Unsupervised Classification

Introduction to Clustering

Data Analysis – juliette.chevallier@insa-toulouse.fr

INSA Toulouse, Applied Mathematics, 4th year

What is clustering? Why is it used?

Principle and First Examples

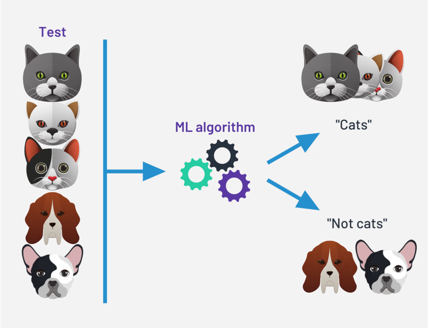
Supervised *vs.* Unsupervised Classification

Cluster Analysis

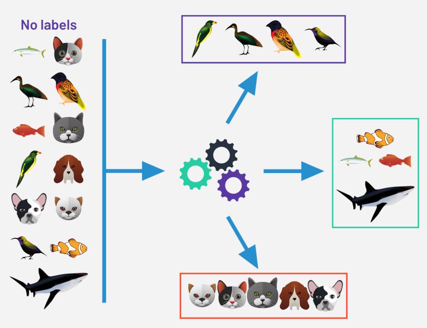
From Wikipedia, the free encyclopedia

Cluster analysis or **clustering** is the task of grouping a set of objects in such a way that objects in the same group (called a **cluster**) are more similar (in some sense) to each other than to those in other groups (clusters).

Images from www.g2.com/articles/supervised-vs-unsupervised-learning



Supervised **classification**

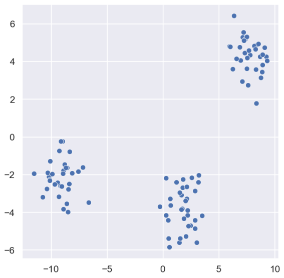


Unsupervised **classification**

Toy Dataset – *Python notebook available on Moodle*

```
n_points = 100
data, labels = make_blobs(n_samples=n_points, n_features=2, centers=3,
                          cluster_std=1, center_box=[-10,10])
```

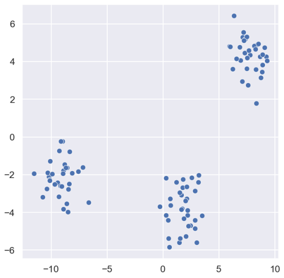
```
sns.scatterplot(x=data[:, 0],
                y=data[:, 1])
plt.show()
```



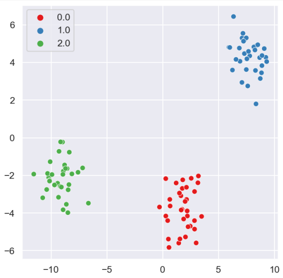
Toy Dataset – *Python notebook available on Moodle*

```python
n_points = 100
data, labels = make_blobs(n_samples=n_points, n_features=2, centers=3,
                          cluster_std=1, center_box=[-10,10])
```

```python
sns.scatterplot(x=data[:, 0],
                y=data[:, 1])
plt.show()
```

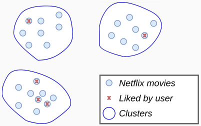



```python
sns.scatterplot(x=data[:, 0],
                y=data[:, 1],
                hue=labels)
plt.show()
```

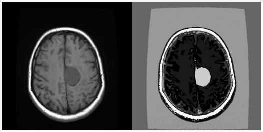


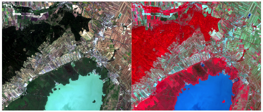
Some Applications in Real Life

- Recommendation systems



- Image segmentation: *Tumor identification, Ecological studies, etc.*



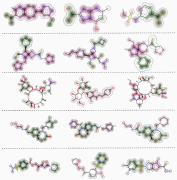


See www.kdnuggets.com/2019/08/introduction-image-segmentation-k-means-clustering.html

- Unsupervised robotic sorting: *Garbage-sorting bot, etc.*

See The Everyday Robot Project from Alphabet

- Data-driven discovery of new chemicals



- Unsupervised image/signal classification

Principle of Clustering

We observe n individuals described by p variables: $x_i = \left(x_{i1}, x_{i2}, \dots, x_{ip}\right) \in \mathcal{X}$

$$X = \begin{pmatrix} x_{11} & \boxed{x_{12}} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}$$

$\mathcal{X} = \mathbb{R}^p\,,\ \{0,1\}^p\,,\ ]-\pi,\pi]^p\,,\ \mathbb{R}^q \times \{0,1\}^{p-q}\,, \dots$

- Initial measurements
- Transformed measurements
- Coordinates after dimension reduction

Principle of Clustering

We observe n individuals described by p variables: $x_i = \left(x_{i1}, x_{i2}, \ldots, x_{ip}\right) \in \mathcal{X}$

$$X = \begin{pmatrix} x_{11} & \boxed{x_{12}} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix}$$

$\mathcal{X} = \mathbb{R}^p\,,\ \{0,1\}^p\,,\ ]-\pi,\pi]^p\,,\ \mathbb{R}^q \times \{0,1\}^{p-q}\,, \ldots$

- Initial measurements
- Transformed measurements
- Coordinates after dimension reduction

Classification: Partitioning a collection of *heterogeneous* individuals into a set of *homogeneous* classes.

Unsupervised: No *a priori* partition of the n individuals, Number of classes K unknown.

Set of data points on which we do not know the labels, but that we want to group together in a smart way.

Principle of Clustering

We observe n individuals described by p variables: $x_i = \left(x_{i1}, x_{i2}, \ldots, x_{ip}\right) \in \mathcal{X}$

$$X = \begin{pmatrix} x_{11} & \boxed{x_{12}} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix}$$

$$\mathcal{X} = \mathbb{R}^p\,,\ \{0,1\}^p\,,\ ]-\pi,\pi]^p\,,\ \mathbb{R}^q \times \{0,1\}^{p-q}\,,\ldots$$

- Initial measurements
- Transformed measurements
- Coordinates after dimension reduction

Classification: Partitioning a collection of *heterogeneous* individuals into a set of *homogeneous* classes.

Unsupervised: No *a priori* partition of the n individuals, Number of classes K unknown.

Set of data points on which we do not know the labels, but that we want to group together in a smart way.

$\Longrightarrow$ Determine K classes $\mathcal{P}_K = \{\mathcal{C}_1, \ldots, \mathcal{C}_K\}$ of the n individuals *from X* such that a **class** is a collection of individuals:

- similar to each other, and
- dissimilar to the individuals of the other classes (well separated classes).

Inertia (for Quantitative Data)

- Assume *quantitative* variables and d_q the Minkowski distance, *i.e.* the distance associated to the norm $\|\cdot\|_q$.
- Let a partition $\mathcal{P}_K = \{\mathcal{C}_1, \ldots, \mathcal{C}_K\}$ into K classes.

Total inertia *Total variance*

$$\boxed{I_{Tot} = \sum_{i=1}^{n} d(\mu, x_i)^q}$$

Let $\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$, center of gravity of the *point cloud*.

Interclass inertia *Variance of class centers*

$$\boxed{I_{Inter} = \sum_{k=1}^{K} |\mathcal{C}_k|\, d(\mu, \mu_k)^q}$$

Let $\mu_k = \frac{1}{|\mathcal{C}_k|}\sum_{i \in \mathcal{C}_k} x_i$, center of gravity of the *class* $\mathcal{C}_k$.

Intraclass inertia *Variance of points in the same class*

$$\boxed{I_{Intra} = \sum_{k=1}^{K} \sum_{i \in \mathcal{C}_k} d(\mu_k, x_i)^q}$$

Inertia (for Quantitative Data)

- Assume *quantitative* variables and d_2 the Euclidean distance, *i.e.* the distance associated to the norm $\|\cdot\|_2$.
- Let a partition $\mathcal{P}_K = \{\mathcal{C}_1, \ldots, \mathcal{C}_K\}$ into K classes.

Total inertia *Total variance*

$$\boxed{I_{Tot} = \sum_{i=1}^{n} d(\mu, x_i)^2}$$

Let $\mu = \dfrac{1}{n}\sum_{i=1}^{n} x_i$, center of gravity of the *point cloud*.

Interclass inertia *Variance of class centers*

$$\boxed{I_{Inter} = \sum_{k=1}^{K} |\mathcal{C}_k|\, d(\mu, \mu_k)^2}$$

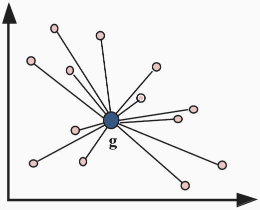
Let $\mu_k = \dfrac{1}{|\mathcal{C}_k|}\sum_{i\in\mathcal{C}_k} x_i$, center of gravity of the *class* $\mathcal{C}_k$.

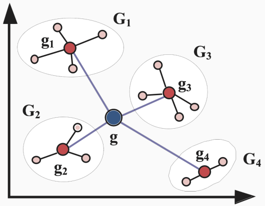
Intraclass inertia *Variance of points in the same class*

$$\boxed{I_{Intra} = \sum_{k=1}^{K}\sum_{i\in\mathcal{C}_k} d(\mu_k, x_i)^2}$$

Huygens' Principle

$$\boxed{I_{Tot} = I_{Inter} + I_{Intra}}$$

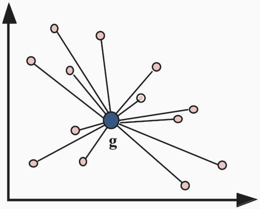


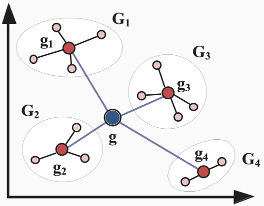


Demonstration: Pythagorean theorem

Huygens' Principle

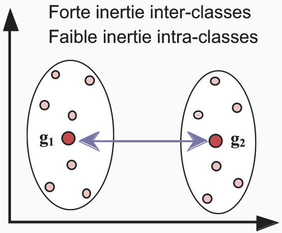
$$\boxed{I_{Tot} = I_{Inter} + I_{Intra}}$$





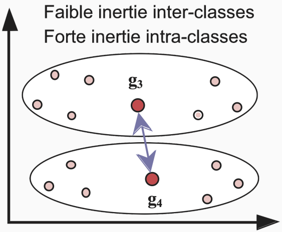
Demonstration: Pythagorean theorem

Principle of clustering: Minimize intraclass inertia
$\Longleftrightarrow$ Maximize interclass inertia



Remark:

$I_{Inter} = BCSS$
$I_{Intra} = WCSS$



Images: Bisson, 2001

Impossibility of an Exhaustive Search

Disclaimer: Here, we only deal with "hard classification" methods: an individual belongs to only one class, *i.e.*

$$\forall i \in [\![1, n]\!], \quad \exists! k \in [\![1, K]\!] \quad \text{such that} \quad i \in \mathcal{C}_k\,.$$

Stirling numbers of the second kind: Number of ways to partition a set of n elements into K nonempty subsets

$$S(n, K) = \left\{ \begin{matrix} n \\ K \end{matrix} \right\} = \frac{1}{K!} \sum_{j=0}^{K} (-1)^{K-j} j^n \binom{K}{j}\,.$$

$\rightarrow$ $S(100, 3) \simeq 10^{47}$ partitions of $n = 100$ individuals into $K = 3$ classes,

$\rightarrow$ $S(100, 5) \simeq 10^{68}$ partitions of $n = 100$ individuals into $K = 5$ classes.

$\rightsquigarrow$ **Impossibility of an Exhaustive Search.**

Quantify the Dissimilarity

- Clustering methods requires to be able to quantify the dissimilarity between observations.
  - $\rightsquigarrow$ Appropriate **dissimilarities** and **distances**

Quantify the Dissimilarity

- Clustering methods requires to be able to quantify the dissimilarity between observations.

  $\rightsquigarrow$ Appropriate **dissimilarities** and **distances**

- **Quantitative data**: Minkowski distance, Euclidean distance, Mahalanobis, *etc.*

- **Qualitative data**: Rogers and Tanimoto dissimilarity, simple dissimilarity, *etc.*

  Example: Let x, y categorical with p features. $d(x, y) = \sum_{j=1}^{p} \mathbb{1}_{\{x_j \neq y_j\}}$

- **Mixed data**: Gower metric, *etc.*

Quantify the Dissimilarity

- Clustering methods requires to be able to quantify the dissimilarity between observations.

  $\rightsquigarrow$ Appropriate **dissimilarities** and **distances**

- **Quantitative data**: Minkowski distance, Euclidean distance, Mahalanobis, *etc.*

- **Qualitative data**: Rogers and Tanimoto dissimilarity, simple dissimilarity, *etc.*

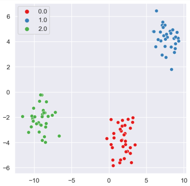
  Example: Let x, y categorical with p features. $d(x,y) = \sum_{j=1}^{p} \mathbb{1}_{\{x_j \neq y_j\}}$

- **Mixed data**: Gower metric, *etc.*

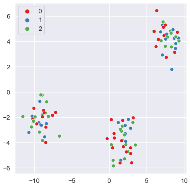
- **Dimension curse**: Beware of the behavior of distances in large dimensions!

How to evaluate it?

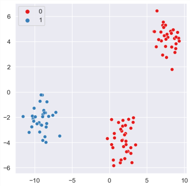
Tools to Evaluate and Compare Clusters



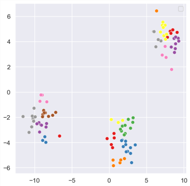
($\mathcal{T}$) Ground truth



($\mathcal{R}$) Random, $K = 3$ classes

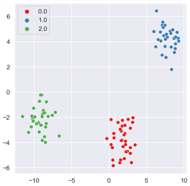


($\mathcal{C}_2$) $K = 2$ classes

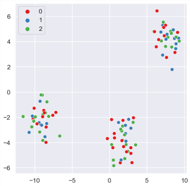


($\mathcal{C}_{20}$) $K = 20$ classes

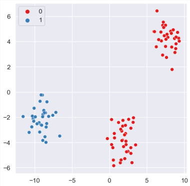
How to Evaluate Clustering Results?



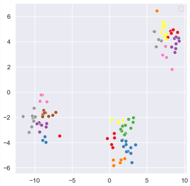
($\mathcal{T}$) Ground truth



($\mathcal{R}$) Random, $K = 3$ classes



($\mathcal{C}_2$) $K = 2$ classes



($\mathcal{C}_{20}$) $K = 20$ classes

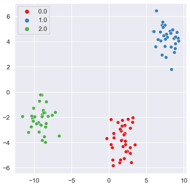
Internal metrics: *Real situation*

No need to know the ground truth.

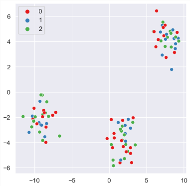
External metrics:

Specific clustering metrics when ground truth is known.

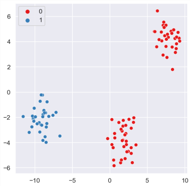
How to Evaluate Clustering Results?



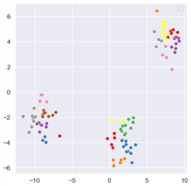
($\mathcal{T}$) Ground truth



($\mathcal{R}$) Random, $K = 3$ classes



($\mathcal{C}_2$) $K = 2$ classes



($\mathcal{C}_{20}$) $K = 20$ classes

Internal metrics: *Real situation*

No need to know the ground truth.

- Silhouette coefficient,
- Davies–Bouldin index,
- Dunn Index,
- R-Square (RSQ) and Semi-Partial R-Square ($SPRSQ$) criteria,
- Calinski-Harabasz score.

External metrics:

Specific clustering metrics when ground truth is known.

- Purity,
- Clustering accuracy,
- Folkes-Mallows index,
- Normalized Mutual Information.

Example of Internal Metric: Silhouette Coefficient

Let x_i, where $i \in \mathcal{C}_k$. n points, K clusters.

- **Cohesion**: Mean distance between x_i and other points in $\mathcal{C}_k$:

$$a(i) = \frac{1}{|\mathcal{C}_k| - 1} \sum_{j \in \mathcal{C}_k, j \neq i} d(x_i, x_j)$$

- **Separation**: Mean distance between x_i and the points of the closest other clusters:

$$b(i) = \min_{\ell \neq k} \frac{1}{|\mathcal{C}_\ell|} \sum_{j \in \mathcal{C}_\ell} d(x_i, x_j)$$

Example of Internal Metric: Silhouette Coefficient

Let x_i, where $i \in \mathcal{C}_k$. n points, K clusters.

- **Cohesion**: Mean distance between x_i and other points in $\mathcal{C}_k$:

$$a(i) = \frac{1}{|\mathcal{C}_k| - 1} \sum_{j \in \mathcal{C}_k, j \neq i} d(x_i, x_j)$$

- **Separation**: Mean distance between x_i and the points of the closest other clusters:

$$b(i) = \min_{\ell \neq k} \frac{1}{|\mathcal{C}_\ell|} \sum_{j \in \mathcal{C}_\ell} d(x_i, x_j)$$

$\rightsquigarrow$ **Silhouette score**:

- Point x_i: $s(i) \in [-1, 1]$

$$\boxed{s(i) = \frac{b(i) - a(i)}{\max\big(a(i), b(i)\big)}}$$

- Entire dataset:

$$\begin{aligned} S &= \frac{1}{n} \sum_{i=1}^{n} s(i) \\ &= \frac{1}{K} \sum_{k=1}^{K} \frac{1}{|\mathcal{C}_k|} \sum_{i \in \mathcal{C}_k} s(i) \end{aligned}$$

Example of Internal Metric: Silhouette Coefficient

Let x_i, where $i \in \mathcal{C}_k$. n points, K clusters.

- **Cohesion**: Mean distance between x_i and other points in $\mathcal{C}_k$:

$$a(i) = \frac{1}{|\mathcal{C}_k| - 1} \sum_{j \in \mathcal{C}_k, j \neq i} d(x_i, x_j)$$

- **Separation**: Mean distance between x_i and the points of the closest other clusters:

$$b(i) = \min_{\ell \neq k} \frac{1}{|\mathcal{C}_\ell|} \sum_{j \in \mathcal{C}_\ell} d(x_i, x_j)$$

	$(\mathcal{T})$	$(\mathcal{R})$	$(\mathcal{C}_2)$	$(\mathcal{C}_{20})$
Silhouette	0.83	-0.03	0.66	0.39

$\rightsquigarrow$ **Silhouette score**:

- Point x_i: $s(i) \in [-1, 1]$

$$\boxed{s(i) = \frac{b(i) - a(i)}{\max\big(a(i), b(i)\big)}}$$

- Entire dataset:

$$\begin{aligned} S &= \frac{1}{n} \sum_{i=1}^{n} s(i) \\ &= \frac{1}{K} \sum_{k=1}^{K} \frac{1}{|\mathcal{C}_k|} \sum_{i \in \mathcal{C}_k} s(i) \end{aligned}$$

Function `silhouette_score` from the `sklearn.metrics` package.

Inertia-Based Criteria

Let a partition $\mathcal{P}_K$.

- **R-Square**: $$\boxed{RSQ(\mathcal{P}_K) = \frac{I_{Inter}(\mathcal{P}_K)}{I_{Tot}} = 1 - \frac{I_{Intra}(\mathcal{P}_K)}{I_{Tot}}}$$
- **Semi-Partial R-Square**: $SPRSQ(\mathcal{P}_K) = \dfrac{I_{Inter}(\mathcal{P}_K) - I_{Inter}(\mathcal{P}_{K-1})}{I_{Tot}}$
- **Calinski-Harabasz** (CH): $$\boxed{PseudoF(\mathcal{P}_K) = \frac{I_{Inter}(\mathcal{P}_K)}{I_{Intra}(\mathcal{P}_K)} \times \frac{n-K}{K-1}}$$

	$(\mathcal{T})$	$(\mathcal{R})$	$(\mathcal{C}_2)$	$(\mathcal{C}_{20})$
Silhouette	0.83	-0.03	0.66	0.39
Calinski-Harabasz	1549.85	0.03	225.78	1009.70
Davies-Bouldin	0.24	64.40	0.45	0.66

Example of External Metric: Purity

Let $\mathcal{P}_L^\star = \{\mathcal{C}_1^\star, \ldots, \mathcal{C}_{K^\star}^\star\}$ be the ground truth partition, n points.

Consider a partition $\mathcal{P}_K = \{\mathcal{C}_1, \ldots, \mathcal{C}_K\}$.

$$\boxed{\mathcal{P}urity(\mathcal{P}_K) = \frac{1}{n} \sum_{k=1}^{K} \max_{\ell \in [\![1, K^\star]\!]} |\mathcal{C}_\ell^\star \cap \mathcal{C}_k|}$$

	$(\mathcal{T})$	$(\mathcal{R})$	$(\mathcal{C}_2)$	$(\mathcal{C}_{20})$
Silhouette	0.83	-0.03	0.66	0.39
Calinski-Harabasz	1549.85	0.03	225.78	1009.70
Davies-Bouldin	0.24	64.40	0.45	0.66
Purity score	1	0.36	0.67	1

Issue: More clusters, better score.

How to evaluate it?

Tools to Evaluate and Compare Clusters

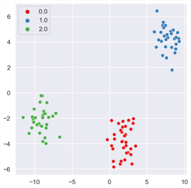
How to Compare two Clusterings?

Let us suppose that we have obtained two partitions from the same data :

$$\mathcal{P}_K = \{\mathcal{C}_1, \dots, \mathcal{C}_K\} \qquad \text{and} \qquad \mathcal{Q}_L = \{\mathcal{D}_1, \dots, \mathcal{D}_L\}$$

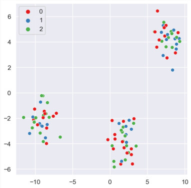
Question: How to compare these two classifications?

- Contingency table,
- Rand Index (RI) and Adjusted Rand Index (ARI),
- Variation of information,
- ...



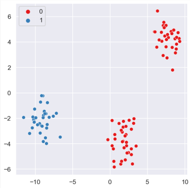
($\mathcal{T}$) Ground truth

vs.



($\mathcal{R}$) Random, $K = 3$ classes

vs.



($\mathcal{C}_2$) $K = 2$ classes

(Adjusted) Rand Index

$\mathcal{P}_K$ vs. $\mathcal{Q}_L$	Grouped in $\mathcal{P}_K$	Separated in $\mathcal{Q}_L$
Grouped in $\mathcal{P}_K$	a	b
Separated in $\mathcal{Q}_L$	c	d

$a + b$: Agreements between $\mathcal{P}_K$ and $\mathcal{Q}_L$.

$c + d$: Disagreements.

- **Rand Index**: Proportion of point pairs that are grouped in the same way in both partitions.

$$\boldsymbol{RI}(\mathcal{P}_K, \mathcal{Q}_L) = \frac{a + d}{a + b + c + d}$$

(Adjusted) Rand Index

$\mathcal{P}_K$ vs. $\mathcal{Q}_L$	Grouped in $\mathcal{P}_K$	Separated in $\mathcal{Q}_L$
Grouped in $\mathcal{P}_K$	a	b
Separated in $\mathcal{Q}_L$	c	d

$a + b$: Agreements between $\mathcal{P}_K$ and $\mathcal{Q}_L$.

$c + d$: Disagreements.

- **Rand Index**: Proportion of point pairs that are grouped in the same way in both partitions.

$$\boxed{\boldsymbol{RI}(\mathcal{P}_K, \mathcal{Q}_L) = \frac{a + d}{a + b + c + d}}$$

- **Adjusted Rand Index**: Let $n_{k\ell} = |\mathcal{C}_k \cap \mathcal{D}_\ell|$, $n_{k+} = \sum_{\ell=1}^{L} n_{k\ell}$, $n_{+\ell} = \sum_{k=1}^{K} n_{k\ell}$.
  - $RI = \sum_{k\ell} \binom{n_{k\ell}}{2}$
  - $\mathbb{E}[RI] = \frac{\sum_k \binom{n_{k+}}{2} \times \sum_\ell \binom{n_{+\ell}}{2}}{\binom{n}{2}}$,

    Indices obtained by randomly partitioning the data
  - $\max(RI) = \frac{1}{2}\left(\sum_k \binom{n_{k+}}{2} + \sum_\ell \binom{n_{+\ell}}{2}\right)$

$$\boxed{\boldsymbol{ARI}(\mathcal{P}_K, \mathcal{Q}_L) = \frac{RI - \mathbb{E}[RI]}{\max(RI) - \mathbb{E}[RI]}}$$

The closer the ARI is to 1, the more similar the two partitions are.

Contingency Table

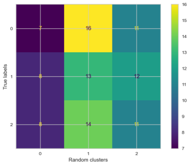
- **Contingency table** to observe if classes are shared, split, *etc.*

$$
\begin{aligned}
n_{k\ell} &= |\mathcal{C}_k \cap \mathcal{D}_\ell| \\
&= \#\left\{ i \in [\![1, n]\!] \,|\, i \in \mathcal{C}_k \cap \mathcal{D}_\ell \right\}
\end{aligned}
$$

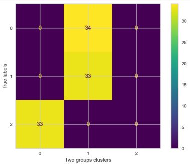
$$
n_{k+} = \sum_{\ell=1}^{L} n_{k\ell}
$$

$$
n_{+\ell} = \sum_{k=1}^{K} n_{k\ell}
$$

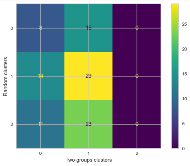
$\mathcal{P}_K$ *vs.* $\mathcal{Q}_L$	$\mathcal{D}_1$	$\mathcal{D}_2$	$\ldots$	$\mathcal{D}_L$	Sums
$\mathcal{C}_1$	n_{11}	n_{12}	$\ldots$	n_{1L}	n_{1+}
$\mathcal{C}_2$	n_{21}	n_{22}	$\ldots$	n_{2L}	n_{2+}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$\mathcal{C}_K$	n_{K1}	n_{K2}	$\ldots$	n_{KL}	n_{K+}
Sums	n_{+1}	n_{+2}	$\ldots$	n_{+L}	n



$(\mathcal{T})$ *vs.* $(\mathcal{R})$



$(\mathcal{T})$ *vs.* $(\mathcal{C}_2)$



$(\mathcal{R})$ *vs.* $(\mathcal{C}_2)$

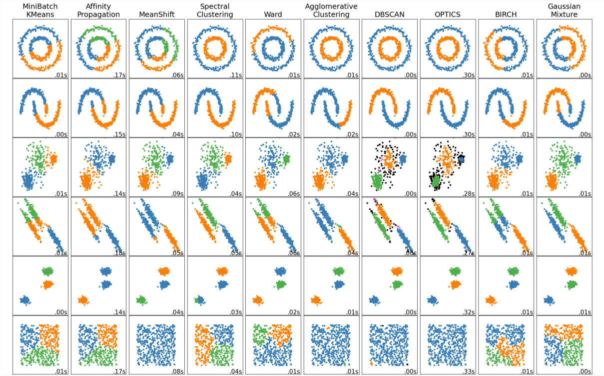
Functions `confusion_matrix` and `ConfusionMatrixDisplay` from `sklearn.metrics`.

How to choose a clustering algorithm?

Course Outline

A Variety of Methods

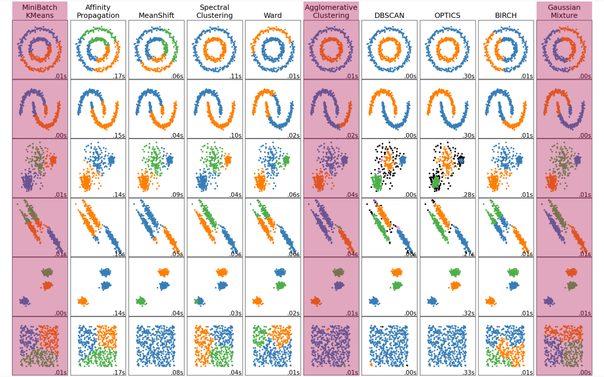
- Clustering methods distinguish by:
  - Type of “similarity” between individuals: Distance, probability distribution, shape, *etc.*
  - Type of “partitioning”: Hard or fuzzy clustering.
- Various categories of methods:
  - Distance-based,
  - Connectivity-based,
  - Density-based,
  - *etc.*



From: scikit-learn.org/stable/auto_examples/cluster/plot_cluster_comparison.html

A Variety of Methods

- Clustering methods distinguish by:
  - Type of “similarity” between individuals: Distance, probability distribution, shape, *etc.*
  - Type of “partitioning”: Hard or fuzzy clustering.
- Various categories of methods:
  - Distance-based,
  - Connectivity-based,
  - Density-based,
  - *etc.*



From: scikit-learn.org/stable/auto_examples/cluster/plot_cluster_comparison.html

Different Families of Clustering Algorithms

	Distance-based	Connectivity-based	Density-based
Pros	• Can conduct inference on new data points • Usually fast	• Does not need access to data points values (only distances) • Can handle non linearly separated clusters	• Does not need access to data points values • Can handle non linearly separable clusters • Does not need number of clusters • Can handle outliers
Cons	• Number of clusters required • No outlier detection • Need access to point values	• Number of clusters required • No outlier detection • Usually slow • Cannot conduct inference	• Usually slow • Cannot be used for inference
Example	K-means	Hierarchical clustering	DBSCAN