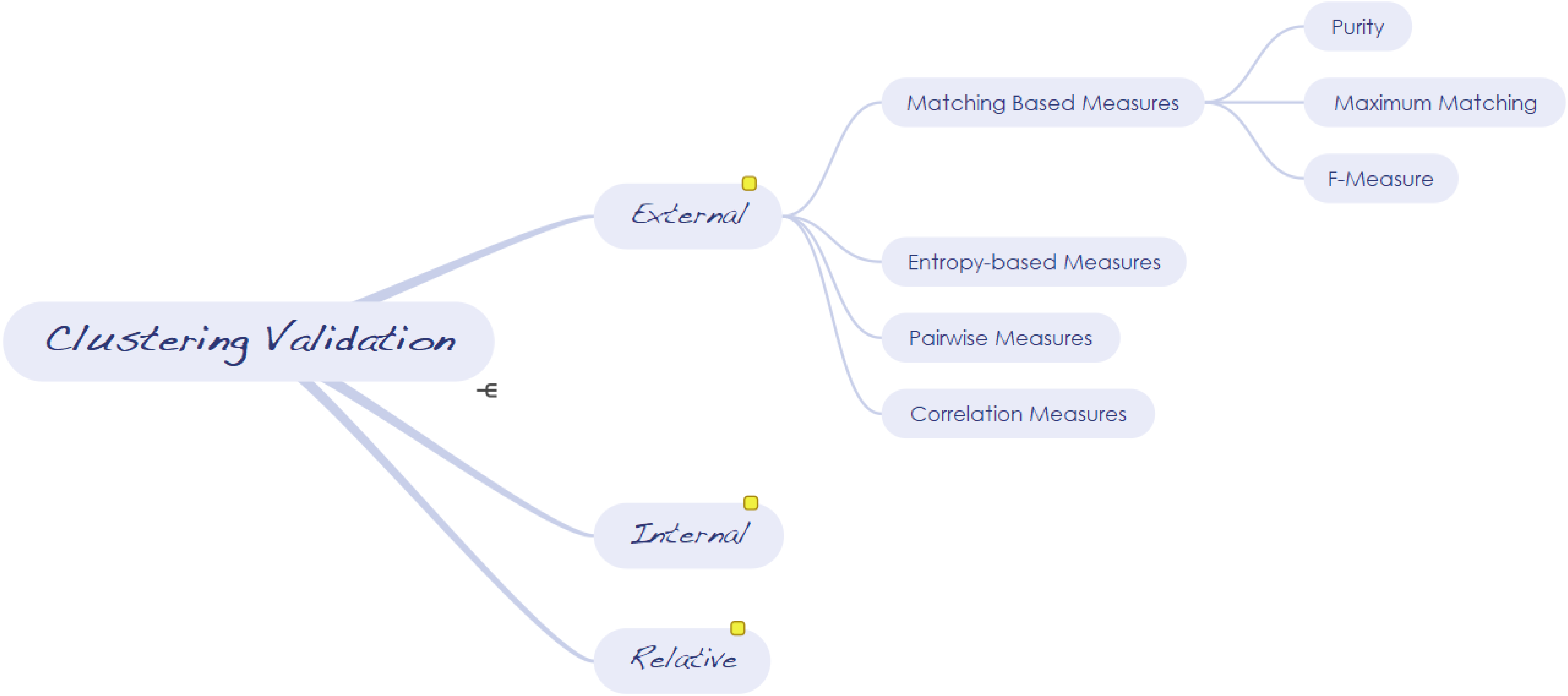
Clustering Validation

ارزیابی خوشه‌بندی


Clustering Validation
External
Matching Based Measures
Purity
Maximum Matching
F-Measure
Entropy-based Measures
Pairwise Measures
Correlation Measures
Internal
Relative

External Measures

Partitioning data into $\boldsymbol{k}$ clusters.

$$D = \{\boldsymbol{x}_i\}_{i=1}^{n}$$

The ground-truth partitioning

$$\mathcal{T} = \{T_1, T_2, \ldots, T_k\}$$

$$T_j = \{\mathbf{x}_i \in \mathbf{D} | y_i = j\}$$

Label information for $\boldsymbol{x}_i$

$$y_i \in \{1, 2, \ldots, k\}$$

Clustering via some clustering algorithm

$$\mathcal{C} = \{C_1, \ldots, C_r\}$$

The cluster label for $\boldsymbol{x}_i$

$$\hat{y}_i \in \{1, 2, \ldots, r\}$$

with the correct number of clusters:

$$r = k$$

All of the external measures rely on the $r \times k$ contingency table N that is induced by a clustering C and the ground-truth partitioning T

$$\mathbf{N}(i, j) = n_{ij} = \left|C_i \cap T_j\right|$$

$$m_j = |T_j| \qquad n_i = |C_i|$$

17.1.1 Matching Based Measures

Purity

$$purity_i = \frac{1}{n_i} \max_{j=1}^{k} \{n_{ij}\}$$

$$purity = \sum_{i=1}^{r} \frac{n_i}{n} purity_i = \frac{1}{n} \sum_{i=1}^{r} \max_{j=1}^{k} \{n_{ij}\}$$

Example 17.1. Figure 17.1 shows two different clusterings obtained via the K-means algorithm on the Iris dataset, using the first two principal components as the two dimensions. Here $n = 150$, and $k = 3$. Visual inspection confirms that Figure 17.1a is a better clustering than that in Figure 17.1b. We now examine how the different contingency table based measures can be used to evaluate these two clusterings.

Consider the clustering in Figure 17.1a. The three clusters are illustrated with different symbols; the gray ones are in the correct partition, whereas the white ones are wrongly clustered compared to the ground-truth Iris types. For instance, C_3 mainly corresponds to partition T_3 (`Iris-virginica`), but it has three points (the white triangles) from T_2. The complete contingency table is as follows:

	iris-setosa T_1	iris-versicolor T_2	iris-virginica T_3	n_i
C_1(squares)	0	47	14	61
C_2(circles)	50	0	0	50
C_3(triangles)	0	3	36	39
m_j	50	50	50	$n = 100$

To compute purity, we first note for each cluster the partition with the maximum overlap. We have the correspondence (C_1, T_2), (C_2, T_1), and (C_3, T_3). Thus, purity is given as

$$purity = \frac{1}{150}(47 + 50 + 36) = \frac{133}{150} = 0.887$$

Maximum Matching

$G = (V, E)$ $\quad (C_i, T_j) \in E \quad V = \mathcal{C} \cup \mathcal{T}$

weight $w(C_i, T_j) = n_{ij}$

for all $C_i \in C$ and $T_j \in T$

where the weight of a matching M is simply the sum of all the edge weights in matching M.

$$w(M) = \sum_{e \in M} w(e)$$

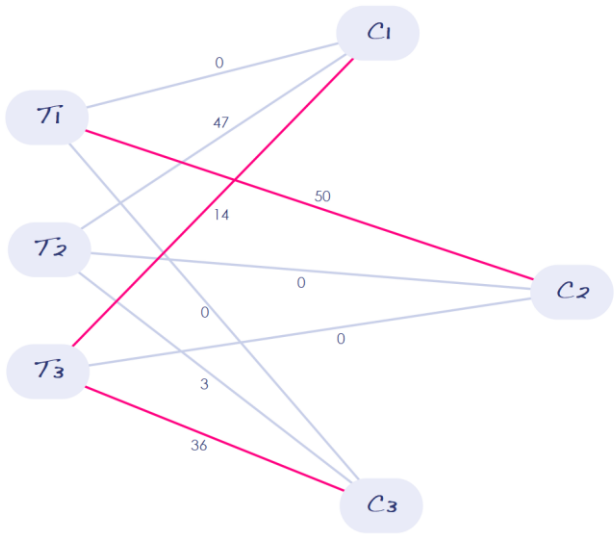
Matching M in G is a subset of E

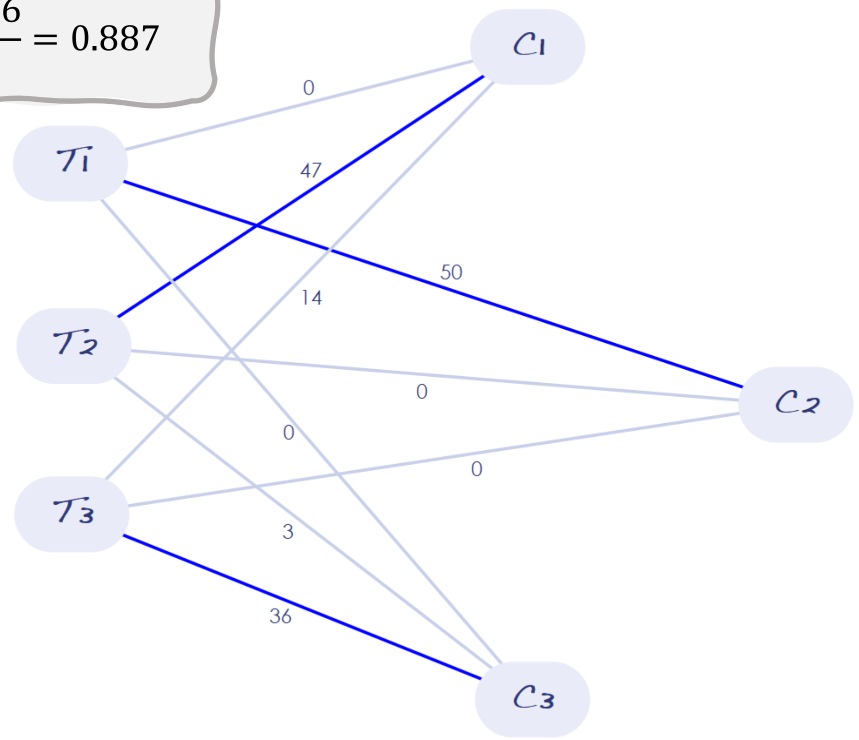
$$match = \arg\max_M \left\{ \frac{w(M)}{n} \right\}$$

	iris-setosa T_1	iris-versicolor T_2	iris-virginica T_3	n_i
C_1(squares)	0	47	14	61
C_2(circles)	50	0	0	50
C_3(triangles)	0	3	36	39
m_j	50	50	50	$n = 100$

For this contingency table, the maximum matching measure gives the same result, as the correspondence above is in fact a maximum weight matching. Thus, $match = 0.887$.

$$match = \frac{47 + 50 + 36}{150} = 0.887$$





F-Measure

$$j_i = \max_{j=1}^{k}\{n_{ij}\}$$

$$prec_i = \frac{1}{n_i}\max_{j=1}^{k}\left\{n_{ij}\right\} = \frac{n_{ij_i}}{n_i}$$

$$recall_i = \frac{n_{ij_i}}{|T_{j_i}|} = \frac{n_{ij_i}}{m_{j_i}}$$

$$F_i = \frac{2}{\frac{1}{prec_i} + \frac{1}{recall_i}} = \frac{2 \cdot prec_i \cdot recall_i}{prec_i + recall_i} = \frac{2\, n_{ij_i}}{n_i + m_{j_i}}$$

$$F = \frac{1}{r}\sum_{i=1}^{r} F_i$$

	iris-setosa T_1	iris-versicolor T_2	iris-virginica T_3	n_i
C_1(squares)	0	47	14	61
C_2(circles)	50	0	0	50
C_3(triangles)	0	3	36	39
m_j	50	50	50	$n = 100$

The cluster C_1 contains $n_1 = 47 + 14 = 61$, whereas its corresponding partition T_2 contains $m_2 = 47 + 3 = 50$ points. Thus, the precision and recall for C_1 are given as

$$prec_1 = \tfrac{47}{61} = 0.77$$
$$recall_1 = \tfrac{47}{50} = 0.94$$

The F-measure for C_1 is therefore

$$F_1 = \frac{2 \cdot 0.77 \cdot 0.94}{0.77 + 0.94} = \tfrac{1.45}{1.71} = 0.85$$

We can also directly compute F_1 using Eq. (17.1)

$$F_1 = \tfrac{2 \cdot n_{12}}{n_1 + m_2} = \frac{2 \cdot 47}{61 + 50} = \frac{94}{111} = 0.85$$

Likewise, we obtain $F_2 = 1.0$ and $F_3 = 0.81$. Thus, the F-measure value for the clustering is given as

$$F = \frac{1}{3}(F_1 + F_2 + F_3) = \frac{2.66}{3} = 0.88$$

17.1.1 Matching Based Measures

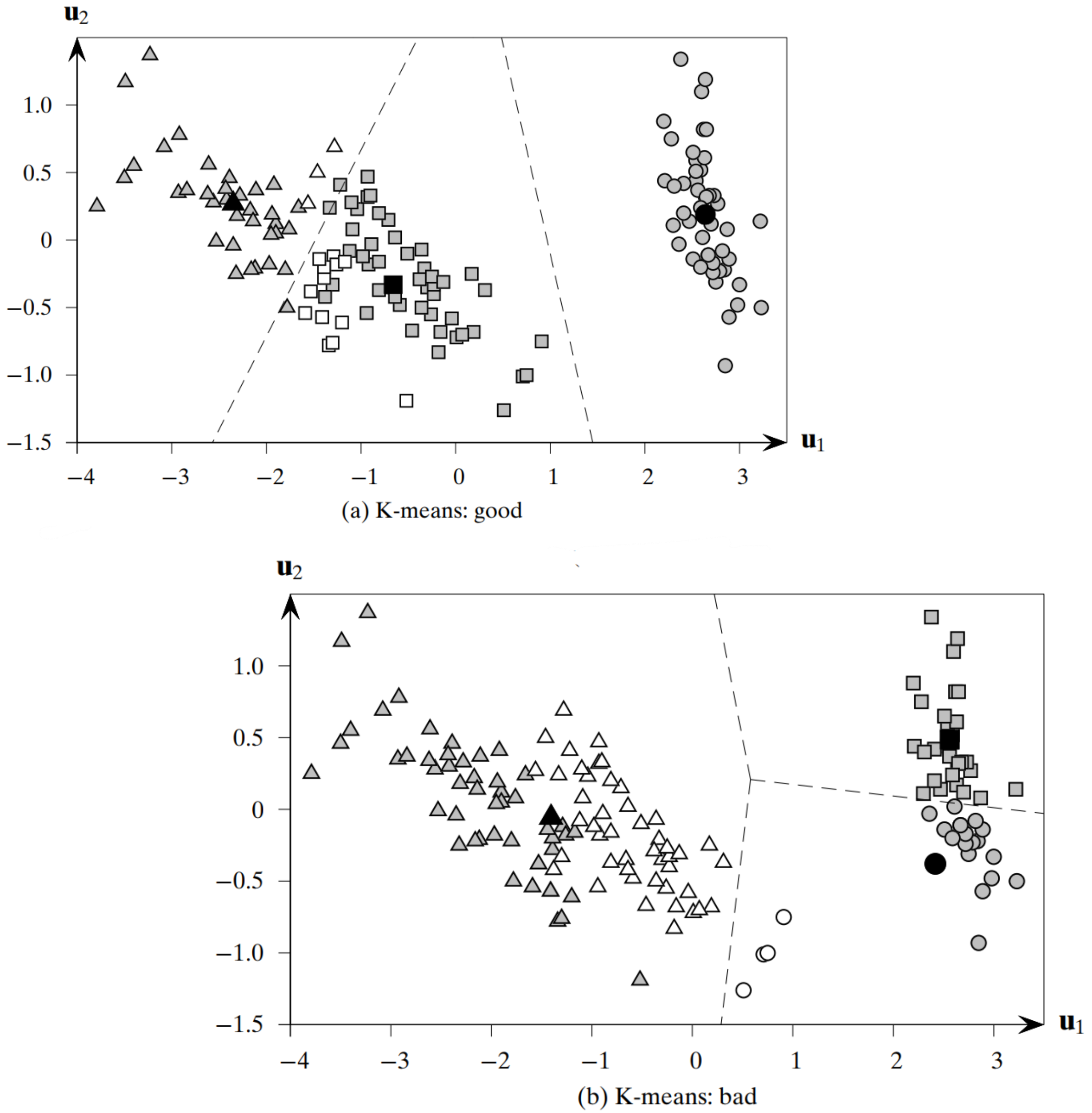


Figure 17.1. K-means: Iris principal components dataset.

For the clustering in Figure 17.1b, we have the following contingency table:

	iris-setosa T_1	iris-versicolor T_2	iris-virginica T_3	n_i
C_1	30	0	0	30
C_2	20	4	0	24
C_3	0	46	50	96
m_j	50	50	50	$n = 150$

For the purity measure, the partition with which each cluster shares the most points is given as (C_1, T_1), (C_2, T_1), and (C_3, T_3). Thus, the purity value for this clustering is

$$purity = \frac{1}{150}(30+20+50) = \frac{100}{150} = 0.67$$

We can see that both C_1 and C_2 choose partition T_1 as the maximum overlapping partition. However, the maximum weight matching is different; it yields the correspondence (C_1, T_1), (C_2, T_2), and (C_3, T_3), and thus

$$match = \frac{1}{150}(30+4+50) = \frac{84}{150} = 0.56$$

The table below compares the different contingency based measures for the two clusterings shown in Figure 17.1.

	purity	*match*	*F*
(a) Good	0.887	0.887	0.885
(b) Bad	0.667	0.560	0.658

The entropy of a clustering C

$$H(\mathcal{C}) = -\sum_{i=1}^{r} p_{C_i} \log p_{C_i} \qquad p_{C_i} = \frac{n_i}{n}$$

C1	C2	C3	Pc1=n1/n	Pc2=n2/n	Pc3=n3/n	H(C)=-Sum(p*log2(p))
61	50	39	0.406667	0.333333	0.26	1.5615
50	50	50	0.333333	0.333333	0.333333	1.58496
75	75	0	0.5	0.5	0	1
150	0	0	1	0	0	0

The entropy of the partitioning T

$$H(\mathcal{T}) = -\sum_{j=1}^{k} p_{T_j} \log p_{T_j} \qquad p_{T_j} = \frac{m_j}{n}$$

The conditional entropy of T with respect to cluster C_i

$$H(\mathcal{T}|C_i) = -\sum_{j=1}^{k} \left(\frac{n_{ij}}{n_i}\right) \log \left(\frac{n_{ij}}{n_i}\right)$$

Example 17.2. We continue with Example 17.1, which compares the two clusterings shown in Figure 17.1. For the entropy-based measures, we use base 2 for the logarithms; the formulas are valid for any base as such.

For the clustering in Figure 17.1a, we have the following contingency table:

	iris-setosa T_1	iris-versicolor T_2	iris-virginica T_3	n_i
C_1	0	47	14	61
C_2	50	0	0	50
C_3	0	3	36	39
m_j	50	50	50	$n = 100$

Consider the conditional entropy for cluster C_1:

$$H(\mathcal{T}|C_1) = -\frac{0}{61}\log_2\left(\frac{0}{61}\right) - \frac{47}{61}\log_2\left(\frac{47}{61}\right) - \frac{14}{61}\log_2\left(\frac{14}{61}\right)$$

$$= -0 - 0.77\log_2(0.77) - 0.23\log_2(0.23) = 0.29 + 0.49 = 0.78$$

	pi1 = ni1/ni	pi2 = ni2/ni	pi3 = ni3/ni
C1	0	0.770492	0.229508
C2	1	0	0
C3	0	0.0769231	0.923077

	-pi1*log2(pi1)	-pi2*log2(pi2)	-pi3*log2(pi3)	H(T|Ci)
C1	0	0.289819	0.487334	0.777153
C2	-0	0	0	0
C3	0	0.284649	0.106594	0.391244

The conditional entropy of T given clustering C

$$H(\mathcal{T}|\mathcal{C}) = \sum_{i=1}^{r} \frac{n_i}{n} H(\mathcal{T}|C_i) = -\sum_{i=1}^{r}\sum_{j=1}^{k} \frac{n_{ij}}{n} \log\left(\frac{n_{ij}}{n_i}\right)$$

$$= -\sum_{i=1}^{r}\sum_{j=1}^{k} p_{ij} \log\left(\frac{p_{ij}}{p_{C_i}}\right) \qquad p_{ij} = \frac{n_{ij}}{n}$$

$$H(\mathcal{T}|\mathcal{C}) = -\sum_{i=1}^{r}\sum_{j=1}^{k} p_{ij}\left(\log p_{ij} - \log p_{C_i}\right)$$

$$= -\left(\sum_{i=1}^{r}\sum_{j=1}^{k} p_{ij}\log p_{ij}\right) + \sum_{i=1}^{r}\left(\log p_{C_i} \sum_{j=1}^{k} p_{ij}\right)$$

$$= -\sum_{i=1}^{r}\sum_{j=1}^{k} p_{ij}\log p_{ij} + \sum_{i=1}^{r} p_{C_i}\log p_{C_i}$$

$$= H(\mathcal{C},\mathcal{T}) - H(\mathcal{C})$$

$H(\mathcal{C},\mathcal{T}) = -\sum_{i=1}^{r}\sum_{j=1}^{k} p_{ij}\log p_{ij}$ is the joint entropy of $\mathcal{C}$ and $\mathcal{T}$

	T1	T2	T3	ni
C1	0	47	14	61
C2	50	0	0	50
C3	0	3	36	39

	pi1 = ni1/ni	pi2 = ni2/ni	pi3 = ni3/ni
C1	0	0.770492	0.229508
C2	1	0	0
C3	0	0.0769231	0.923077

	-pi1*log2(pi1)	-pi2*log2(pi2)	-pi3*log2(pi3)	H(T|Ci)
C1	0	0.289819	0.487334	0.777153
C2	-0	0	0	0
C3	0	0.284649	0.106594	0.391244

H(T|C) = 0.42

In a similar manner, we obtain $H(\mathcal{T}|C_2) = 0$ and $H(\mathcal{T}|C_3) = 0.39$. The conditional entropy for the clustering $\mathcal{C}$ is then given as

$$H(\mathcal{T}|\mathcal{C}) = \frac{61}{150}\cdot 0.78 + \frac{50}{150}\cdot 0 + \frac{39}{150}\cdot 0.39 = 0.32 + 0 + 0.10 = 0.42$$

$$H(T|C) = 1.98 - 1.56 = 0.42$$

	T1	T2	T3	pi1=ni1/n	pi2=ni2/n	pi3=ni3/n	-pi1*log2(pi1)	-pi2*log2(pi2)	-pi3*log2(pi3)
C1	0	47	14	0	0.313333	0.0933333	0	0.524592	0.319337
C2	50	0	0	0.333333	0	0	0.528321	0	0
C3	0	3	36	0	0.02	0.24	0	0.112877	0.494134

H(C,T) = 1.98

Corresponding to the ideal clustering:
$H(T|C) = 0$ if and only if T is completely determined by C.

	T1	T2	T3
C1	0	50	0
C2	50	0	0
C3	0	0	50

H(C)=1.58
H(T,C)=1.58
H(T|C)=0.0

	T1	T2	T3
C1	50	0	0
C2	50	0	0
C3	0	0	50

H(C)=1.58
H(T,C)=1.58
H(T|C)=0.0

If C and T are independent of each other, then $H(T|C) = H(T)$, which means that C provides no information about T .

	T1	T2	T3
C1	0	50	0
C2	50	0	0
C3	0	0	50

H(C)=1.58
H(T,C)=1.58
H(T|C)=0.0
H(T)=1.58

	T1	T2	T3
C1	0	47	14
C2	50	0	0
C3	0	3	36

H(C)=1.56
H(T,C)=1.98
H(T|C)=0.42
H(T)=1.58

	T1	T2	T3
C1	15	15	15
C2	15	15	15
C3	15	15	15

H(C)=1.58
H(T,C)=3.17
H(T|C)=1.58
H(T)=1.58

Mutual Information

The amount of shared information between the clustering C and partitioning T

$$I(\mathcal{C},\mathcal{T}) = \sum_{i=1}^{r}\sum_{j=1}^{k} p_{ij} \log\left(\frac{p_{ij}}{p_{C_i} \cdot p_{T_j}}\right)$$

When C and T are independent then $p_{i,j} = p_{C_i} \cdot p_{T_j}$, and thus $I(C,T) = 0$.

$$I(\mathcal{C},\mathcal{T}) = H(\mathcal{C}) + H(\mathcal{T}) - H(\mathcal{C},\mathcal{T})$$

$$I(\mathcal{C},\mathcal{T}) = H(\mathcal{T}) - H(\mathcal{T}|\mathcal{C})$$

$$I(\mathcal{C},\mathcal{T}) = H(\mathcal{C}) - H(\mathcal{C}|\mathcal{T})$$

	T1	T2	T3
C1	0	50	0
C2	50	0	0
C3	0	0	50

H(C)=1.58
H(T,C)=1.58
H(T|C)=0.0
H(T)=1.58
I(T,C)=1.58

	T1	T2	T3
C1	0	47	14
C2	50	0	0
C3	0	3	36

H(C)=1.56
H(T,C)=1.98
H(T|C)=0.42
H(T)=1.58
I(T,C)=1.17

	T1	T2	T3
C1	15	15	15
C2	15	15	15
C3	15	15	15

H(C)=1.58
H(T,C)=3.17
H(T|C)=1.58
H(T)=1.58
I(T,C)=0.0

Normalized Mutual Information

$$NMI(\mathcal{C},\mathcal{T}) = \sqrt{\frac{I(\mathcal{C},\mathcal{T})}{H(\mathcal{C})} \cdot \frac{I(\mathcal{C},\mathcal{T})}{H(\mathcal{T})}} = \frac{I(\mathcal{C},\mathcal{T})}{\sqrt{H(\mathcal{C}) \cdot H(\mathcal{T})}}$$

The NMI value lies in the range $[0,1]$.
Values close to 1 indicate a good clustering.

Variation of Information

$$VI(\mathcal{C},\mathcal{T}) = (H(\mathcal{T}) - I(\mathcal{C},\mathcal{T})) + (H(\mathcal{C}) - I(\mathcal{C},\mathcal{T}))$$
$$= H(\mathcal{T}) + H(\mathcal{C}) - 2I(\mathcal{C},\mathcal{T})$$

The VI value is zero only when C and T are identical.
Thus, the lower the VI value the better the clustering C.

$$I(\mathcal{C},\mathcal{T}) = H(\mathcal{T}) - H(\mathcal{T}|\mathcal{C}) = H(\mathcal{C}) - H(\mathcal{C}|\mathcal{T})$$

$$VI(\mathcal{C},\mathcal{T}) = H(\mathcal{T}|\mathcal{C}) + H(\mathcal{C}|\mathcal{T})$$

$$VI(\mathcal{C},\mathcal{T}) = 2H(\mathcal{T},\mathcal{C}) - H(\mathcal{T}) - H(\mathcal{C})$$

	T1	T2	T3
C1	0	50	0
C2	50	0	0
C3	0	0	50

H(C)=1.58
H(T,C)=1.58
H(T|C)=0.0
H(T)=1.58
I(T,C)=1.58
NMI(T,C)=1.0
VI(T,C)=0.0

	T1	T2	T3
C1	0	47	14
C2	50	0	0
C3	0	3	36

H(C)=1.56
H(T,C)=1.98
H(T|C)=0.42
H(T)=1.58
I(T,C)=1.17
NMI(T,C)=0.74
VI(T,C)=0.81

	T1	T2	T3
C1	15	15	15
C2	15	15	15
C3	15	15	15

H(C)=1.58
H(T,C)=3.17
H(T|C)=1.58
H(T)=1.58
I(T,C)=0.0
NMI(T,C)=0.0
VI(T,C)=3.17

	T1	T2	T3
C1	30	0	0
C2	20	4	0
C3	0	46	50

H(C)=1.3
H(T,C)=2.04
H(T|C)=0.74
H(T)=1.58
I(T,C)=0.84
NMI(T,C)=0.59
VI(T,C)=1.2

	T1	T2	T3
C1	0	47	14
C2	50	0	0
C3	0	3	36

H(C)=1.56
H(T,C)=1.98
H(T|C)=0.42
H(T)=1.58
I(T,C)=1.17
NMI(T,C)=0.74
VI(T,C)=0.81

	T1	T2
C1	5	5
C2	5	5

H(C)=1.0
H(T,C)=2.0
H(T|C)=1.0
H(T)=1.0
I(T,C)=0.0
NMI(T,C)=0.0
VI(T,C)=2.0

To compute the normalized mutual information, note that

$$H(\mathcal{T}) = -3\left(\frac{50}{150}\log_2\left(\frac{50}{150}\right)\right) = 1.585$$

$$H(\mathcal{C}) = -\left(\frac{61}{150}\log_2\left(\frac{61}{150}\right) + \frac{50}{150}\log_2\left(\frac{50}{150}\right) + \frac{39}{150}\log_2\left(\frac{39}{150}\right)\right)$$

$$= 0.528 + 0.528 + 0.505 = 1.561$$

$$I(\mathcal{C},\mathcal{T}) = \frac{47}{150}\log_2\left(\frac{47\cdot 150}{61\cdot 50}\right) + \frac{14}{150}\log_2\left(\frac{14\cdot 150}{61\cdot 50}\right) + \frac{50}{150}\log_2\left(\frac{50\cdot 150}{50\cdot 50}\right)$$

$$+\frac{3}{150}\left(\log_2\frac{3\cdot 150}{39\cdot 50}\right) + \frac{36}{150}\log_2\left(\frac{36\cdot 150}{39\cdot 50}\right)$$

$$= 0.379 - 0.05 + 0.528 - 0.042 + 0.353 = 1.167$$

Thus, the NMI and VI values are

$$NMI(\mathcal{C},\mathcal{T}) = \frac{I(\mathcal{C},\mathcal{T})}{\sqrt{H(\mathcal{T})\cdot H(\mathcal{C})}} = \frac{1.167}{\sqrt{1.585\times 1.561}} = 0.742$$

$$VI(\mathcal{C},\mathcal{T}) = H(\mathcal{T}) + H(\mathcal{C}) - 2I(\mathcal{C},\mathcal{T}) = 1.585 + 1.561 - 2\cdot 1.167 = 0.812$$

We can likewise compute these measures for the other clustering in Figure 17.1b, whose contingency table is shown in Example 17.1.

The table below compares the entropy based measures for the two clusterings shown in Figure 17.1.

	$H(\mathcal{T}\mid\mathcal{C})$	NMI	VI
(a) Good	0.418	0.742	0.812
(b) Bad	0.743	0.587	1.200

As expected, the good clustering in Figure 17.1a has a higher score for normalized mutual information, and lower scores for conditional entropy and variation of information.

True Positives: $TP = |\{(\mathbf{x}_i, \mathbf{x}_j) : y_i = y_j \text{ and } \hat{y}_i = \hat{y}_j\}|$

False Negatives: $FN = |\{(\mathbf{x}_i, \mathbf{x}_j) : y_i = y_j \text{ and } \hat{y}_i \neq \hat{y}_j\}|$

False Positives: $FP = |\{(\mathbf{x}_i, \mathbf{x}_j) : y_i \neq y_j \text{ and } \hat{y}_i = \hat{y}_j\}|$

True Negatives: $TN = |\{(\mathbf{x}_i, \mathbf{x}_j) : y_i \neq y_j \text{ and } \hat{y}_i \neq \hat{y}_j\}|$

$$TP = \sum_{i=1}^{r}\sum_{j=1}^{k}\binom{n_{ij}}{2} = \sum_{i=1}^{r}\sum_{j=1}^{k}\frac{n_{ij}(n_{ij}-1)}{2} = \frac{1}{2}\left(\sum_{i=1}^{r}\sum_{j=1}^{k}n_{ij}^2 - \sum_{i=1}^{r}\sum_{j=1}^{k}n_{ij}\right)$$

$$= \frac{1}{2}\left(\left(\sum_{i=1}^{r}\sum_{j=1}^{k}n_{ij}^2\right) - n\right)$$

$$FN = \sum_{j=1}^{k}\binom{m_j}{2} - TP = \frac{1}{2}\left(\sum_{j=1}^{k}m_j^2 - \sum_{j=1}^{k}m_j - \sum_{i=1}^{r}\sum_{j=1}^{k}n_{ij}^2 + n\right)$$

$$= \frac{1}{2}\left(\sum_{j=1}^{k}m_j^2 - \sum_{i=1}^{r}\sum_{j=1}^{k}n_{ij}^2\right)$$

Example 17.3. Let us continue with Example 17.1. Consider again the contingency table for the clustering in Figure 17.1a:

	iris-setosa T_1	iris-versicolor T_2	iris-virginica T_3
C_1	0	47	14
C_2	50	0	0
C_3	0	3	36

Using Eq. (17.7), we can obtain the number of true positives as follows:

$$TP = \binom{47}{2} + \binom{14}{2} + \binom{50}{2} + \binom{3}{2} + \binom{36}{2}$$
$$= 1081 + 91 + 1225 + 3 + 630 = 3030$$

$FN = 645$ $\quad FP = 766$ $\quad TN = 6734$

Note that there are a total of $N = \binom{150}{2} = 11175$ point pairs.

$$FP = \sum_{i=1}^{r}\binom{n_i}{2} - TP = \frac{1}{2}\left(\sum_{i=1}^{r}n_i^2 - \sum_{i=1}^{r}\sum_{j=1}^{k}n_{ij}^2\right)$$

$$TN = N - (TP + FN + FP) = \frac{1}{2}\left(n^2 - \sum_{i=1}^{r}n_i^2 - \sum_{j=1}^{k}m_j^2 + \sum_{i=1}^{r}\sum_{j=1}^{k}n_{ij}^2\right)$$

17.1.3 Pairwise Measures

Jaccard Coefficient

$$Jaccard = \frac{TP}{TP + FN + FP}$$

Rand Statistic

$$Rand = \frac{TP + TN}{N}$$

Fowlkes-Mallows Measure

$$prec = \frac{TP}{TP + FP} \qquad recall = \frac{TP}{TP + FN}$$

$$FM = \sqrt{prec \cdot recall} = \frac{TP}{\sqrt{(TP + FN)(TP + FP)}}$$

$$Jaccard = \frac{3030}{3030 + 645 + 766} = \frac{3030}{4441} = 0.68$$

$$Rand = \frac{3030 + 6734}{11175} = \frac{9764}{11175} = 0.87$$

$$FM = \frac{3030}{\sqrt{3675 \cdot 3796}} = \frac{3030}{3735} = 0.81$$

	T1	T2	T3
C1	0	47	14
C2	50	0	0
C3	0	3	36

```
TP=3030
FN=645
FP=766
TN=6734
Jaccard=0.68
Rand=0.87
Prec=0.80
Recall=0.82
FM=0.81
```

	T1	T2	T3
C1	30	0	0
C2	20	4	0
C3	0	46	50

```
TP=2891
FN=784
FP=2380
TN=5120
Jaccard=0.48
Rand=0.72
Prec=0.55
Recall=0.79
FM=0.66
```

$TP = 2891 \qquad FN = 784 \qquad FP = 2380 \qquad TN = 5120$

The table below compares the different contingency based measures on the two clusterings in Figure 17.1.

	Jaccard	*Rand*	*FM*
(a) Good	0.682	0.873	0.811
(b) Bad	0.477	0.717	0.657

As expected, the clustering in Figure 17.1a has higher scores for all three measures.

Proximity Matrix

$$\mathbf{W} = \left\{\delta(\mathbf{x}_i, \mathbf{x}_j)\right\}_{i,j=1}^{n} \qquad \delta(\mathbf{x}_i, \mathbf{x}_j) = \|\mathbf{x}_i - \mathbf{x}_j\|_2$$

$$w_{ij} = \mathbf{W}(\mathbf{x}_i, \mathbf{x}_j) \qquad \mathcal{C} = \{C_1 \dots C_k\} \qquad V = \{\mathbf{x}_i \mid \mathbf{x}_i \in \mathbf{D}\}$$

$$S, R \subset V \qquad C_i \cap C_j = \emptyset \text{ for all } i, j, \text{ and } \bigcup_i C_i = V$$

$$W(S, R) = \sum_{\mathbf{x}_i \in S} \sum_{\mathbf{x}_j \in R} w_{ij}$$

given $S \subseteq V$, we denote by $\overline{S}$ the complementary set of vertices, that is, $\overline{S} = V - S$.

The sum of all the intracluster and intercluster weights

$$W_{in} = \frac{1}{2}\sum_{i=1}^{k} W(C_i, C_i)$$

$$W_{out} = \frac{1}{2}\sum_{i=1}^{k} W(C_i, \overline{C_i}) = \sum_{i=1}^{k-1}\sum_{j>i} W(C_i, C_j)$$

The number of distinct intracluster and intercluster edges

$$N_{in} = \sum_{i=1}^{k}\binom{n_i}{2} = \frac{1}{2}\sum_{i=1}^{k} n_i(n_i - 1)$$

$$N_{out} = \sum_{i=1}^{k-1}\sum_{j=i+1}^{k} n_i \cdot n_j = \frac{1}{2}\sum_{i=1}^{k}\sum_{\substack{j=1 \\ j \neq i}}^{k} n_i \cdot n_j$$

The total number of distinct pairs of points N

$$N = N_{in} + N_{out} = \binom{n}{2} = \frac{1}{2}n(n-1)$$

Example 17.6. Consider the two clusterings for the Iris principal components dataset shown in Figure 17.1, along with their corresponding graph representations in Figure 17.2. Let us evaluate these two clusterings using internal measures.

The good clustering shown in Figure 17.1a and Figure 17.2a has clusters with the following sizes:

$$n_1 = 61 \qquad n_2 = 50 \qquad n_3 = 39$$

Thus, the number of intracluster and intercluster edges (i.e., point pairs) is given as

$$N_{in} = \binom{61}{2} + \binom{50}{2} + \binom{31}{2} = 1830 + 1225 + 741 = 3796$$

$$N_{out} = 61 \cdot 50 + 61 \cdot 39 + 50 \cdot 39 = 3050 + 2379 + 1950 = 7379$$

In total there are $N = N_{in} + N_{out} = 3796 + 7379 = 11175$ distinct point pairs.

The weights on edges within each cluster $W(C_i, C_i)$, and those from a cluster to another $W(C_i, C_j)$, are as given in the intercluster weight matrix

$$\begin{pmatrix} W & C_1 & C_2 & C_3 \\ \hline C_1 & 3265.69 & 10402.30 & 4418.62 \\ C_2 & 10402.30 & 1523.10 & 9792.45 \\ C_3 & 4418.62 & 9792.45 & 1252.36 \end{pmatrix} \qquad (17.29)$$

Thus, the sum of all the intracluster and intercluster edge weights is

$$W_{in} = \frac{1}{2}(3265.69 + 1523.10 + 1252.36) = 3020.57$$

$$W_{out} = (10402.30 + 4418.62 + 9792.45) = 24613.37$$

BetaCV Measure

$$BetaCV = \frac{W_{in}/N_{in}}{W_{out}/N_{out}} = \frac{N_{out}}{N_{in}} \cdot \frac{W_{in}}{W_{out}} = \frac{N_{out}}{N_{in}} \frac{\sum_{i=1}^{k} W(C_i, C_i)}{\sum_{i=1}^{k} W(C_i, \overline{C_i})}$$

C-index

$$Cindex = \frac{W_{in} - W_{\min}(N_{in})}{W_{\max}(N_{in}) - W_{\min}(N_{in})}$$

Dunn Index

$$Dunn = \frac{W_{out}^{\min}}{W_{in}^{\max}}$$

$$W_{out}^{\min} = \min_{i,j>i} \{w_{ab} | \mathbf{x}_a \in C_i, \mathbf{x}_b \in C_j\}$$

$$W_{in}^{\max} = \max_{i} \{w_{ab} | \mathbf{x}_a, \mathbf{x}_b \in C_i\}$$

The BetaCV measure can then be computed as

$$BetaCV = \frac{N_{out} \cdot W_{in}}{N_{in} \cdot W_{out}} = \frac{7379 \times 3020.57}{3796 \times 24613.37} = 0.239$$

For the C-index, we first compute the sum of the N_{in} smallest and largest pair-wise distances, given as

$$W_{\min}(N_{in}) = 2535.96 \qquad W_{\max}(N_{in}) = 16889.57$$

Thus, C-index is given as

$$Cindex = \frac{W_{in} - W_{\min}(N_{in})}{W_{\max}(N_{in}) - W_{\min}(N_{in})} = \frac{3020.57 - 2535.96}{16889.57 - 2535.96} = \frac{484.61}{14535.61} = 0.0338$$

The Dunn index can be computed from the minimum and maximum intercluster distances:

$$\left(\begin{array}{c|ccc} W^{\min} & C_1 & C_2 & C_3 \\ \hline C_1 & 0 & 1.62 & 0.198 \\ C_2 & 1.62 & 0 & 3.49 \\ C_3 & 0.198 & 3.49 & 0 \end{array}\right) \qquad \left(\begin{array}{c|ccc} W^{\max} & C_1 & C_2 & C_3 \\ \hline C_1 & 2.50 & 4.85 & 4.81 \\ C_2 & 4.85 & 2.33 & 7.06 \\ C_3 & 4.81 & 7.06 & 2.55 \end{array}\right)$$

The Dunn index value for the clustering is given as

$$Dunn = \frac{W_{out}^{\min}}{W_{in}^{\max}} = \frac{0.198}{2.55} = 0.078$$

Davies–Bouldin Index

$$\mu_i = \frac{1}{n_i}\sum_{\mathbf{x}_j \in C_i} \mathbf{x}_j \qquad \sigma_{\mu_i} = \sqrt{\frac{\sum_{\mathbf{x}_j \in C_i} \delta(\mathbf{x}_j, \mu_i)^2}{n_i}} = \sqrt{var(C_i)}$$

$$DB_{ij} = \frac{\sigma_{\mu_i} + \sigma_{\mu_j}}{\delta(\mu_i, \mu_j)} \qquad DB = \frac{1}{k}\sum_{i=1}^{k} \max_{j \neq i}\{DB_{ij}\}$$

Silhouette Coefficient

$$s_i = \frac{\mu_{out}^{\min}(\mathbf{x}_i) - \mu_{in}(\mathbf{x}_i)}{\max\left\{\mu_{out}^{\min}(\mathbf{x}_i), \mu_{in}(\mathbf{x}_i)\right\}}$$

$$\mu_{in}(\mathbf{x}_i) = \frac{\sum_{\mathbf{x}_j \in C_{\hat{y}_i}, j \neq i} \delta(\mathbf{x}_i, \mathbf{x}_j)}{n_{\hat{y}_i} - 1}$$

$$\mu_{out}^{\min}(\mathbf{x}_i) = \min_{j \neq \hat{y}_i}\left\{\frac{\sum_{\mathbf{y} \in C_j} \delta(\mathbf{x}_i, \mathbf{y})}{n_j}\right\} \qquad SC = \frac{1}{n}\sum_{i=1}^{n} s_i$$

To compute the Davies–Bouldin index, we compute the cluster mean and dispersion values:

$$\mu_1 = \begin{pmatrix} -0.664 \\ -0.33 \end{pmatrix} \qquad \mu_2 = \begin{pmatrix} 2.64 \\ 0.19 \end{pmatrix} \qquad \mu_3 = \begin{pmatrix} -2.35 \\ 0.27 \end{pmatrix}$$

$$\sigma_{\mu_1} = 0.723 \qquad \sigma_{\mu_2} = 0.512 \qquad \sigma_{\mu_3} = 0.695$$

and the DB_{ij} values for pairs of clusters:

$$\left(\begin{array}{c|ccc} DB_{ij} & C_1 & C_2 & C_3 \\ \hline C_1 & - & 0.369 & 0.794 \\ C_2 & 0.369 & - & 0.242 \\ C_3 & 0.794 & 0.242 & - \end{array}\right)$$

For example, $DB_{12} = \frac{\sigma_{\mu_1} + \sigma_{\mu_2}}{\delta(\mu_1, \mu_2)} = \frac{1.235}{3.346} = 0.369$. Finally, the DB index is given as

$$DB = \frac{1}{3}(0.794 + 0.369 + 0.794) = 0.652$$

The silhouette coefficient [Eq. (17.26)] for a chosen point, say $\mathbf{x}_1$, is given as

$$s_i = \frac{1.902 - 0.701}{\max\{1.902, 0.701\}} = \frac{1.201}{1.902} = 0.632$$

The average value across all points is $SC = 0.598$

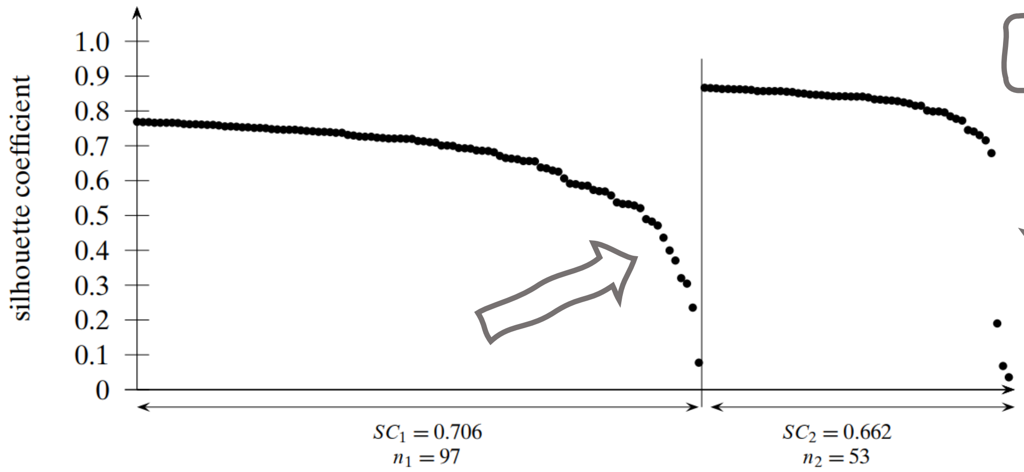
	Lower better				Higher better				
	$BetaCV$	$Cindex$	Q	DB	NC	$Dunn$	SC	Γ	Γ_n
(a) Good	0.24	0.034	−0.23	0.65	2.67	0.08	0.60	8.19	0.92
(b) Bad	0.33	0.08	−0.20	1.11	2.56	0.03	0.55	7.32	0.83

to compare different clusterings obtained by varying different parameters for the same algorithm, for example, to choose the number of clusters k.

Silhouette Coefficient

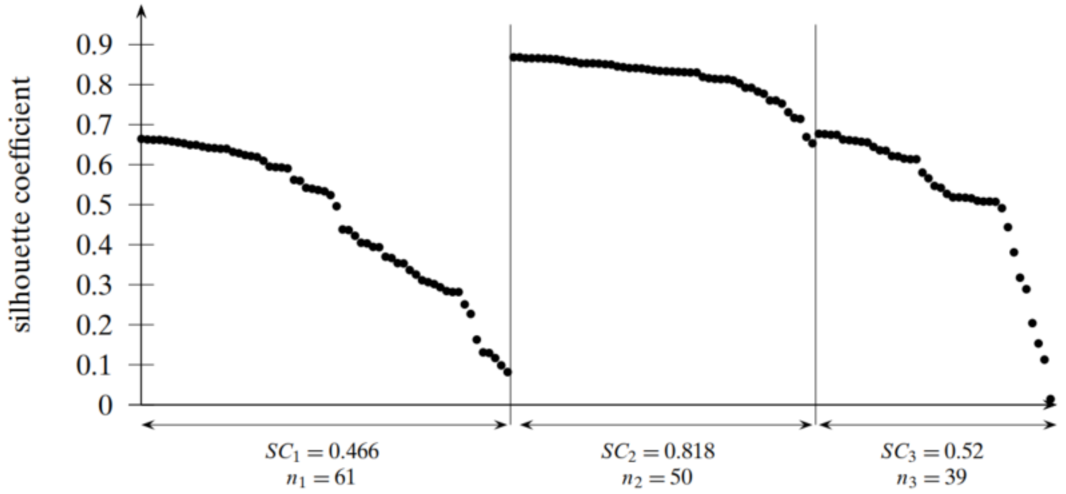
$$\mu_{in}(\mathbf{x}_i) = \frac{\sum_{\mathbf{x}_j \in C_{\hat{y}_i}, j \neq i} \delta(\mathbf{x}_i, \mathbf{x}_j)}{n_{\hat{y}_i} - 1} \qquad \mu_{out}^{\min}(\mathbf{x}_i) = \min_{j \neq \hat{y}_i} \left\{ \frac{\sum_{\mathbf{y} \in C_j} \delta(\mathbf{x}_i, \mathbf{y})}{n_j} \right\}$$

$$s_i = \frac{\mu_{out}^{\min}(\mathbf{x}_i) - \mu_{in}(\mathbf{x}_i)}{\max\left\{\mu_{out}^{\min}(\mathbf{x}_i), \mu_{in}(\mathbf{x}_i)\right\}} \qquad SC_i = \frac{1}{n_i} \sum_{\mathbf{x}_j \in C_i} s_j \qquad SC = \frac{1}{n} \sum_{i=1}^{n} s_i$$

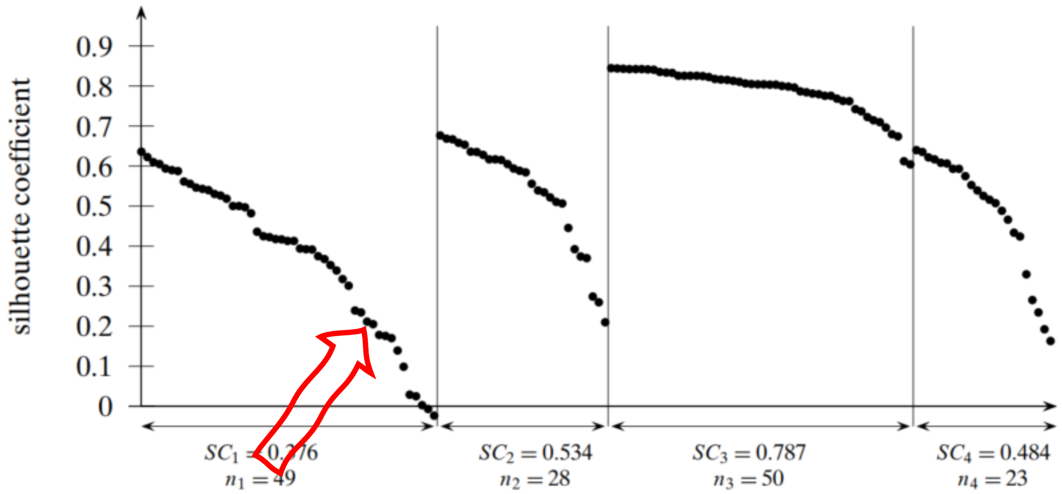


(a) $k = 2$, $SC = 0.706$

Border Points



(b) $k = 3$, $SC = 0.598$



(c) $k = 4$, $SC = 0.559$

Because k = 2 yields the highest silhouette coefficient, and the two clusters are essentially well separated, in the absence of prior knowledge, we would choose k = 2 as the best number of clusters for this dataset.

17.3.1 Cluster Stability

The main idea behind cluster stability is that the clusterings obtained from several datasets sampled from the same underlying distribution as D should be similar or "stable."

The joint probability distribution for D is typically unknown. Therefore, to sample a dataset from the same distribution we can try a variety of methods, including random perturbations, subsampling, or bootstrap resampling.

Considering the **bootstrapping** approach; we generate t samples of size n by sampling from D with replacement, which allows the same point to be chosen possibly multiple times, and thus each sample D_i will be different. Next, for each sample D_i we run the same clustering algorithm with different cluster values k ranging from 2 to k_{max}.

Several of the external cluster evaluation measures can be used as distance measures, by setting, for example, $C = C_k(D_i)$ and $T = C_k(D_j)$, or vice versa.

The points common to both D_i and D_j, denoted as D_{ij}.
For each point $\boldsymbol{x}_a$ in the input dataset $\boldsymbol{D}$, let m_i^a and m_j^a denote the number of occurrences of $\boldsymbol{x}_a$in D_i and D_j, respectively.

$$\mathbf{D}_{ij} = \mathbf{D}_i \cap \mathbf{D}_j = \left\{ m^a \text{ instances of } \mathbf{x}_a \mid \mathbf{x}_a \in \mathbf{D}, m^a = \min\{m_i^a, m_j^a\} \right\}$$

ALGORITHM 17.1. Clustering Stability Algorithm for Choosing k

ClusteringStability $(A, t, k^{\max}, \mathbf{D})$:

$n \leftarrow |\mathbf{D}|$
`// Generate t samples`
for $i = 1, 2, \ldots, t$ **do**
  $\mathbf{D}_i \leftarrow$ sample n points from $\mathbf{D}$ with replacement
`// Generate clusterings for different values of k`
for $i = 1, 2, \ldots, t$ **do**
  for $k = 2, 3, \ldots, k^{\max}$ **do**
    $\mathcal{C}_k(\mathbf{D}_i) \leftarrow$ cluster $\mathbf{D}_i$ into k clusters using algorithm A
`// Compute mean difference between clusterings for each k`
foreach *pair* $\mathbf{D}_i, \mathbf{D}_j$ *with* $j > i$ **do**
  $\mathbf{D}_{ij} \leftarrow \mathbf{D}_i \cap \mathbf{D}_j$ `// create common dataset using` (17.30)
  for $k = 2, 3, \ldots, k^{\max}$ **do**
    $d_{ij}(k) \leftarrow d\big(\mathcal{C}_k(\mathbf{D}_i), \mathcal{C}_k(\mathbf{D}_j), \mathbf{D}_{ij}\big)$ `// distance between clusterings`
for $k = 2, 3, \ldots, k^{\max}$ **do**
  $\mu_d(k) \leftarrow \frac{2}{t(t-1)} \sum_{i=1}^{t} \sum_{j>i} d_{ij}(k)$ `// expected pairwise distance`
`// Choose best k`
$k^* \leftarrow \arg\min_k \{\mu_d(k)\}$

Instead of the distance function d, we can also evaluate clustering stability via a similarity measure, in which case, after computing the average similarity between pairs of clusterings for a given k, we can choose the best value k^* as the one that maximizes the expected similarity $\mu_s(k)$.
Examples of similarity functions include Jaccard, Fowlkes–Mallows and so on.

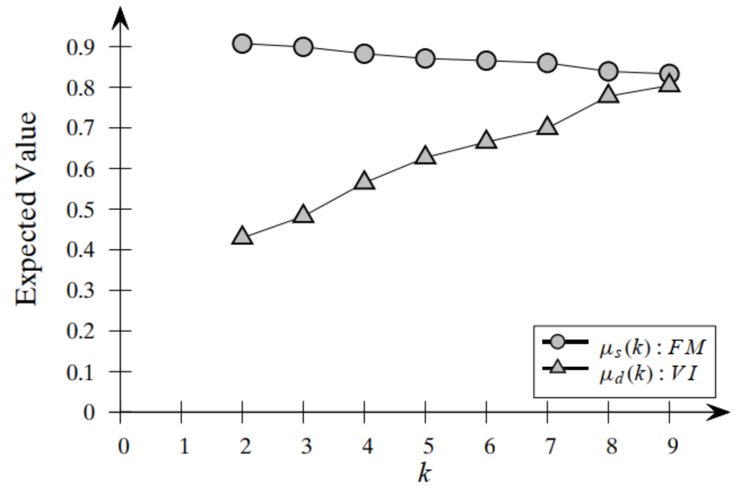
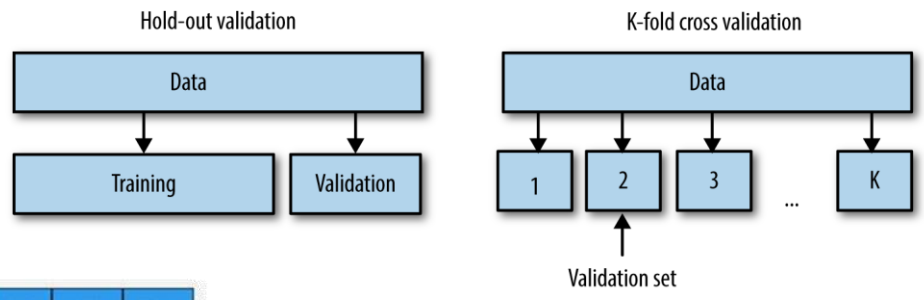
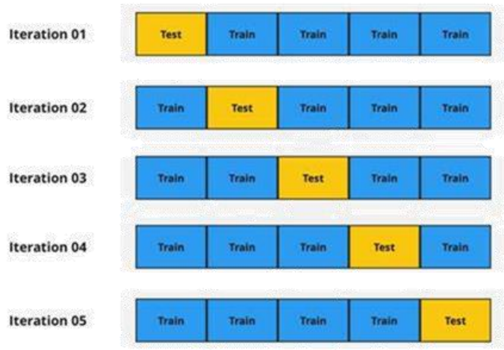
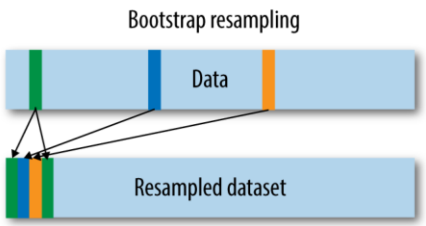


Figure 17.6. Clustering stability: Iris dataset.

In summary, cross-validation splits the available dataset to create multiple datasets, and bootstrapping method uses the original dataset to create multiple datasets after resampling with replacement. However, bootstrapping is not as strong as cross-validation when it is used for model validation.







Clustering tendency or **clusterability** aims to determine whether the dataset D has any meaningful groups to begin with.

Spatial Histogram

we divide each dimension X j into b equi-width bins, and simply count how many points lie in each of the bd d-dimensional cells.

The empirical joint probability mass function (EPMF)

$$f(\mathbf{i}) = P(\mathbf{x}_j \in \text{cell } \mathbf{i}) = \frac{|\{\mathbf{x}_j \in \text{cell } \mathbf{i}\}|}{n}$$

$$\mathbf{i} = (i_1, i_2, \ldots, i_d)$$

Next, we generate t random samples, each comprising n points within the same d-dimensional space as the input dataset $\mathbf{D}$. That is, for each dimension X_j, we compute its range $[\min(X_j), \max(X_j)]$, and generate values uniformly at random within the given range. Let $\mathbf{R}_j$ denote the jth such random sample. We can then compute the corresponding EPMF $g_j(\mathbf{i})$ for each $\mathbf{R}_j$, $1 \leq j \leq t$.

Finally, we can compute how much the distribution f differs from g_j (for $j = 1, \ldots, t$), using the Kullback–Leibler (KL) divergence from f to g_j, defined as

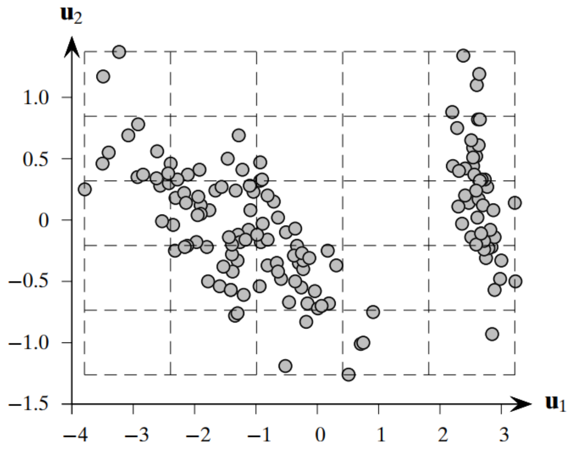
$$KL(f|g_j) = \sum_{\mathbf{i}} f(\mathbf{i}) \log\left(\frac{f(\mathbf{i})}{g_j(\mathbf{i})}\right)$$

The KL divergence is zero only when f and g_j are the same distributions. Using these divergence values, we can compute how much the dataset $\mathbf{D}$ differs from a random dataset.

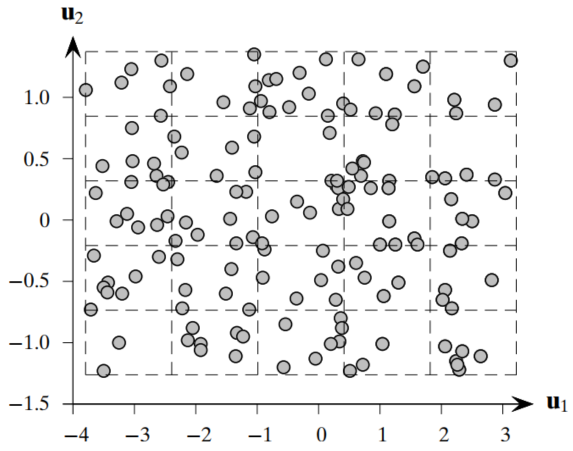
17.3.2 Clustering Tendency

Example 17.11. Figure 17.7c shows the empirical joint probability mass function for the Iris principal components dataset that has $n = 150$ points in $d = 2$ dimensions. It also shows the EPMF for one of the datasets generated uniformly at random in the same data space. Both EPMFs were computed using $b = 5$ bins in each dimension, for a total of 25 spatial cells. The spatial grids/cells for the Iris dataset $\mathbf{D}$, and the random sample $\mathbf{R}$, are shown in Figures 17.7a and 17.7b, respectively. The cells are numbered starting from 0, from bottom to top, and then left to right. Thus, the bottom left cell is 0, top left is 4, bottom right is 19, and top right is 24. These indices are used along the x-axis in the EPMF plot in Figure 17.7c.

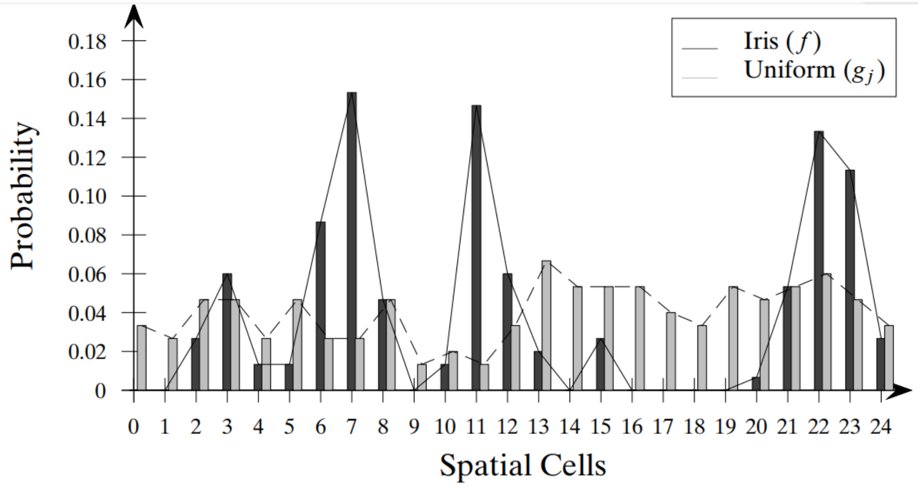
We generated $t = 500$ random samples from the null distribution, and computed the KL divergence from f to g_j for each $1 \le j \le t$ (using logarithm with base 2). The distribution of the KL values is plotted in Figure 17.7d. The mean KL value was $\mu_{KL} = 1.17$, with a standard deviation of $\sigma_{KL} = 0.18$, indicating that the Iris data is indeed far from the randomly generated data, and thus is clusterable.



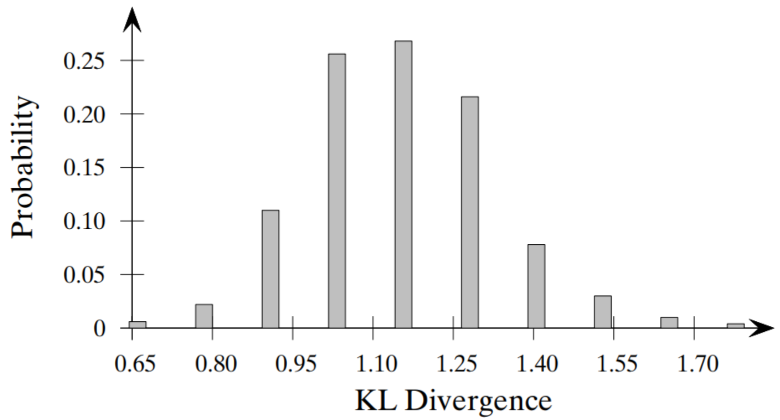
(a) Iris: spatial cells



(b) Uniform: spatial cells



(c) Empirical probability mass function



(d) KL-divergence distribution

Distance Distribution

Instead of trying to estimate the density, another approach to determine clusterability is to compare the pairwise point distances from $\mathbf{D}$, with those from the randomly generated samples $\mathbf{R}_i$ from the null distribution. That is, we create the EPMF from the proximity matrix $\mathbf{W}$ for $\mathbf{D}$ [Eq. (17.22)] by binning the distances into b bins:

$$f(i) = P(w_{pq} \in \text{ bin } i \mid \mathbf{x}_p, \mathbf{x}_q \in \mathbf{D}, p > q) = \frac{|\{w_{pq} \in \text{ bin } i\}|}{n(n-1)/2}$$

Likewise, for each of the samples $\mathbf{R}_j$, we can determine the EPMF for the pairwise distances, denoted g_j. Finally, we can compute the KL divergences between f and g_j using Eq. (17.31). The expected divergence indicates the extent to which $\mathbf{D}$ differs from the null (random) distribution.

Example 17.12. Figure 17.8a shows the distance distribution for the Iris principal components dataset $\mathbf{D}$ and the random sample $\mathbf{R}_j$ from Figure 17.7b. The distance distribution is obtained by binning the edge weights between all pairs of points using $b = 25$ bins.

We then compute the KL divergence from $\mathbf{D}$ to each $\mathbf{R}_j$, over $t = 500$ samples. The distribution of the KL divergences (using logarithm with base 2) is shown in Figure 17.8b. The mean divergence is $\mu_{KL} = 0.18$, with standard deviation $\sigma_{KL} = 0.017$. Even though the Iris dataset has a good clustering tendency, the KL divergence is not very large. We conclude that, at least for the Iris dataset, the distance distribution is not as discriminative as the spatial histogram approach for clusterability analysis.