



UNIVERSITAT DE  
BARCELONA

# *Introduction to Machine Learning*

Master in Artificial Intelligence  
UPC, UB, URV





# **Course. Introduction to Machine Learning**

## **Work 3. Clustering exercise**

### **Session 3**

**Dr. Maria Salamó Llorente**  
Dept. Mathematics and Informatics,  
Faculty of Mathematics and Informatics,  
University of Barcelona

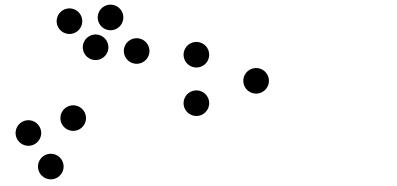


1. Introduction (session 1)
2. OPTICS with sklearn (session 1)
3. Spectral Clustering with sklearn (session 1)
4. K-Means and Improved K-Means (your own code) (session 2)
5. Fuzzy clustering (your own code) (session 2)
6. Validation techniques (using sklearn validation metrics) (session 3)

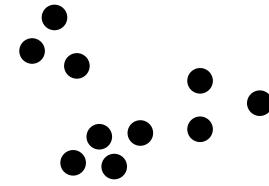


## 6. VALIDATION TECHNIQUES

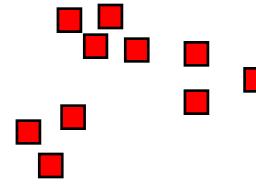
# Clustering Validation



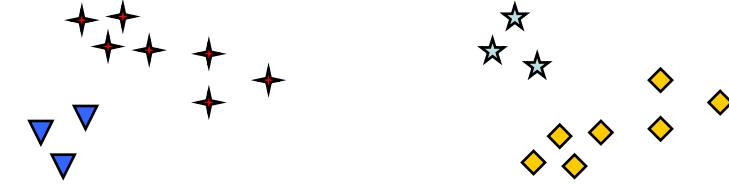
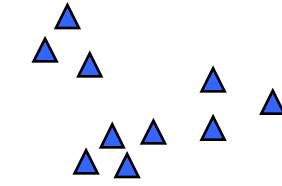
*How many clusters?*



*Six Clusters*



*Two Clusters*



*Four Clusters*

**Which is the best clustering?**

## Supervised classification:

- Class labels known for ground truth
- Accuracy, precision, recall

## Cluster analysis

- No class labels

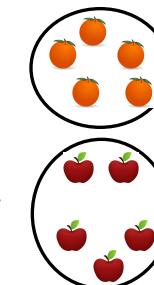
## Validation need to:

- Compare clustering algorithms
- Solve number of clusters
- Avoid finding patterns in noise

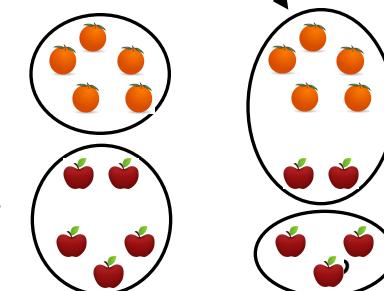
$$\text{Precision} = 5/5 = 100\%$$

$$\text{Recall} = 5/7 = 71\%$$

Oranges:



Apples:



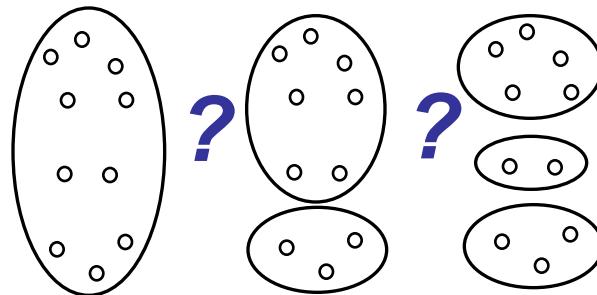
$$\text{Precision} = 3/5 = 60\%$$

$$\text{Recall} = 3/3 = 100\%$$

- **Internal criterion:** A good clustering will produce high quality clusters in which:
  - the intra-class (that is, intra-cluster) similarity is high
  - the inter-class similarity is low
  - The measured quality of a clustering depends on both the example representation and the similarity measure used
- **External criterion:** The quality of a clustering is also measured by its ability to discover some or all of the hidden patterns or latent classes
  - Assessable with gold standard data

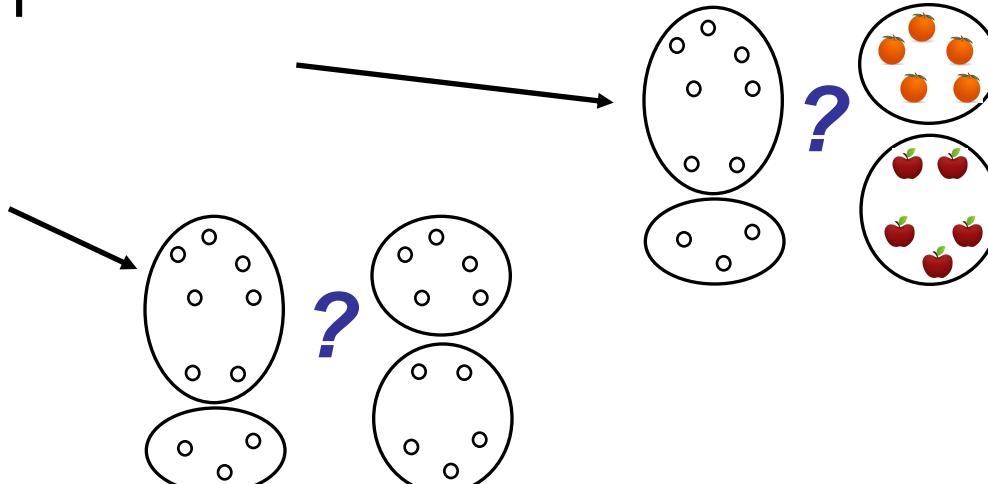
## Internal Index

- Validate *without* external info
- With different number of clusters
- Solve the number of clusters



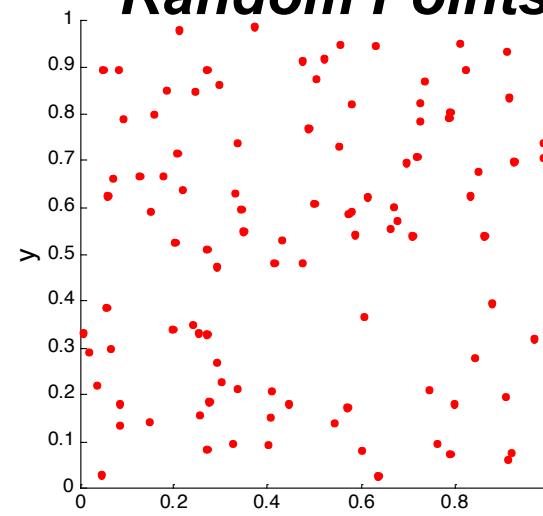
## External Index

- Validate against ground truth
- Compare two clusters:  
(how similar)

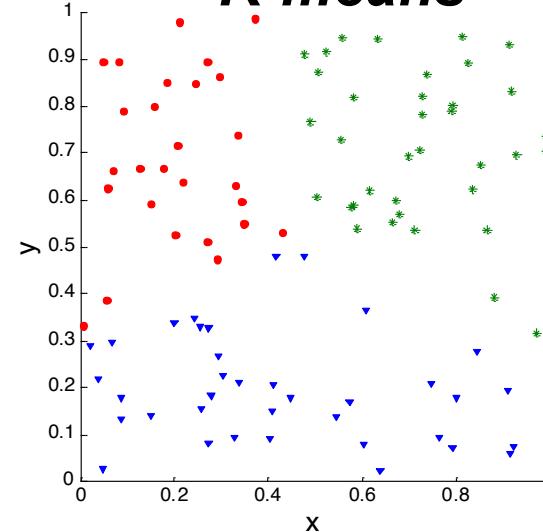


# Clustering of random data

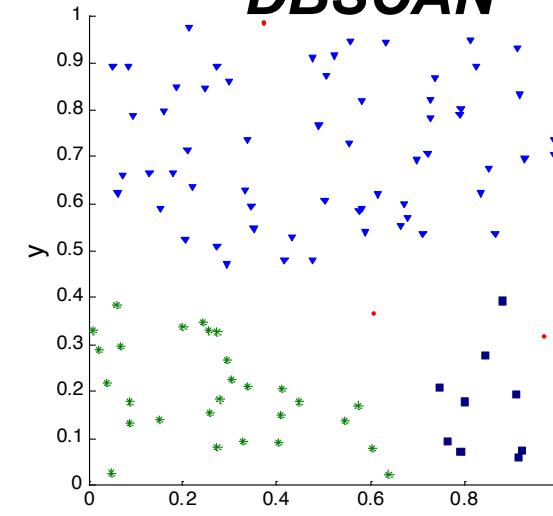
**Random Points**



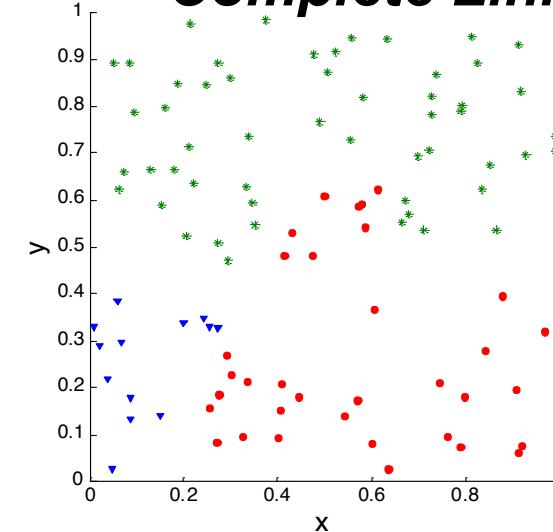
**K-means**



**DBSCAN**

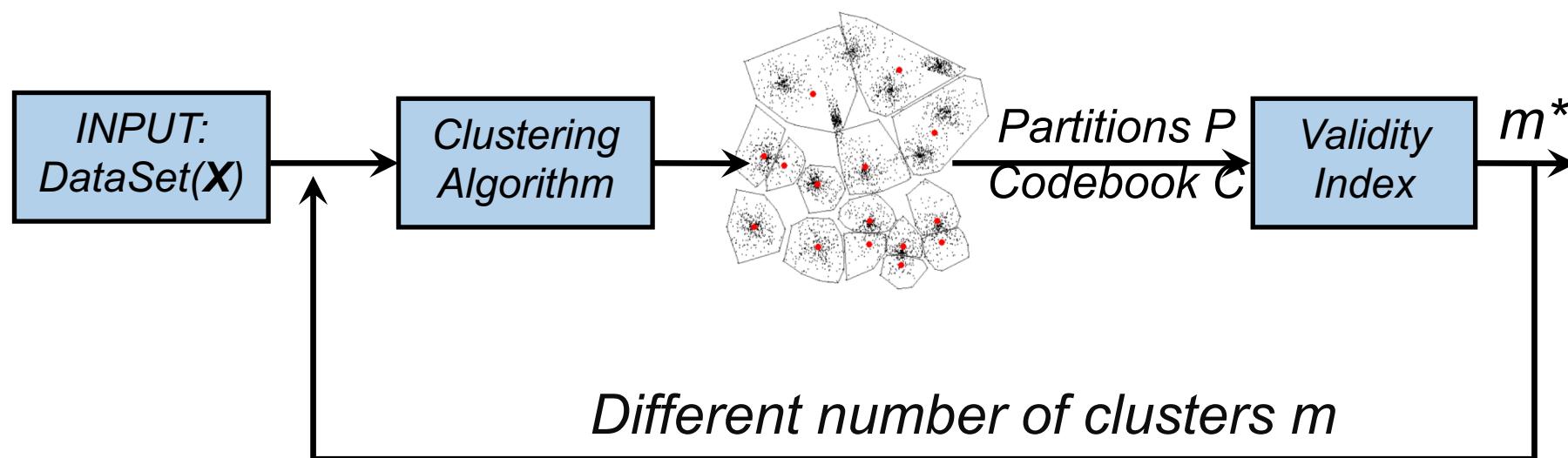


**Complete Link**



# Cluster validation process

- **Cluster validation** refers to procedures that evaluate the results of clustering in a **quantitative** and **objective** fashion [Jain & Dubes, 1988]
  - How to be “quantitative”: To employ the measures.
  - How to be “objective”: To validate the measures!



# Internal indexes

- Ground truth is rarely available but unsupervised validation must be done.
- Minimizes (or maximizes) internal index:
  - Variances of within cluster and between clusters
  - Rate-distortion method
  - F-ratio
  - Davies-Bouldin index (DBI)
  - Bayesian Information Criterion (BIC)
  - Silhouette Coefficient
  - Minimum description principle (MDL)
  - Stochastic complexity (SC)

# Internal indexes

Table B.1: Formulas for internal indexes

Name	Formula
SSW	$SSW = \frac{1}{N} \sum_{i=1}^N \ x_i - C_{p_i}\ ^2$
SSB	$SSB = \frac{2}{M(M-1)} \sum_{i=1}^M \sum_{j=1, j \neq i}^M \ C_i - C_j\ ^2$
Calinski-Harabasz index	$CH = \frac{SSB/(M-1)}{SSW(N-M)}$
Hartigan	$H_M = \left( \frac{SSW_M}{SSW_{M+1}} - 1 \right) (N - M - 1)$ <i>or : </i> $H_M = \log (SSB_M / SSW_M)$
Krzanowski-Lai index	$diff_M = (M-1)^{2/D} SSW_{M-1} - M^{2/D} SSW_M$ $KL_M =  diff_M  /  diff_{M+1} $
Ball&Hall	$BH_M = SSW_M / M$
Xu-index	$Xu = D \log (\sqrt{SSW_M / (DN^2)}) + \log M$
Dunn's index	$Dunn = \sum_{i=1}^M \frac{\max(\ x_j - C_i\ ^2)_{j \in C_i}}{S_i + S_j}, i \neq j$
Davies&Bouldin index	$where : d_{ij} = \ C_i - C_j\ ^2, S_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \ x_j - C_i\ ^2$ <i>and, </i> $R_i = \max_{j=1, \dots, M} R_{ij}, i = 1, \dots, M$ $DBI = \frac{1}{M} \sum_{i=1}^M R_i$

# Internal indexes

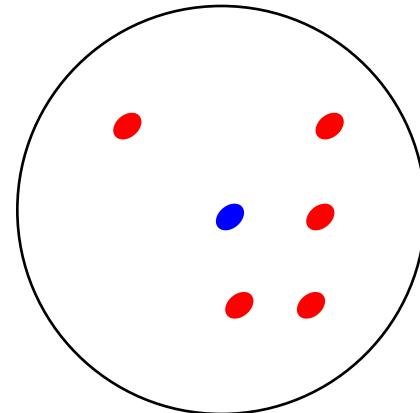
Silhouette Coefficients	$a(x_i) = \frac{1}{n_m - 1} \sum_{j=1, j \neq i}^{n_m} \ x_i - x_j\ _{x_i, x_j \in C_m}^2$ $b(x_i) = \min_t \left\{ \frac{1}{n_t} \sum_{j \in C_t} \ x_i - x_j\ ^2 \right\}_{x_i \notin C_t}$ $s(x_i) = \frac{b(x_i) - a(x_i)}{\max(a(x_i), b(x_i))}$ $SC = \frac{1}{N} \sum_{i=1}^N s(x_i)$ $b(x_i) = \min \left\{ \sum_{t \neq m} \ C_t - C_m\ ^2 \right\}_{x_i \notin C_t} (SC' 2008)$
RMSSTD	$RMSSTD = \frac{\sum_{k=1, \dots, M} \sum_{i=1}^{n_{kd}} (x_i - \bar{x}^d)^2}{\sum_{k=1, \dots, M} \sum_{d=1, \dots, D} (n_{kd} - 1)}$
R-square	$RS = \frac{SST - SSW}{SST} = \frac{\sum_{d=1, \dots, D} \sum_{i=1}^{n_d} (x_i - \bar{x}^d)^2 - \sum_{k=1, \dots, M} \sum_{d=1, \dots, D} (x_i - \bar{x}^d)^2}{\sum_{d=1, \dots, D} \sum_{i=1}^{n_d} (x_i - \bar{x}^d)^2}$
Bayesian Information Criterion	$BIC = L * N - \frac{1}{2} M(D + 1) \sum_{i=1}^M \log(n_i)$
Xie-Beni	$XB = \frac{\sum_{i=1}^N \sum_{k=1}^M u_{ik}^2 \ x_i - C_k\ ^2}{N \min_{t \neq s} \{ \ C_t - C_s\ ^2 \}}$
Partition Coefficient	$PC = \sum_{i=1}^N \sum_{k=1}^M u_{ik}^2 / N$
Partition Entropy	$PE = - \left( \sum_{i=1}^N \sum_{k=1}^M u_{ik} \log(u_{ik}) \right) / N$

Soft partitions

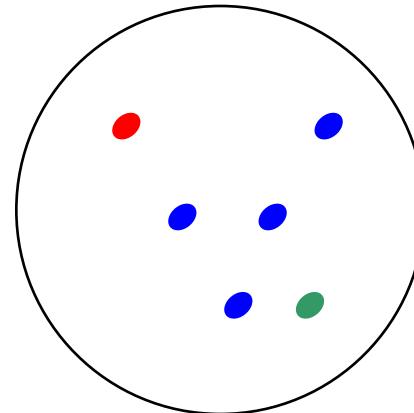
- Assesses clustering with respect to ground truth
- Assume that there are  $C$  gold standard classes, while our clustering algorithms produce  $k$  clusters,  $\pi_1, \pi_2, \dots, \pi_k$  with  $n_i$  members.
- **Simple measure:** purity, the ratio between the dominant class in the cluster  $\pi_i$  and the size of cluster  $\pi_i$

$$\text{Purity}(\pi_i) = \frac{1}{n_i} \max_j (n_{ij}) \quad j \in C$$

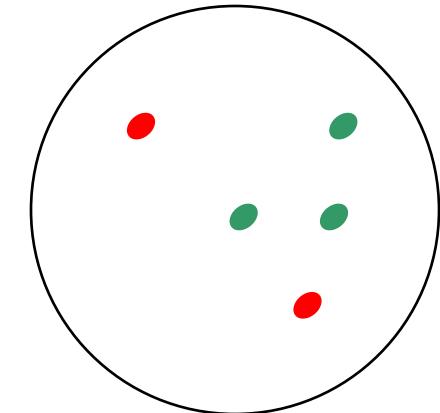
# Purity Example



*Cluster I*



*Cluster II*



*Cluster III*

**Cluster I:** Purity =  $1/6 (\max(5, 1, 0)) = 5/6 (0,83)$

**Cluster II:** Purity =  $1/6 (\max(1, 4, 1)) = 4/6 (0,66)$

**Cluster III:** Purity =  $1/5 (\max(2, 0, 3)) = 3/5 (0,60)$

Measure the number of pairs that are in:

Same class **both** in  $P$  and  $G$ .

$$a = \frac{1}{2} \sum_{i=1}^K \sum_{j=1}^{K'} n_{ij} (n_{ij} - 1)$$

Same class in  $P$  but different in  $G$ .

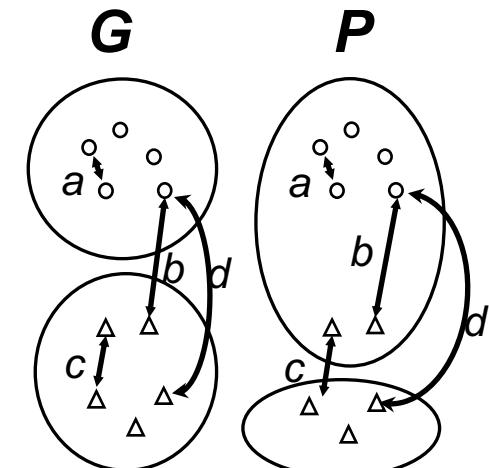
$$b = \frac{1}{2} \left( \sum_{j=1}^{K'} n_{\cdot j}^2 - \sum_{i=1}^K \sum_{j=1}^{K'} n_{ij}^2 \right)$$

Different classes in  $P$  but same in  $G$ .

$$c = \frac{1}{2} \left( \sum_{i=1}^K n_{i \cdot}^2 - \sum_{i=1}^K \sum_{j=1}^{K'} n_{ij}^2 \right)$$

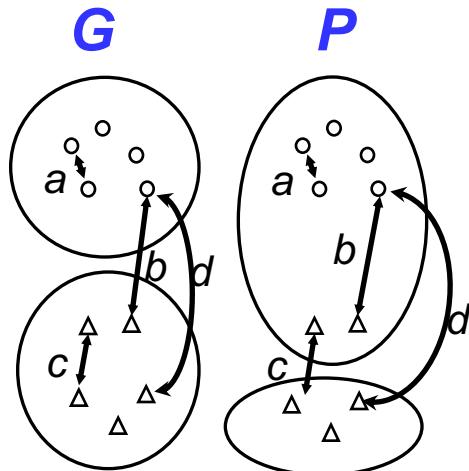
Different classes **both** in  $P$  and  $G$ .

$$d = \frac{1}{2} \left( N^2 + \sum_{i=1}^K \sum_{j=1}^{K'} n_{ij}^2 - \left( \sum_{i=1}^K n_{i \cdot}^2 + \sum_{j=1}^{K'} n_{\cdot j}^2 \right) \right)$$



# Rand and Adjusted Rand index

[Rand, 1971] [Hubert and Arabie, 1985]



Agreement:  $a, d$

Disagreement:  $b, c$

$$RI(P, G) = \frac{a + d}{a + b + c + d}$$

$$ARI = \frac{RI - E(RI)}{1 - E(RI)}$$

If true class labels (*ground truth*) are known, the validity of a clustering can be verified by comparing the class labels and clustering labels.

$N$	.	=	$n_{11}$	$n_{12}$	...	$n_{1l}$	$n_{1..}$
.	$n_{..}$		$n_{21}$	$n_{22}$	...	$n_{2l}$	$n_{2..}$
			:	:	..	:	:
			$n_{k1}$	$n_{k2}$	...	$n_{kl}$	$n_{k..}$
			$n_{.1}$	$n_{.2}$	...	$n_{.l}$	$n_{..}$

$n_{ij}$  = number of objects in class  $i$  and cluster  $j$



- Pair counting
  - Chi-Squared Coefficient
  - Rand Index
  - Adjusted Rand Index
  - Fowlkes-Mallows Index
  - Mirkin Metric
- Other measures
  - Information theoretic
    - Mutual Information Metric (MI), Normalized Mutual Information, Variation of Information
  - Set matching
    - Jaccard Index, Normalized Van Dongen, Pair Set Index

# Summary of external indexes

Table 1: External Cluster Validation Measures.

Measure	Notation	Definition	Range
1 Entropy	$E$	$-\sum_i p_i (\sum_j \frac{p_{ij}}{p_i} \log \frac{p_{ij}}{p_i})$	$[0, \log K']$
2 Purity	$P$	$\sum_i p_i (\max_j \frac{p_{ij}}{p_j})$	$(0,1]$
3 F-measure	$F$	$\sum_j p_j \max_i [2 \frac{p_{ij}}{p_i} \frac{p_{ij}}{p_j} / (\frac{p_{ij}}{p_i} + \frac{p_{ij}}{p_j})]$	$(0,1]$
4 Variation of Information	$VI$	$-\sum_i p_i \log p_i - \sum_j p_j \log p_j - 2 \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{p_i p_j}$	$[0, 2 \log \max(K, K')]$
5 Mutual Information	$MI$	$\sum_i \sum_j p_{ij} \log \frac{p_{ij}}{p_i p_j}$	$(0, \log K')$
6 Rand statistic	$R$	$[(\binom{n}{2}) - \sum_i (\binom{n_i}{2}) - \sum_j (\binom{n_j}{2}) + 2 \sum_{ij} (\binom{n_{ij}}{2})] / (\binom{n}{2})$	$(0,1]$
7 Jaccard coefficient	$J$	$\sum_{ij} (\binom{n_{ij}}{2}) / [\sum_i (\binom{n_i}{2}) + \sum_j (\binom{n_j}{2}) - \sum_{ij} (\binom{n_{ij}}{2})]$	$[0,1]$
8 Fowlkes and Mallows index	$FM$	$\sum_{ij} (\binom{n_{ij}}{2}) / \sqrt{\sum_i (\binom{n_i}{2}) \sum_j (\binom{n_j}{2})}$ $\frac{(\binom{n}{2}) \sum_{ij} (\binom{n_{ij}}{2}) - \sum_i (\binom{n_i}{2}) \sum_j (\binom{n_j}{2})}{\sqrt{\sum_i (\binom{n_i}{2}) \sum_j (\binom{n_j}{2})}  (\binom{n}{2}) - \sum_i (\binom{n_i}{2})   (\binom{n}{2}) - \sum_j (\binom{n_j}{2}) }}$	$[0,1]$ $(-1,1]$
9 Hubert $\Gamma$ statistic I	$\Gamma$	$\frac{[(\binom{n}{2}) - 2 \sum_i (\binom{n_i}{2}) - 2 \sum_j (\binom{n_j}{2}) + 4 \sum_{ij} (\binom{n_{ij}}{2})] / (\binom{n}{2})}{\sqrt{\sum_i (\binom{n_i}{2}) \sum_j (\binom{n_j}{2})}}$	$[0,1]$
10 Hubert $\Gamma$ statistic II	$\Gamma'$	$\sqrt{\sum_i (\binom{n_i}{2}) + \sum_j (\binom{n_j}{2}) - 2 \sum_{ij} (\binom{n_{ij}}{2})} / \sqrt{\sum_j (\binom{n_j}{2})}$	$[0, +\infty)$
11 Minkowski score	$MS$	$1 - \frac{1}{n} \max_{\sigma} \sum_j n_{\sigma(j),j}$	$[0,1]$
12 classification error	$\varepsilon$	$(2n - \sum_i \max_j n_{ij} - \sum_j \max_i n_{ij}) / 2n$	$[0, 1)$
13 van Dongen criterion	$VD$	$\sum_i p_i (\max_j \frac{p_{ij}}{p_i})$	$(0,1]$
14 micro-average precision	$MAP$	$\sum_i p_i (1 - \max_j \frac{p_{ij}}{p_i})$	$[0,1)$
15 Goodman-Kruskal coefficient	$GK$	$\sum_i p_i (1 - \max_j \frac{p_{ij}}{p_i})$	$[0,1)$
16 Mirkin metric	$M$	$\sum_i n_i^2 + \sum_j n_j^2 - 2 \sum_i \sum_j n_{ij}^2$	$[0, 2(\binom{n}{2})]$

Note:  $p_{ij} = n_{ij} / n$ ,  $p_i = n_i / n$ ,  $p_j = n_j / n$ .

- Clustering performance evaluation

```
from sklearn import metrics
```

- Adjusted Rand index
- Mutual information based scores
- Homogeneity, completeness and V-measure
- Fowlkes-Mallows scores
- Silhouette Coefficient
- Calinski-Harabaz Index
- Davies-Bouldin Index
- Contingency Matrix

# References on cluster validation

1. G.W. Milligan, and M.C. Cooper, "An examination of procedures for determining the number of clusters in a data set", *Psychometrika*, Vol.50, 1985, pp. 159-179.
2. E. Dimitriadou, S. Dolnicar, and A. Weingassel, "An examination of indexes for determining the number of clusters in binary data sets", *Psychometrika*, Vol.67, No.1, 2002, pp. 137-160.
3. D.L. Davies and D.W. Bouldin, "A cluster separation measure ", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1(2), 224-227, 1979.
4. J.C. Bezdek and N.R. Pal, "Some new indexes of cluster validity ", *IEEE Transactions on Systems, Man and Cybernetics*, 28(3), 302-315, 1998.
5. H. Bischof, A. Leonardis, and A. Selb, "MDL Principle for robust vector quantization", *Pattern Analysis and Applications*, 2(1), 59-72, 1999.
6. P. Fränti, M. Xu and I. Kärkkäinen, "Classification of binary vectors by using DeltaSC-distance to minimize stochastic complexity", *Pattern Recognition Letters*, 24 (1-3), 65-73, January 2003.



# References on cluster validation

7. G.M. James, C.A. Sugar, "Finding the Number of Clusters in a Dataset: An Information-Theoretic Approach". *Journal of the American Statistical Association*, vol. 98, 397-408, 2003.
8. P.K. Ito, Robustness of ANOVA and MANOVA Test Procedures. In: Krishnaiah P. R. (ed), *Handbook of Statistics 1: Analysis of Variance*. North-Holland Publishing Company, 1980.
9. I. Kärkkäinen and P. Fränti, "Dynamic local search for clustering with unknown number of clusters", *Int. Conf. on Pattern Recognition (ICPR'02)*, Québec, Canada, vol. 2, 240-243, August 2002.
10. D. Pellag and A. Moore, "X-means: Extending K-Means with Efficient Estimation of the Number of Clusters", *Int. Conf. on Machine Learning (ICML)*, 727-734, San Francisco, 2000.
11. S. Salvador and P. Chan, "Determining the Number of Clusters/Segments in Hierarchical Clustering/Segmentation Algorithms", *IEEE Int. Con. Tools with Artificial Intelligence (ICTAI)*, 576-584, Boca Raton, Florida, November, 2004.
12. M. Gyllenberg, T. Koski and M. Verlaan, "Classification of binary vectors by stochastic complexity ". *Journal of Multivariate Analysis*, 63(1), 47-72, 1997.

13. M. Gyllenberg, T. Koski and M. Verlaan, "Classification of binary vectors by stochastic complexity ". *Journal of Multivariate Analysis*, 63(1), 47-72, 1997.
14. X. Hu and L. Xu, "A Comparative Study of Several Cluster Number Selection Criteria", *Int. Conf. Intelligent Data Engineering and Automated Learning (IDEAL)*, 195-202, Hong Kong, 2003.
15. Kaufman, L. and P. Rousseeuw, 1990. Finding Groups in Data: An Introduction to Cluster Analysis. *John Wiley and Sons, London*. ISBN: 10:0471878766.
16. [1.3] M.Halkidi, Y.Batistakis and M.Vazirgiannis: Cluster validity methods: part 1, SIGMOD Rec., Vol.31, No.2, pp.40-45, 2002
17. R. Tibshirani, G. Walther, T. Hastie. Estimating the number of clusters in a data set via the gap statistic. *J.R.Statist. Soc. B*(2001) 63, Part 2, pp.411-423.
18. T. Lange, V. Roth, M, Braun and J. M. Buhmann. Stability-based validation of clustering solutions. *Neural Computation*. Vol. 16, pp. 1299-1323. 2004.

19. Q. Zhao, M. Xu and P. Fränti, "Sum-of-squares based clustering validity index and significance analysis", *Int. Conf. on Adaptive and Natural Computing Algorithms (ICANNGA'09)*, Kuopio, Finland, LNCS 5495, 313-322, April 2009.
20. Q. Zhao, M. Xu and P. Fränti, "Knee point detection on bayesian information criterion", *IEEE Int. Conf. Tools with Artificial Intelligence (ICTAI)*, Dayton, Ohio, USA, 431-438, November 2008.
21. W.M. Rand, "Objective criteria for the evaluation of clustering methods," *Journal of the American Statistical Association*, 66, 846–850, 1971
22. L. Hubert and P. Arabie, "Comparing partitions", *Journal of Classification*, 2(1), 193-218, 1985.
23. P. Fränti, M. Rezaei and Q. Zhao, "Centroid index: Cluster level similarity measure", *Pattern Recognition*, 2014. (accepted)



# **Course. Introduction to Machine Learning**

## **Work 3. Clustering exercise**

### **Session 3**

**Dr. Maria Salamó Llorente**  
Dept. Mathematics and Informatics,  
Faculty of Mathematics and Informatics,  
University of Barcelona