Course. Introduction to Machine Learning
Lecture 3. Introduction to unsupervised learning and Cluster Analysis (Part II)

Dr. Maria Salamó Llorente

Dept. Mathematics and Informatics,

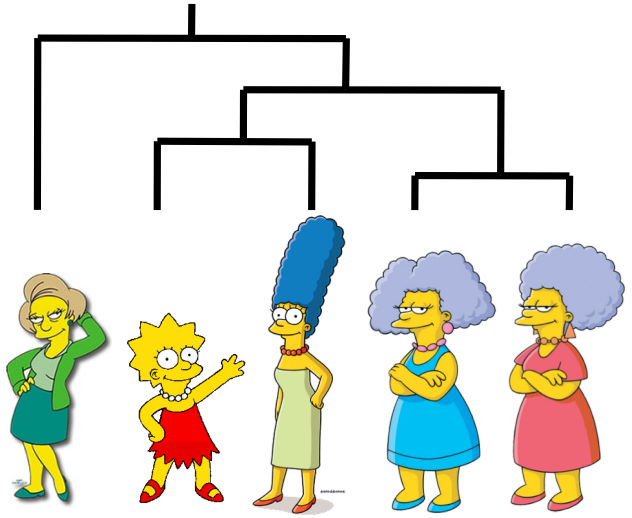
Faculty of Mathematics and Informatics,

University of Barcelona

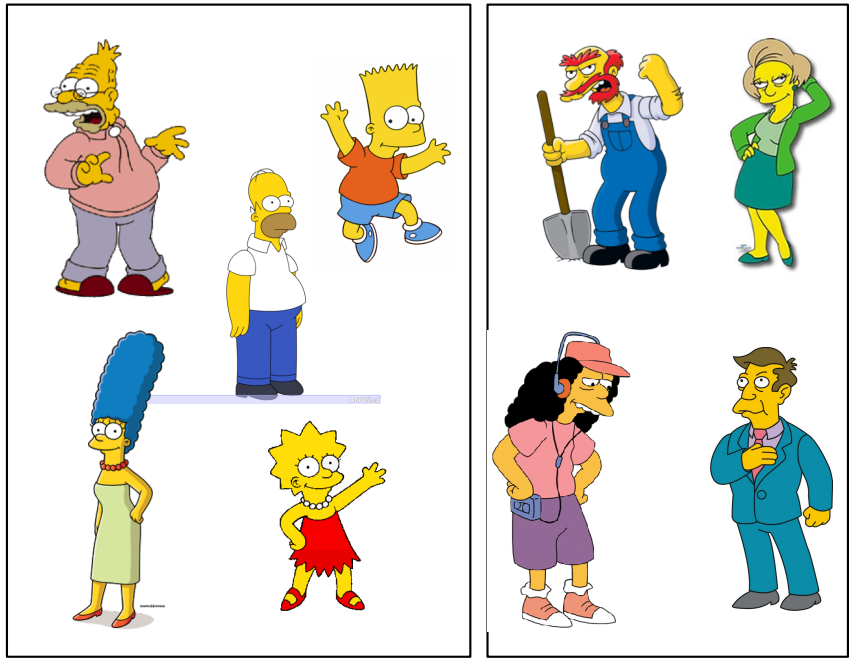
UNIVERSITAT DE BARCELONA

Recall the Taxonomy

- **Hierarchical algorithms**
  - Examples are organized as a binary tree
  - No explicit division in groups
    - **Bottom-up**
    - Top-down



- **Partitional algorithms**
  - Usually start with a random (partial) partitioning
  - Refine it iteratively:
    - **K-means clustering**
    - **Mixture-model based clustering**



Partitional algorithms

- **Method**: construct a partition of ***n*** objects into a set of ***k*** clusters
- **Given**: a set of objects (training set) and typically must provide the number of desired clusters, ***K***.
- **Basic process**:
  - Randomly choose *K* instances as *seeds*, one per cluster
  - Form initial clusters based on these seeds
  - Iterate, repeatedly reallocating instances to different clusters to improve the overall clustering
  - Stop when clustering converges or after a fixed number of iterations

K-Means basis

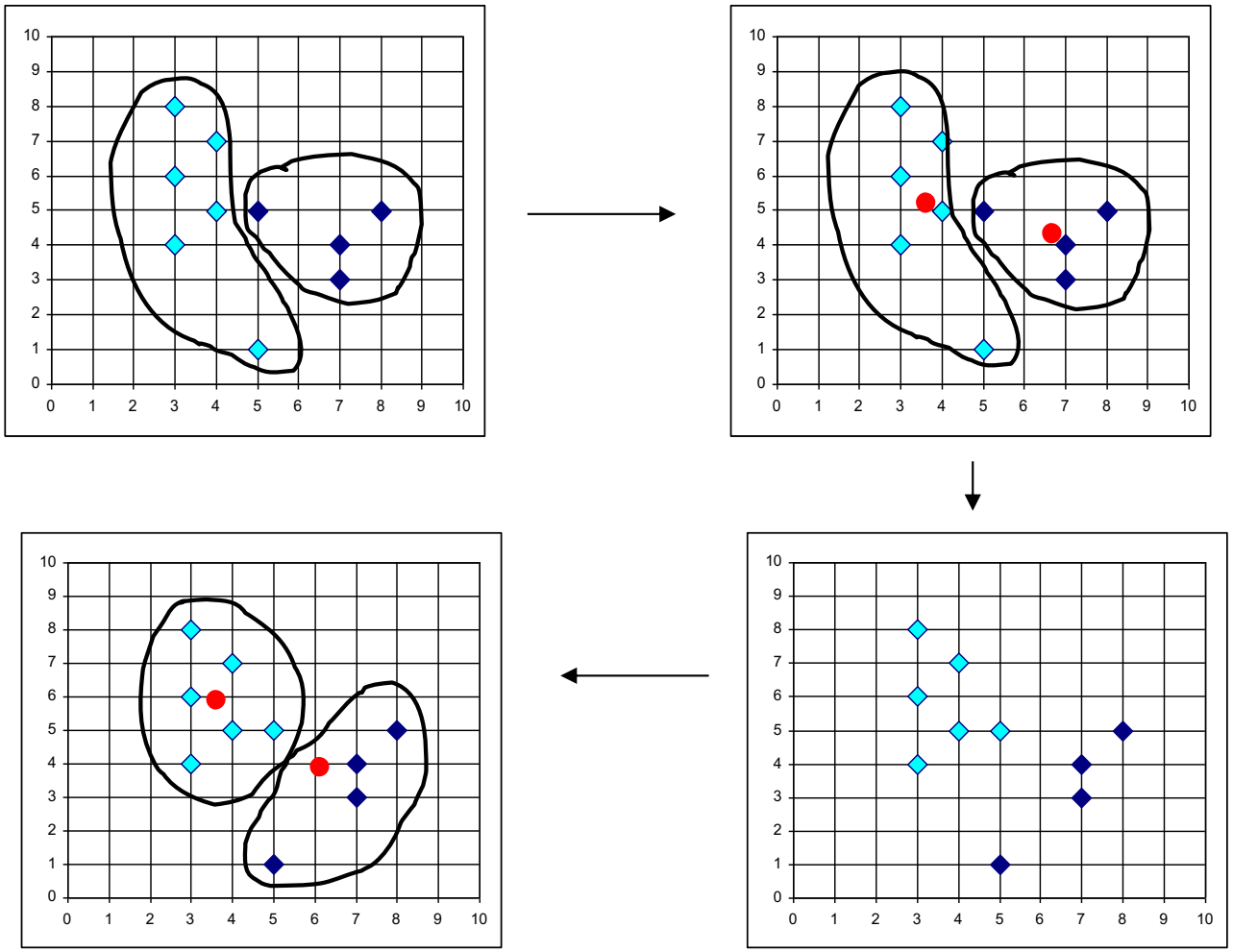
- Assumes instances are **real-valued vectors**
- Clusters based on ***centroids***, ***center of gravity***, or mean of points in a cluster, ***c***:

$$\vec{\mu}(c) = \frac{1}{|c|}\sum_{\vec{x}\in c}\vec{x}$$

- Reassignment of instances to clusters is **based on distance** to the current cluster centroids

The *K-Means* Clustering Method

Example



Distance Metrics

- Euclidean distance (L_2 norm):

$$L_2(\vec{x},\vec{y}) = \sum_{i=1}^{m}(x_i - y_i)^2$$

- L_1 norm:

$$L_1(\vec{x},\vec{y}) = \sum_{i=1}^{m}|x_i - y_i|$$

- Cosine Similarity (transform to a distance by subtracting from 1):

$$1 - \frac{\vec{x} \bullet \vec{y}}{|\vec{x}| \cdot |\vec{y}|}$$

K-Means Algorithm

*Let **d** be the distance measure between instances.*

1. *Decide on a value for k*
2. *Select k random instances $\{s_1, s_2, \ldots s_k\}$ as seeds.*
3. *(Decide the class membership)*

For each instance x_i:

Assign x_i to the cluster c_j such that $d(x_i, s_j)$ is minimal.

4. *(Update the seeds to the centroid of each cluster)*

For each cluster c_j

$s_j = \mu(c_j)$

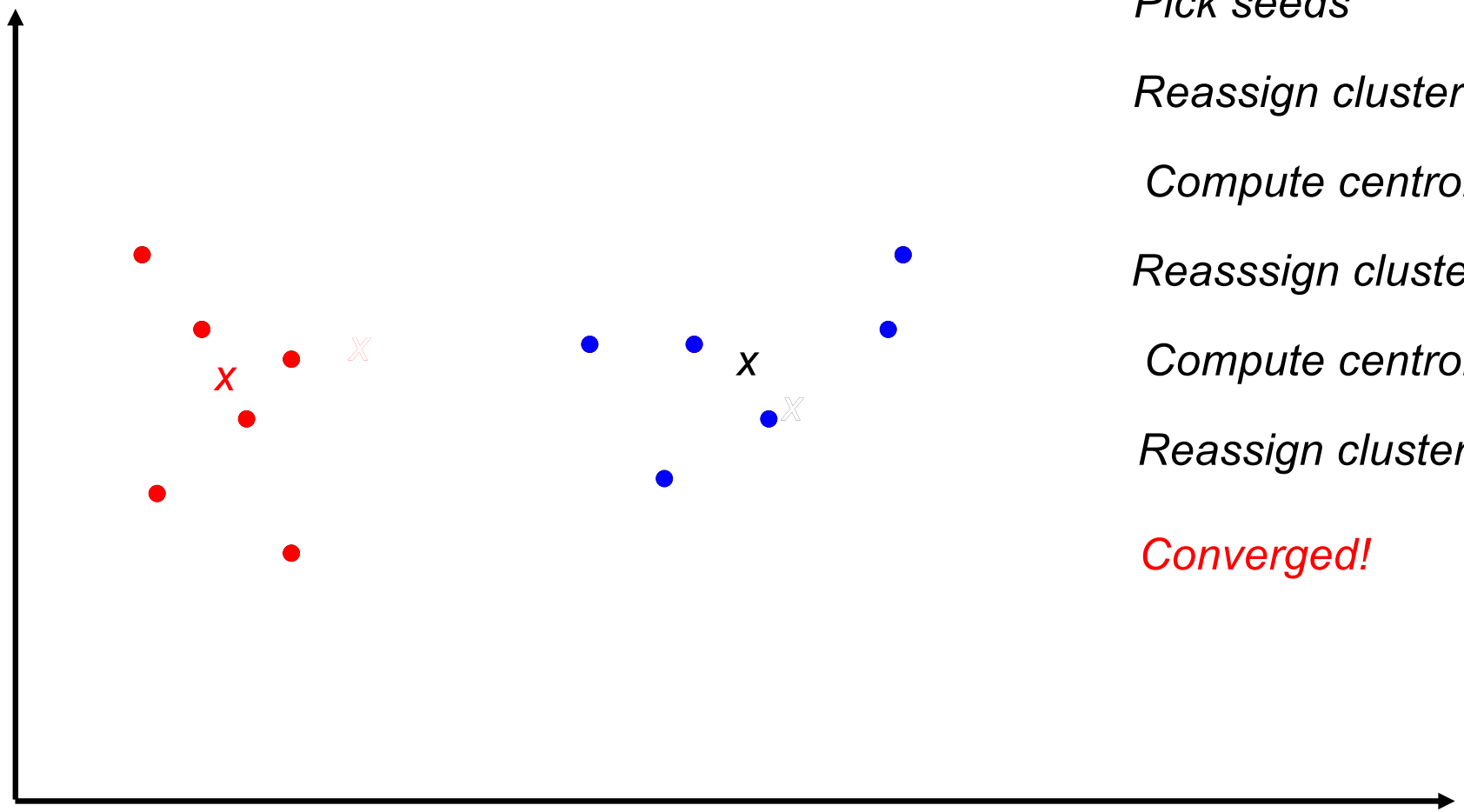
$$\vec{\mu}_k = \frac{1}{C_k} \sum_{i \in C_k} \vec{x}_i$$

5. *(Until clustering converges or other stopping criterion):*

If none of the N instances changed membership, exit

Otherwise, go to step 3

K Means Example (K=2)



Pick seeds

Reassign clusters

Compute centroids

Reasssign clusters

Compute centroids

Reassign clusters

Converged!

Time Complexity

- Assume computing distance between two instances is **O(*m*)** where ***m*** is the dimensionality of the vectors
- **Reassigning clusters**: **O(*kn*)** distance computations, or **O(*knm*)**.
- **Computing centroids**: Each instance vector gets added once to some centroid: **O(*nm*)**.
- Assume these two steps are each done once for *I* iterations: O(*Iknm*).
- Linear in all relevant factors, assuming a fixed number of iterations, more efficient than $O(n^2)$ or $O(n^3)$ HAC.

K-Means Objective

- The **objective** of k-means is to **minimize the total sum of the squared distance of every point to its corresponding cluster** centroid

$$\textit{Goodness measure (SD)} = \sum_{l=1}^{K} \sum_{x_i \in X_l} \| x_i - \mu_l \|^2$$

  - *Finding the global optimum is NP-hard.*
  - *The k-means algorithm is guaranteed to converge a local optimum.*

Comments on the *K-Means* Method

- **Strength**
  - *Relatively efficient*: *O*(*tkn*), where *n* is # instances, k is # clusters, and *t* is # iterations. Normally, *k*, *t* << *n*.
  - Often terminates at a *local optimum*. The *global optimum* may be found using techniques such as: *simulated annealing* or *genetic algorithms*
- **Weakness**
  - Applicable only when *mean* is defined; what about categorical data?
  - Need to specify k, the *number* of clusters, in advance
  - Unable to handle noisy data and *outliers*
  - Not suitable to discover clusters with *non-convex shapes*

Seed Choice

- Results can vary based on random seed selection

- Some seeds can result in poor convergence rate, or convergence to sub-optimal clusterings.
  - Select good seeds using a heuristic or the results of another method.
  - Try out multiple starting points (very important!)
  - Initialize with the results of another method

Variations of the *K-Means* method

- A few variants of the *k-means* which differ in
  - Selection of the initial *k* means
  - Dissimilarity calculations
  - Strategies to calculate cluster means
- Handling categorical data: *k-modes* (Huang'98)
  - Replacing means of clusters with modes
  - Using new dissimilarity measures to deal with categorical objects
  - Using a frequency-based method to update modes of clusters
  - A mixture of categorical and numerical data: *k-prototype* method

Extension of k-Means (towards hierarchical clustering): Bisecting k-Means

Start with a single cluster of all objects

1. Pick a cluster to split
2. Find two subclusters using the basic k-Means algorithm (bisecting step)
3. Repeat step 2 for *n* times and take the split that produces the clustering with the highest overall similarity
4. Repeat steps 1, 2 and 3 until the desired number of clusters is reached

Different ways to choose the cluster which is to be split: (a) largest cluster, (b) the most heterogeneous cluster, (c) the largest cluster which has a predetermined degree of heterogeneity, ...

Fuzzy Clustering

- **Introduced by Zadeh** in 1969 to overcome the idea that all things can be absolutetely *True* or *False*. Zadeh was the first to claim that something can be 0.70 true
- According to fuzzy Algebra every element of the universe can belong to any **fuzzy set (FS)** with a **degree of membership** that varies from 0 to 1 taking real values
- If an element of the universe belongs to a FS with a degree of $\mu 1$ then it belongs to its complement with a degree $(1-\mu 1)$

Example of Fuzzy Set

$$A : \boldsymbol{\mu}_A = [\, \mu_A(\boldsymbol{x}_1)\ \ \mu_A(\boldsymbol{x}_2) \ldots \mu_A(\boldsymbol{x}_n)\,]$$

$$\mu_A(x_1) = .5 \quad \mu_A(x_2) = .8 \quad \mu_A(x_3) = .2$$

$$\mu_A(x_3) = .2 \quad \mu_A(x_4) = .3 \quad \mu_A(x_5) = .2$$

Example

Or equivalently

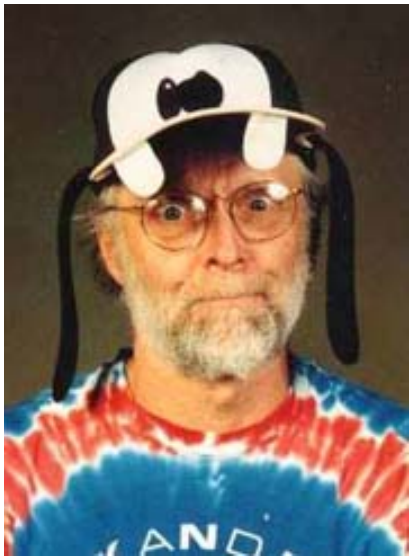
$$\boldsymbol{\mu}_A = [\, .5 \quad .8 \quad .2 \quad .3 \quad .2\,]$$

Fuzzy Clustering

- “linguistic” terms like “much”, “much more”, “less”, “more or less”, “more than” and others can use in fuzzy clustering
- Data points are given partial **degree of membership** in multiple nearby clusters
- Central point in the fuzzy clustering is always no unique partitioning of the data in a collection of clusters. In this membership value is assigned to each cluster. Sometimes this membership has been used to decide whether the data points belong to the cluster or not

Fuzzy C-Means Clustering (FCM)

- Useful in Fuzzy Modeling
  - Identification of the fuzzy rules needed to describe a "black box" system, on the basis of observed vectors of inputs and outputs
- Several approximations
  - FCM: Fuzzy C-Means Clustering (Bezdek, 1981)
  - PCM: Possibililistic C-Means Clustering (Krishnapuram - Keller, 1993)
  - FPCM: Fuzzy Possibililistic C-Means (N. Pal - K. Pal - Bezdek, 1997)



Prof. Bezdek

UNIVERSITAT DE BARCELONA

Fuzzy C-Means Clustering

- **Input**: Unlabeled data set
  - N is the number of data points in X, $X = \{x_1, x_2, \ldots, x_n\}$
  - $x_i \in \mathcal{R}^p$ where *p* is the number of features in each vector
- Main **output**:
  - A *c*-partition of *X*, which is *c x n* matrix *U*
- Additional **output**
  - Set of vectors $V = \{v_1, v_2, \ldots, v_c\} \subset \mathcal{R}^p$
  - v_i is called "cluster center"

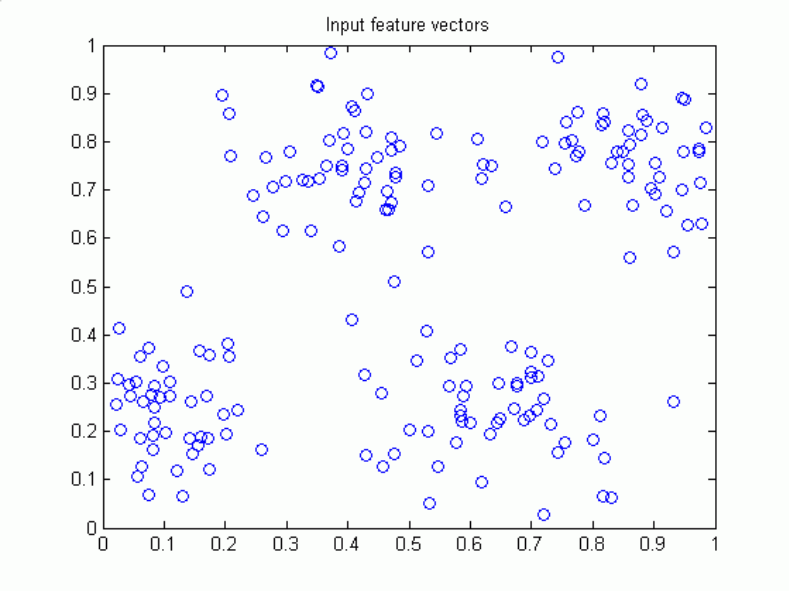
UNIVERSITAT DE BARCELONA

Fuzzy C-Means Clustering

p = 2

X

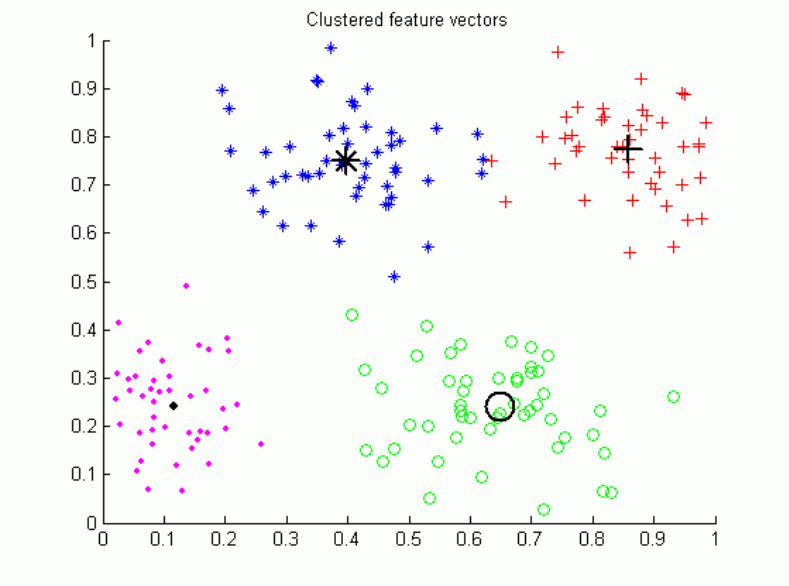
n = 188



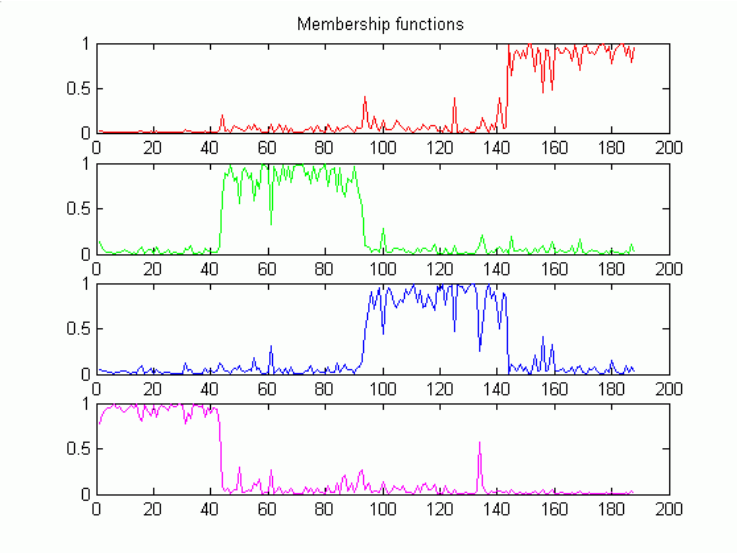
Sample Illustration

U

c = 4



Rows of U (Membership Functions)



Fuzzy C-Means Clustering

- **Goal**: Optimization of an "objective function" or "performance index"

$$J_m(U,V) = \sum_{k=1}^{n}\sum_{i=1}^{c}(U_{ik})^m \| x_k - v_i \|^2,$$

$$1 \le m \le \infty$$

- Constraint $\sum_{i=1}^{c} u_{ik} = 1, \forall k$

- Degree of fuzzification $m \geq 1$

Fuzzy C-Means Clustering

GOAL: Minimizing Objective Function

- Zeroing the gradient of $\mathcal{J}_m(U,V)$ with respect to *V*

$$U_{ik} = \left(\sum_{j=1}^{c} \left(\frac{\| x_k - v_i \|}{\| x_k - v_j \|} \right)^{2/(m-1)} \right)^{-1} \quad \forall i, \forall k \qquad \textbf{(Eq. 1)}$$

- Zeroing the gradient of $\mathcal{J}_m(U,V)$ with respect to *U*

$$v_i = \frac{\sum_{k=1}^{n} (U_{ik})^m x_k}{\sum_{k=1}^{n} (U_{ik})^m} \qquad \textbf{(Eq. 2)}$$

Note: It is the Center of Gravity

Fuzzy C-Means Clustering

Pick

- Initial Choices
  - Number of clusters , *1 < c < n*
  - Maximum number of iterations (Typ.: 100), *T*
  - Weighting exponent (Fuzziness degree), *m*
    - m=1: crisp
    - m=2: Typical
  - Termination measure $E_t = \| V_t - V_{t-1} \|$, ← 1-norm
  - Termination threshold (Typ. 0.01) (0 < ε)

Fuzzy C-Means Clustering

Iterative FCM algorithm

- Guess Initial Cluster Centers $V_0 = (V_{1,0}, \ldots, V_{c,0}) \in \mathcal{R}^{cp}$
- Alternating Optimization (AO)

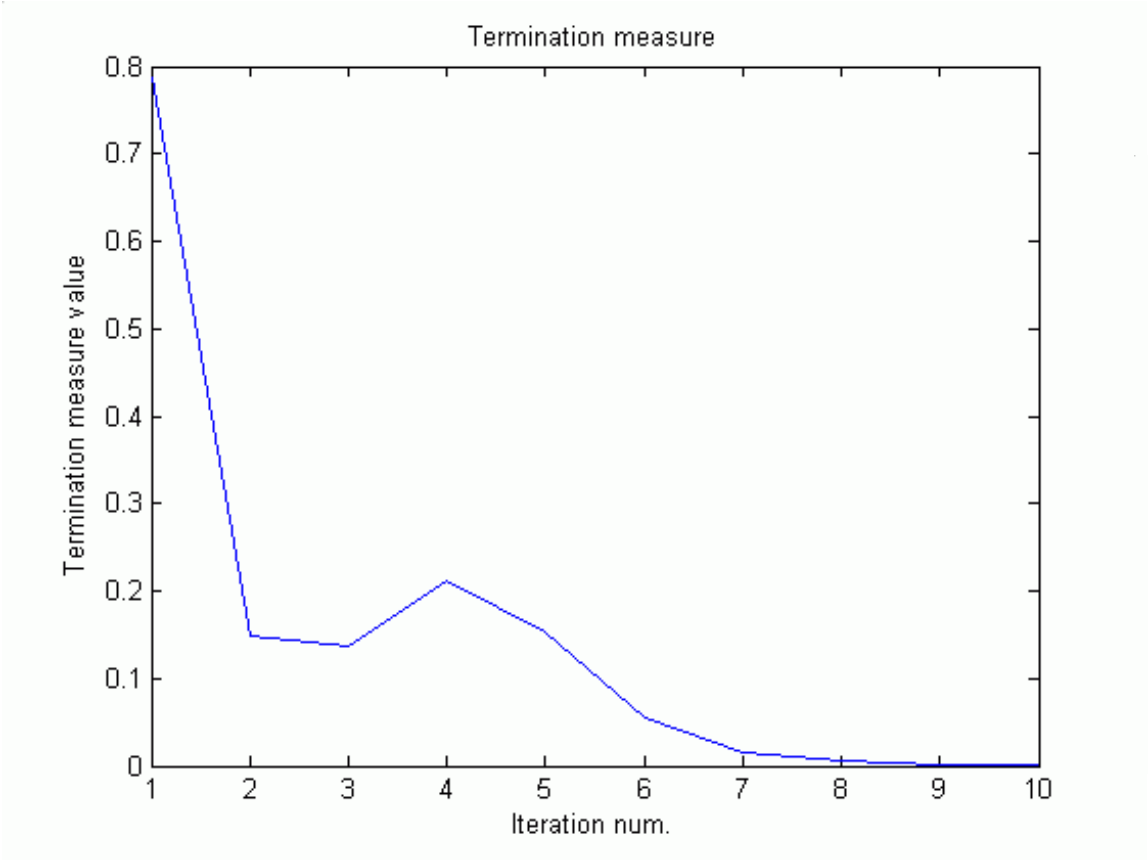
```
t ← 0
REPEAT
    t ← t +1
    Compute matrix U_t   (Eq.1)
    Compute associated clusters centers V_t (Eq.2)
UNTIL ( t = T   or ‖ V_t − V_{t−1} ‖ ≤ ε )
(U,V) ← (U_t,V_t)
```

Fuzzy C-Means Clustering

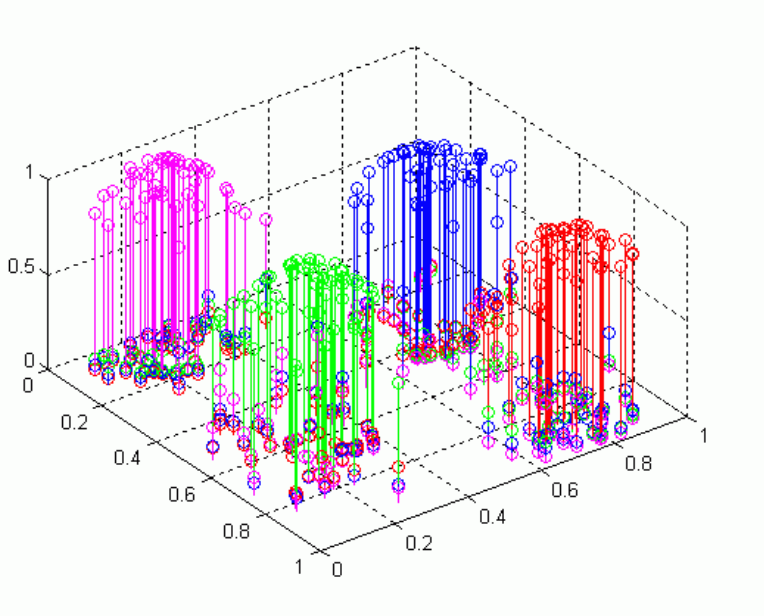
Sample Termination Measure Plot

m = 2.0

Termination Measure Values



Final Membership Degrees



Example Fuzzy clusters to Crisp clusters

Fuzzy Clusters ***(this is U matrix)***

$$
\begin{aligned}
Cl_1 &: [\ 0.6 \quad 0.7 \quad 0.3 \quad 0.1 \quad 0.4 \quad 0.2 \quad 0.1\ ] \\
Cl_2 &: [\ 0.1 \quad 0.1 \quad 0.3 \quad 0.5 \quad 0.1 \quad 0.7 \quad 0.1\ ] \\
Cl_3 &: [\ 0.3 \quad 0.2 \quad 0.4 \quad 0.4 \quad 0.5 \quad 0.1 \quad 0.8\ ]
\end{aligned}
$$

Example Fuzzy clusters to Crisp clusters

Fuzzy Clusters

Cl_1 : [0.6 0.7 0.3 0.1 0.4 0.2 0.1]
Cl_2 : [0.1 0.1 0.3 0.5 0.1 0.7 0.1]
Cl_3 : [0.3 0.2 0.4 0.4 0.5 0.1 0.8]

Crisp Clusters from Fuzzy Clusters

Cl_1 : [1 1 0 0 0 0 0]
Cl_2 : [0 0 0 1 0 1 0]
Cl_3 : [0 0 1 0 1 0 1]

Example Fuzzy clusters to Crisp clusters

Cl_1 : [1 1 0 0 0 0 0]
Cl_2 : [0 0 0 1 0 1 0]
Cl_3 : [0 0 1 0 1 0 1]

Cl_1 : { x_1 x_2 }
Cl_2 : { x_4 x_6 }
Cl_3 : { x_3 x_5 x_7 }

Crisp Clusters

Cl_1 : { x_1, x_2 }
Cl_2 : { x_4, x_6 }
Cl_3 : { x_3, x_5, x_7 }

Fuzzy C-Means Clustering

- **Advantages**
  - Unsupervised
  - Always converges
- **Disadvantages**
  - Long computational time
  - Sensitivity to the initial guess (speed, local minima)
  - Sensitivity to noise
    - One expects low (or even no) membership degree for outliers (noisy points)

Fuzzy C-Means Clustering

Optimal Number of Clusters

- Performance Index

$$\min_{(c)} \left\{ P(c) = \sum_{i=1}^{c} \sum_{k=1}^{n} u_{ik}^{m} (\|x_k - v_i\|^2 - \|v_i - \overline{x}\|^2) \right.$$

Average of all feature vectors $\overline{x} = \frac{1}{n} \sum_{k=1}^{n} x_k$

$$\sum_{i=1}^{c} \sum_{k=1}^{n} u_{ik}^{m} (\|x_k - v_i\|^2) \quad - \quad \sum_{i=1}^{c} \sum_{k=1}^{n} u_{ik}^{m} (\|v_i - \overline{x}\|^2)$$

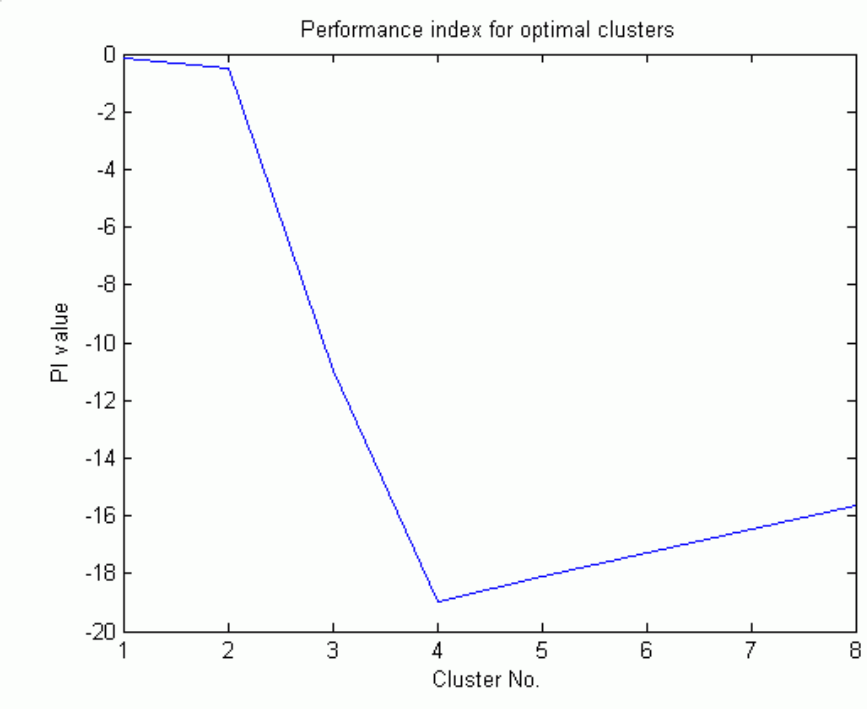
Sum of the within fuzzy cluster fluctuations (small value for optimal c)

Sum of the between fuzzy cluster fluctuations (big value for optimal c)

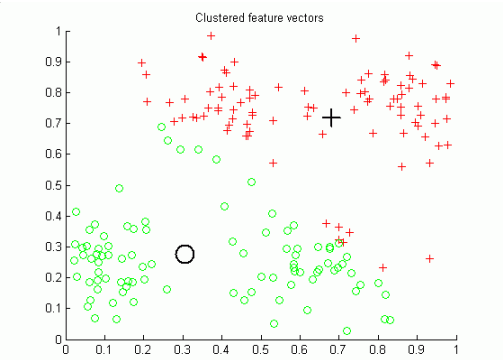
Fuzzy C-Means Clustering

Optimal Cluster No. (Example)

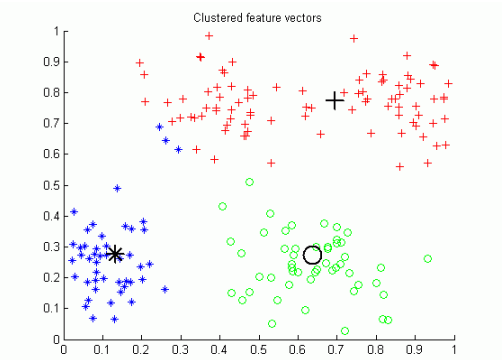
Performance index for optimal clusters (is minimum for c = 4)



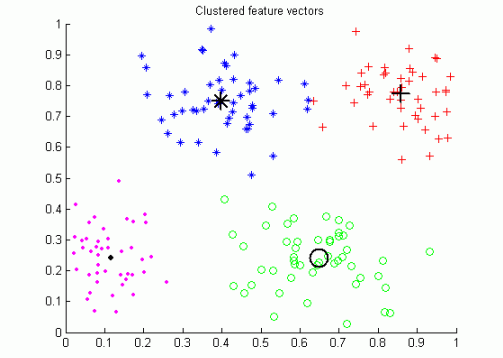
c = 2



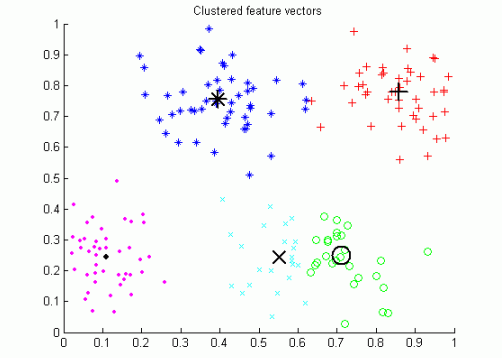
c = 3



c = 4



c = 5



UNIVERSITAT DE BARCELONA

References of FCM

- J. C. Dunn (1973): "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters", *Journal of Cybernetics 3: 32-57J.*
- ***C. Bezdek (1981): "Pattern Recognition with Fuzzy Objective Function Algoritms", Plenum Press, New York***
- **J. C. Bezdek, R. Ehrlich, W. Full (1984). FCM: The fuzzy c-Means Algorithm.**
- James C. Bezdek, James Keller, Raghu Krishnapuram and Nikhil R. Pal (1999), *Fuzzy Models and Algorithms for Pattern Recognition and Image Processing*, Kluwer Academic Publishers, TA 1650.F89.
- R. Krishnapuram and J. M. Keller (1993) A possibilistic approach to clustering," *IEEE Transactions on Fuzzy Systems*, Vol. 1, No. 2, pp. 98-110.
- N. R. Pal, K. Pal and J. C. Bezdek (1997), "A mixed c-means clustering model," *Proceedings of the Sixth IEEE International Conference on Fuzzy Systems*, Vol. 1, pp. 11-21.
- Jun Yan, Michael Ryan and James Power, *Using fuzzy logic Towards intelligent systems*, Prentice Hall, 1994.

Soft Clustering

- Clustering typically assumes that each instance is given a “hard” assignment to exactly one cluster
  - Does not allow uncertainty in class membership or for an instance to belong to more than one cluster

- ***Soft clustering*** **gives probabilities** that an instance belongs to each of a set of clusters
- Each instance is assigned a probability distribution across a set of discovered categories
  - probabilities of all categories must sum to 1

Expectation Maximization (EM)

- The EM algorithm was explained and given its name in a classic 1977 paper by Arthur Dempster, Nan Laird, and Donald Rubin
  - They pointed out that the method had been "proposed in many times in special circumstances" by earlier authors.
- EM is typically used to compute maximum likelihood estimates given incomplete samples

Applications of EM

- Filling in missing data in samples
- Discovering the value of latent variables
- Estimating the parameters of HMMs
- Estimating parameters of finite mixtures
- **Unsupervised learning of clusters**
- **Semi-supervised classification and clustering**

Expectation Maximization

- **Main idea:** Use probabilities instead of distances
- **Goal**:
  - Find the most likely clusters given the data
  - Determine the probability with which an object belongs to a certain cluster

$$\Pr(C \mid x) = \frac{\Pr(x \mid C)\Pr(C)}{\Pr(x)} = \frac{\Pr(x \mid C)\Pr(C)}{\sum_C \Pr(C)\Pr(x \mid C)}$$

where *Pr(C)* is the probability that a randomly selected object belongs to cluster *C*, and *Pr(x|C)* is the probability of observing the object *x* given the cluster *C*

  - Let us assume that we know that there are *k* clusters
  - To learn the clusters, we need to determine their parameters (means and standard deviations)

Expectation Maximization (EM)

- Probabilistic method for soft clustering
  - It uses probabilities instead of distances!
- Direct method that assumes *k* clusters: $\{c_1, c_2, \ldots c_k\}$
- Soft version of *k*-means
- Assumes a probabilistic model of categories that allows computing P(C | x) for each category or cluster, C, for a given example, x
- For text, typically assume a naïve-Bayes category model

EM Algorithm

- **Iterative method** for learning probabilistic categorization model from unsupervised data.
- Initially assume random assignment of examples to categories.
- Learn an initial probabilistic model by estimating model parameters θ from this randomly labeled data.
- **Algorithm**: iterate following two steps until convergence:
  - ***Expectation*** *(E-step):* Compute P(C | x) for each example given the current model, and probabilistically re-label the examples based on these posterior probability estimates.
  - ***Maximization*** *(M-step):* Re-estimate the model parameters, θ, from the probabilistically re-labeled data.

EM algorithm (cont.)

1. Calculate cluster probability (represented as object weights) for each object (**Estimation step**)
2. Estimate distribution parameters based on the cluster probabilities (**Maximization step**)
3. Procedure stops when log-likelihood saturates

Log-likelihood: $\Sigma_i log(p_A Pr(x_i|A) + p_B Pr(x_i|B))$

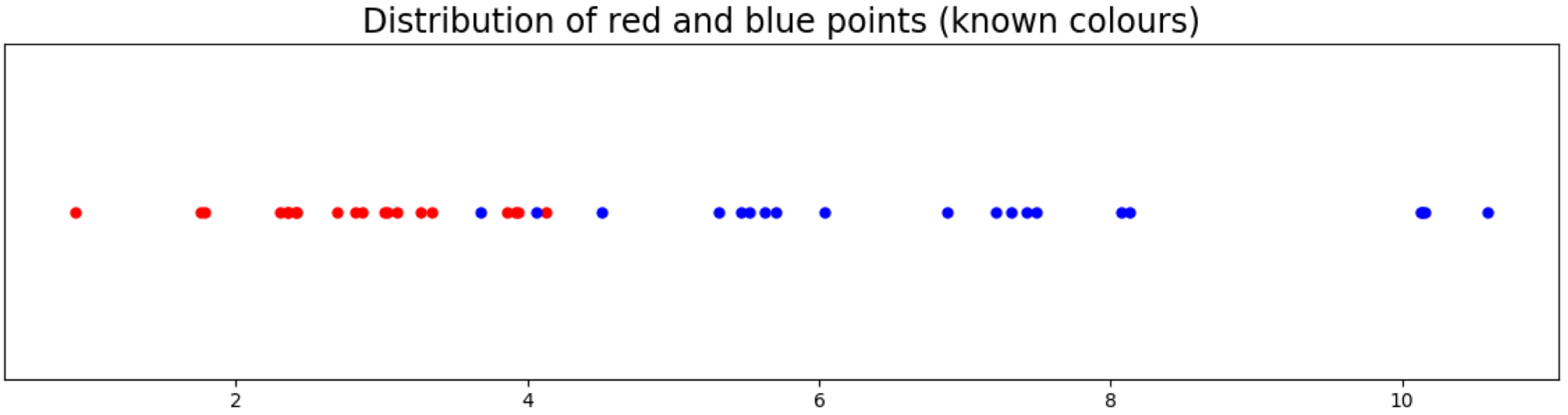
Estimating parameters from weighted objects:

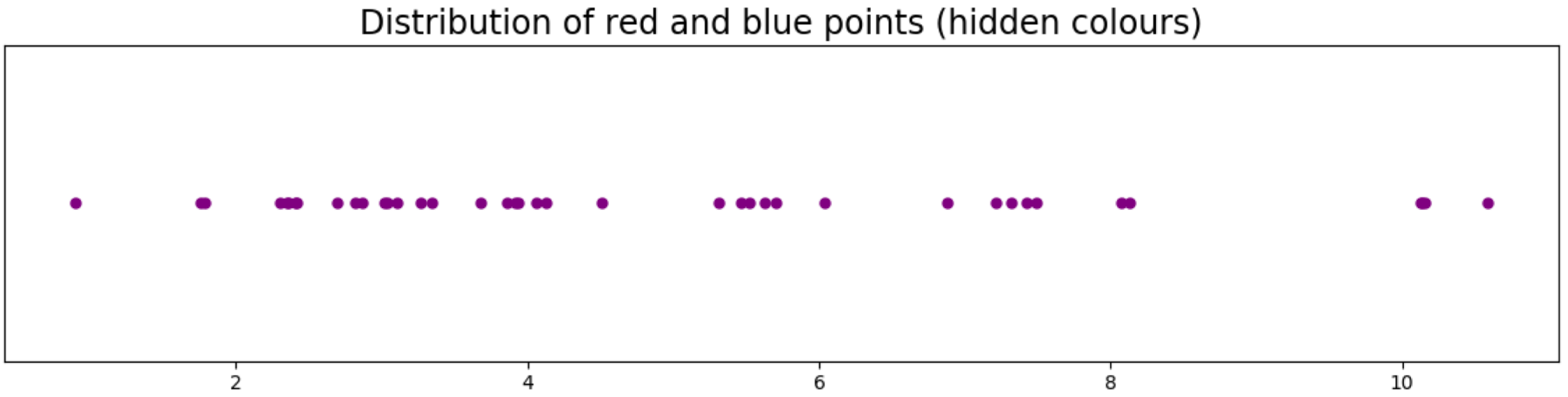
Mean of cluster A: $\mu_A = \frac{w_1 x_1 + w_2 x_2 + ... + w_n x_n}{w_1 + w_2 + ... + w_n}$

Standard deviation: $\sigma_A^2 = \frac{w_1 (x_1 - \mu)^2 + w_2 (x_1 2 - \mu)^2 + ... + w_n (x_n - \mu)^2}{w_1 + w_2 + ... + w_n}$

UNIVERSITAT DE BARCELONA

EM example

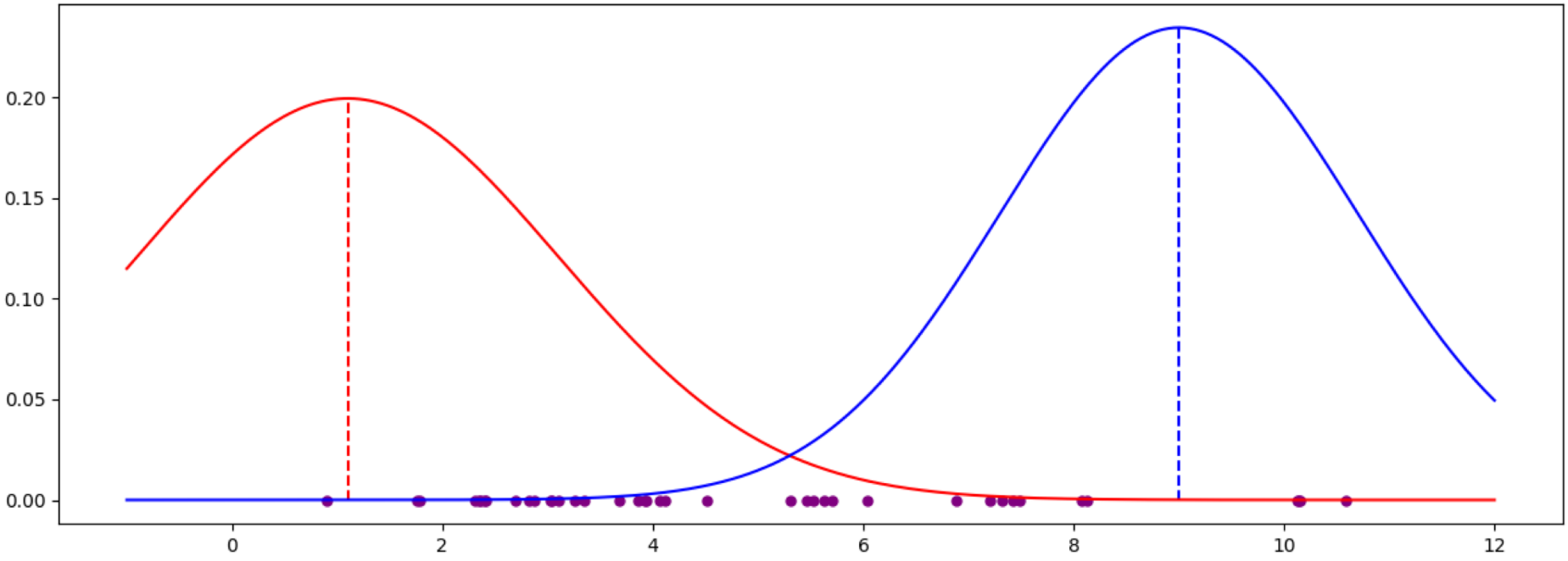




UNIVERSITAT DE BARCELONA

EM example

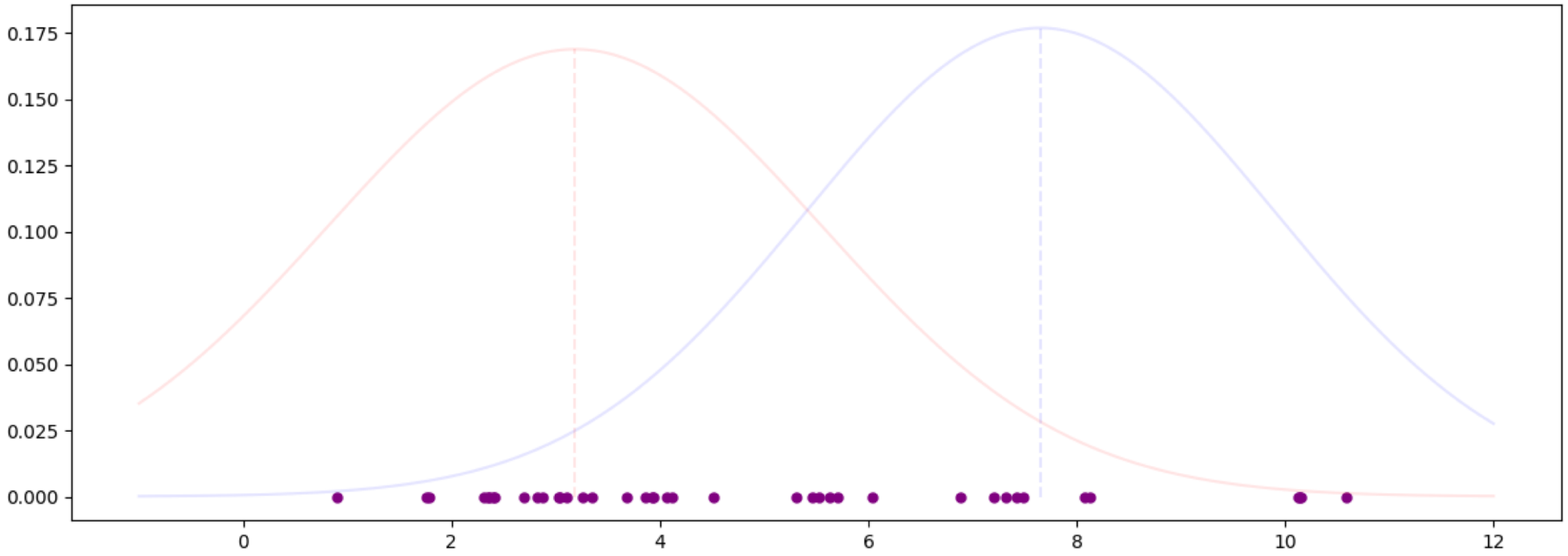
- We know that we have two groups of points, each drawn from a normal distribution
- We also have a likelihood function and we would like to find values for the mean and standard deviation that maximise this function (maximum likelihood estimation)



EM example

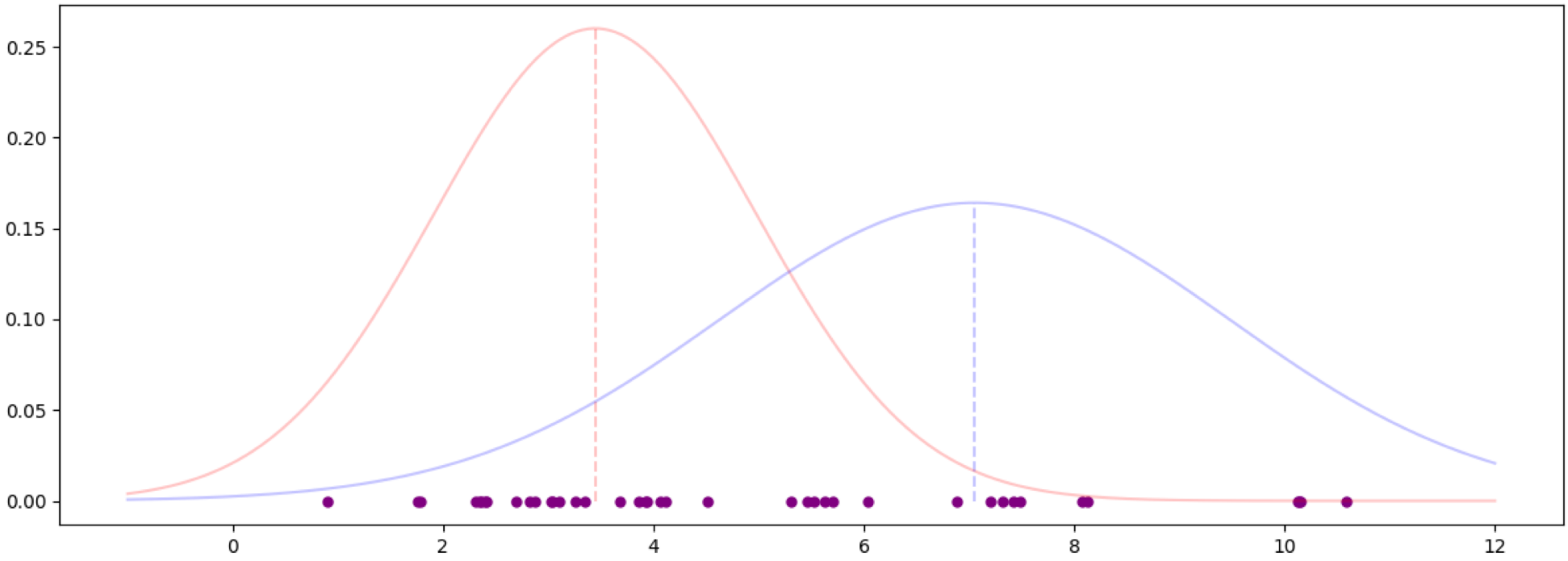
1. Start with initial estimates of the mean and standard deviation for the red and blue groups
2. Check how likely each (mean, standard deviation) estimate is to produce each sequence each of the data points (using the likelihood function)
3. Produce a weighting for each (mean, stand deviation) pair for each data point (the **Expectation step**)
    1. These weights will allow us to "rescale" the data points along the axis
4. Use formulae to compute new maximum likelihood estimates of each parameter based on the rescaled data points (the **Maximisation step**)
5. Repeat steps 2-4 until each parameter estimate has converged, or a set number of iterations has been reached

EM example

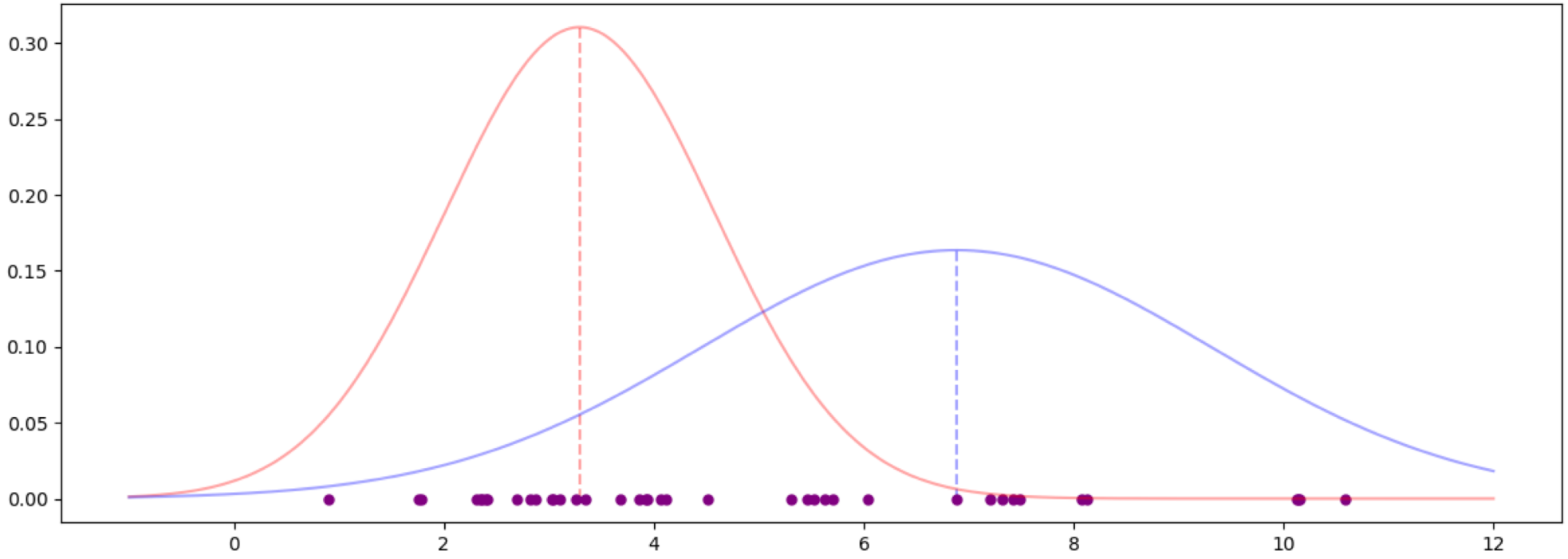


UNIVERSITAT DE BARCELONA

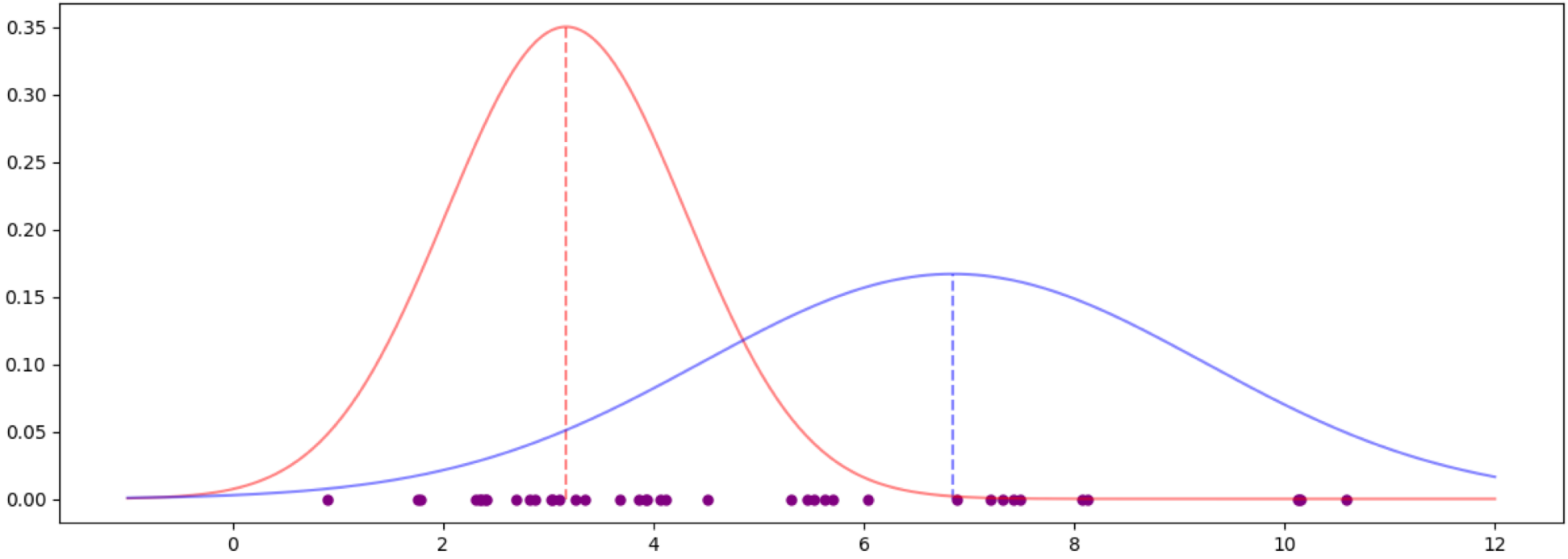
EM example



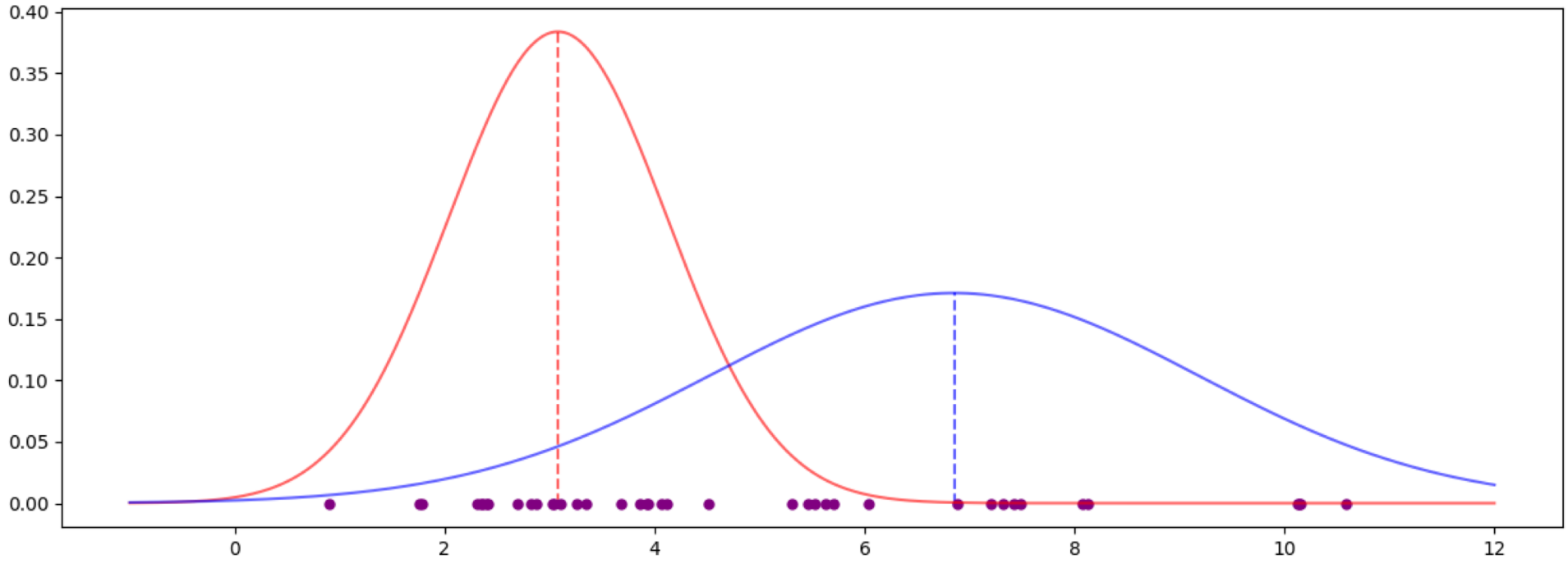
EM example



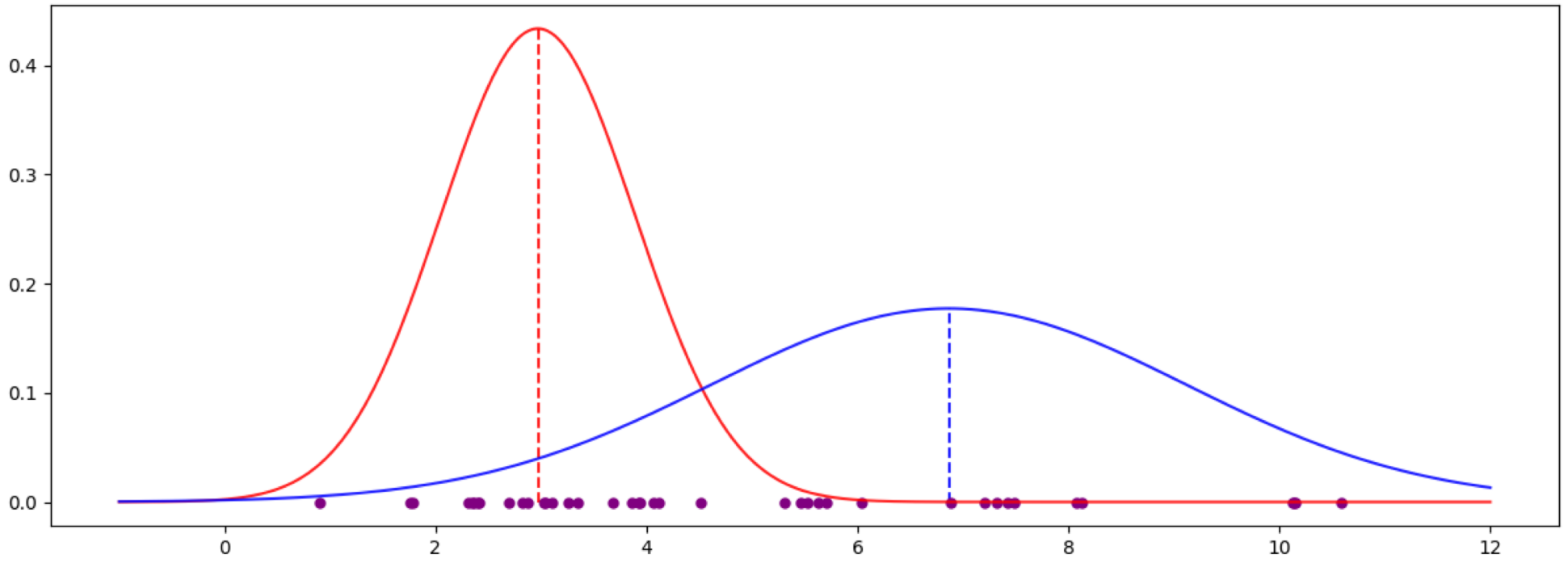
EM example



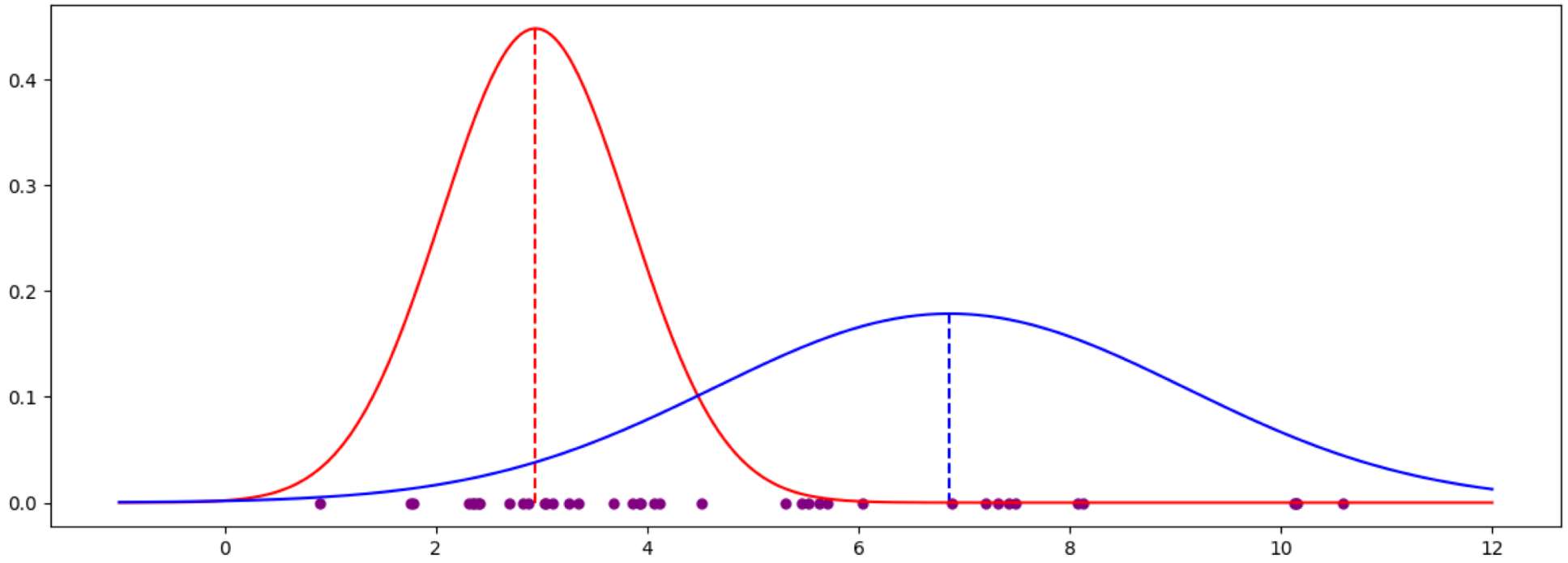
EM example



EM example

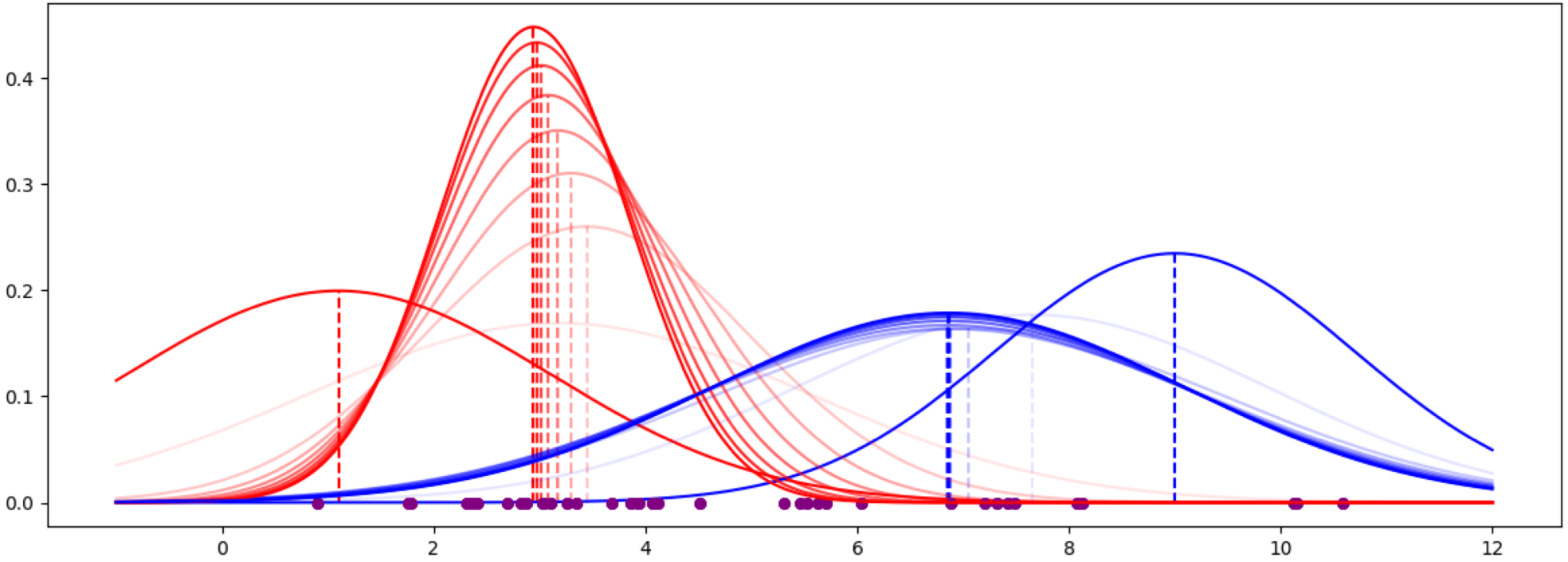


EM example



UNIVERSITAT DE BARCELONA

EM example



UNIVERSITAT DE BARCELONA

Clustering Validation

How many clusters?

Six Clusters

Two Clusters

Four Clusters

Which is the best clustering?

Cluster validation

Supervised classification:

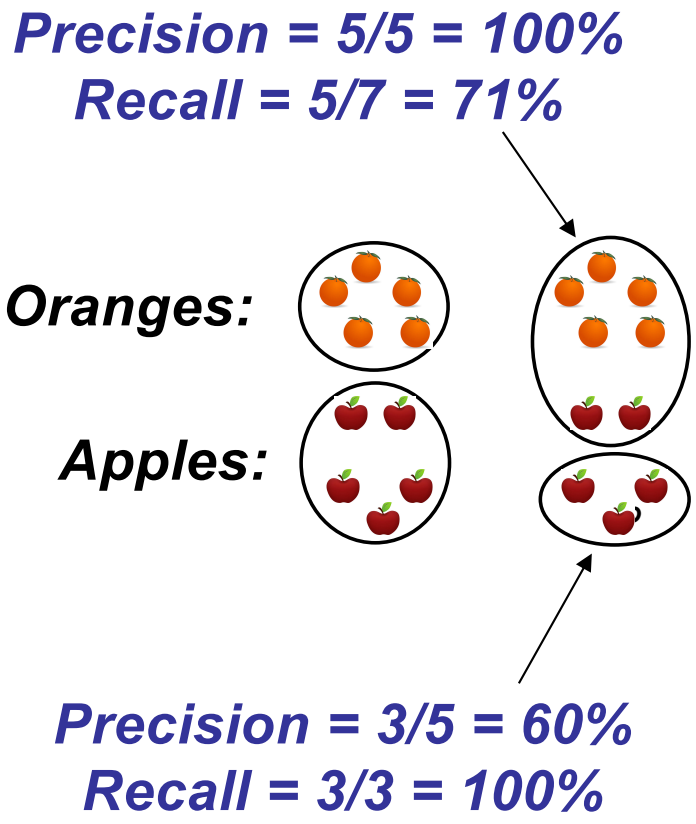
- Class labels known for ground truth
- Accuracy, precision, recall

Cluster analysis

- No class labels

Validation need to:

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



UNIVERSITAT DE BARCELONA

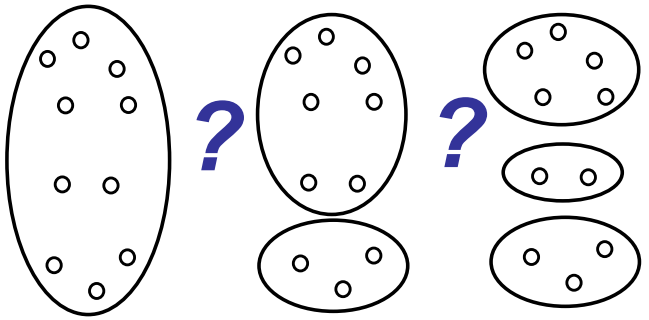
What Is A Good Clustering?

- **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

Measuring clustering validity

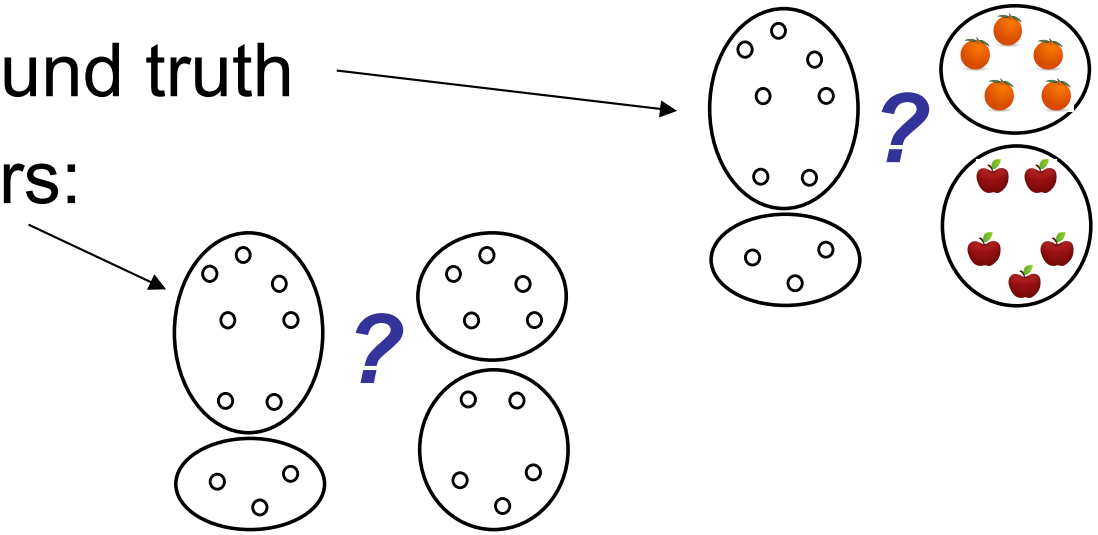
Internal Index

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



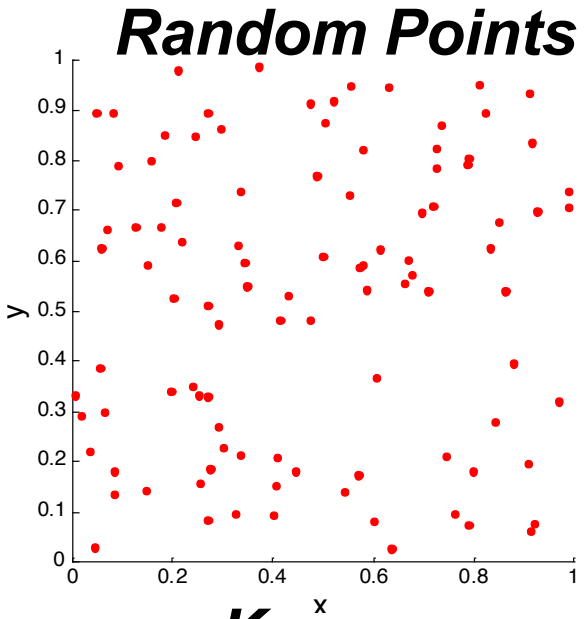
External Index

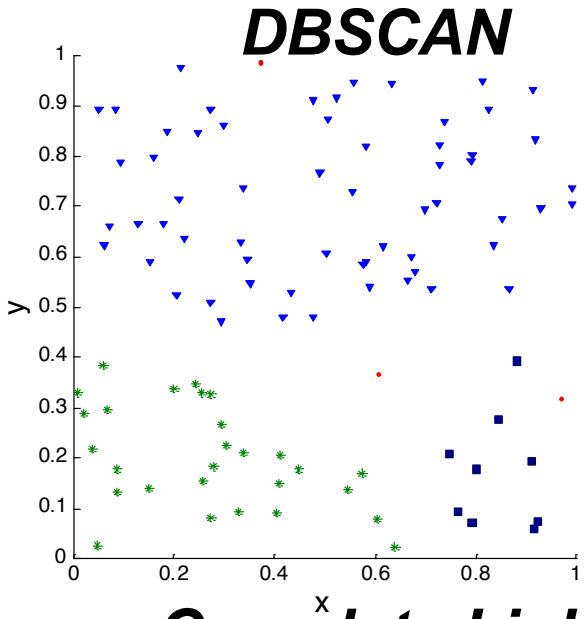
- Validate against ground truth
- Compare two clusters: (how similar)

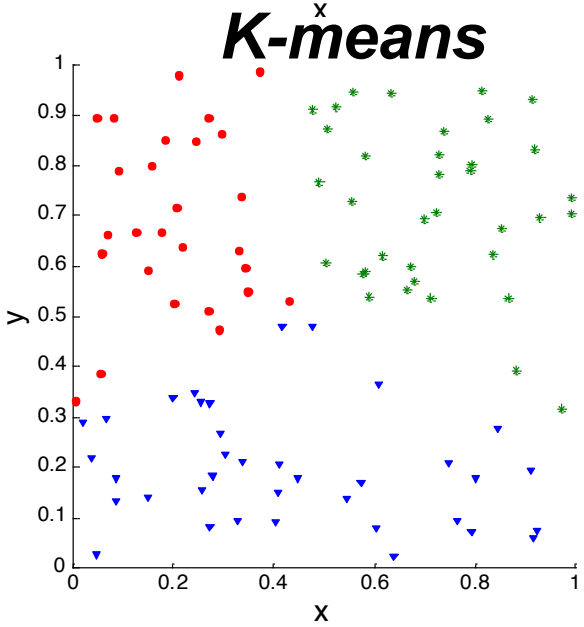


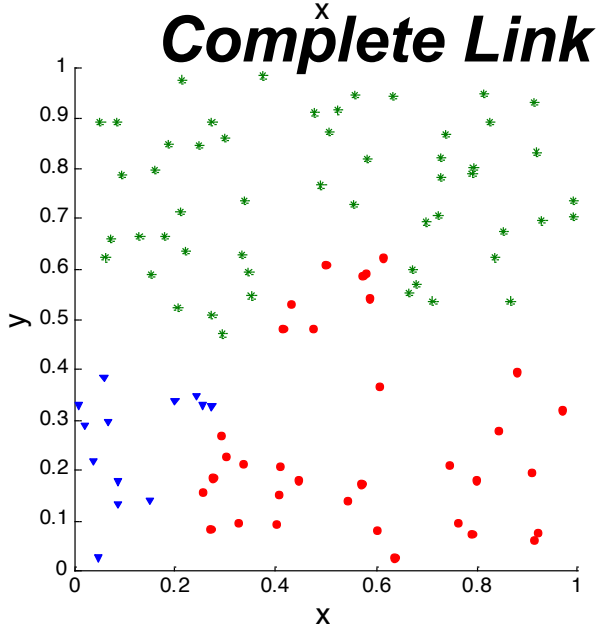
UNIVERSITAT DE BARCELONA

Clustering of random data



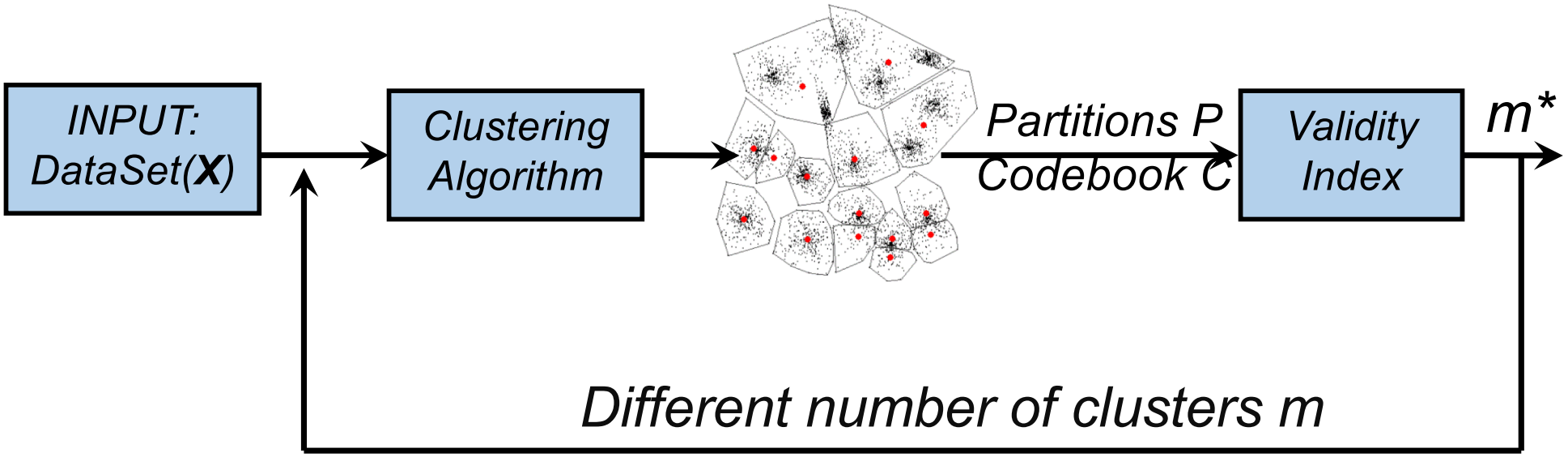






Cluster validation process

- **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!



UNIVERSITAT DE BARCELONA

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

Internal indexes

Silhouette Coefficients	$a(x_i) = \frac{1}{n_m - 1} \sum_{j=1, j \neq i}^{n_m} \|x_i - x_j\|^2_{x_i, x_j \in C_m}$ $b(x_i) = \min_t \{\frac{1}{n_t} \sum_{j \in C_t} \|x_i - x_j\|^2\}_{x_i \notin C_t}$ $s(x_i) = \frac{b(x_i) - a(x_i)}{\max(a(x_i), b(x_i))}$ $SC = \frac{1}{N} \sum_{i=1}^{N} s(x_i)$ $b(x_i) = \min \{\sum_{t \neq m} \|C_t - C_m\|^2\}_{x_i \notin C_t} (SC'2008)$
RMSSTD	$RMSSTD = \frac{\sum_{\substack{k=1,\ldots,M \\ d=1,\ldots,D}} \sum_{i=1}^{n_{kd}} (x_i - \overline{x^d})^2}{\sum_{\substack{k=1,\ldots,M \\ d=1,\ldots,D}} (n_{kd} - 1)}$
R-square	$RS = \frac{SST - SSW}{SST} = \frac{\sum_{d \ 1,\ldots,D} \sum_{i=1}^{n_d} (x_i - \overline{x^d})^2 - \sum_{\substack{k=1,\ldots,M \\ d=1,\ldots,D}} \sum_{i=1}^{n_{kd}} (x_i - \overline{x^d})^2}{\sum_{d=1,\ldots,D} \sum_{i-1}^{n_d} (x_i - \overline{x^d})^2}$
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	$PE = -(\sum_{i=1}^{N} \sum_{k=1}^{M} u_{ik} \log(u_{ik}))/N$

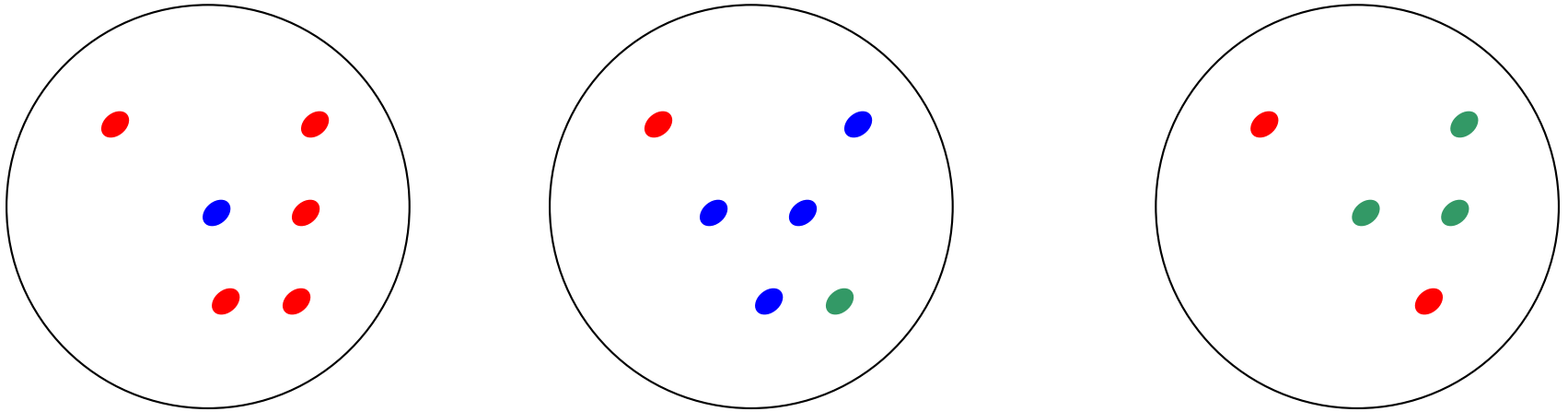
Soft partitions

External Indexing

- Assesses clustering with respect to ground truth
- Assume that there are C gold standard classes, while our clustering algorithms produce *k* clusters, π_1, π_2, …, π_k with n_i members.
- **Simple measure**: purity, the ratio between the dominant class in the cluster π_i and the size of cluster π_i

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

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)

UNIVERSITAT DE BARCELONA

Pair-counting measures

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*.

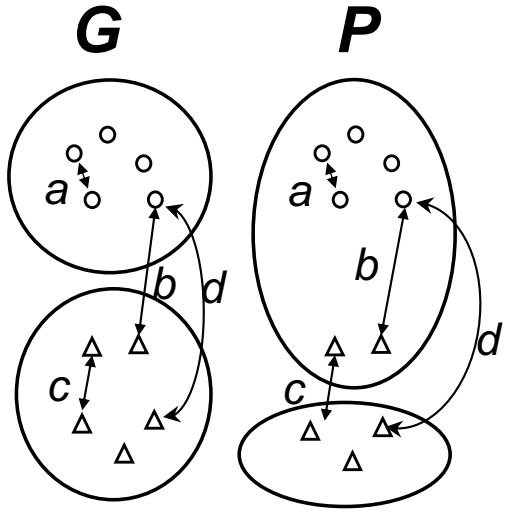
$$b = \frac{1}{2}\left(\sum_{j=1}^{K'} n_{.j}^2 - \sum_{i=1}^{K}\sum_{j=1}^{K'} n_{ij}^2\right)$$

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

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

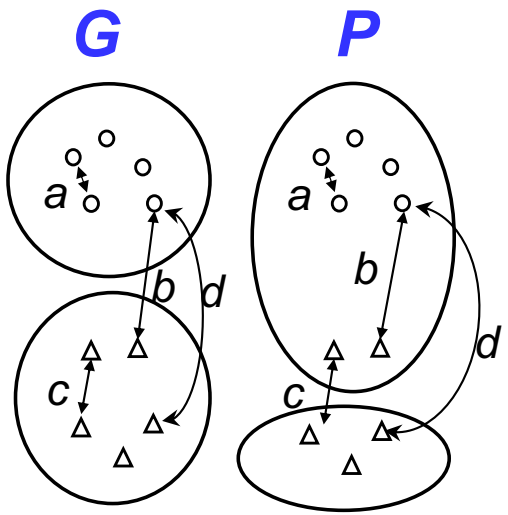
Different classes **both** in *P* and *G*.

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



Rand and Adjusted Rand index

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



Agreement: a, d
Disagreement: b, c

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

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

External indexes

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

$$\begin{array}{|c|c|} \hline N & \cdot \\ \hline \cdot & n_{..} \\ \hline \end{array} = \begin{array}{|cccc|c|} \hline n_{11} & n_{12} & \cdots & n_{1l} & n_{1.} \\ n_{21} & n_{22} & \cdots & n_{2l} & n_{2.} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ n_{k1} & n_{k2} & \cdots & n_{kl} & n_{k.} \\ \hline n_{.1} & n_{.2} & \cdots & n_{.l} & n_{..} \\ \hline \end{array}$$

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

External indexes

- 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

UNIVERSITAT DE BARCELONA

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	$\sum_i p_i(\max_j \frac{p_{ij}}{p_i})$	(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	$[\binom{n}{2} - \sum_i \binom{n_{i\cdot}}{2} - \sum_j \binom{n_{\cdot j}}{2} + 2\sum_{ij}\binom{n_{ij}}{2}]/\binom{n}{2}$	(0,1]
7	Jaccard coefficient	J	$\sum_{ij}\binom{n_{ij}}{2}/[\sum_i \binom{n_{i\cdot}}{2} + \sum_j \binom{n_{\cdot j}}{2} - \sum_{ij}\binom{n_{ij}}{2}]$	[0,1]
8	Fowlkes and Mallows index	FM	$\sum_{ij}\binom{n_{ij}}{2}/\sqrt{\sum_i \binom{n_{i\cdot}}{2}\sum_j \binom{n_{\cdot j}}{2}}$	[0,1]
9	Hubert Γ statistic I	Γ	$\frac{\binom{n}{2}\sum_{ij}\binom{n_{ij}}{2} - \sum_i \binom{n_{i\cdot}}{2}\sum_j \binom{n_{\cdot j}}{2}}{\sqrt{\sum_i \binom{n_{i\cdot}}{2}\sum_j \binom{n_{\cdot j}}{2}[\binom{n}{2} - \sum_i \binom{n_{i\cdot}}{2}][\binom{n}{2} - \sum_j \binom{n_{\cdot j}}{2}]}}$	(-1,1]
10	Hubert Γ statistic II	Γ'	$[\binom{n}{2} - 2\sum_i \binom{n_{i\cdot}}{2} - 2\sum_j \binom{n_{\cdot j}}{2} + 4\sum_{ij}\binom{n_{ij}}{2}]/\binom{n}{2}$	[0,1]
11	Minkowski score	MS	$\sqrt{\sum_i \binom{n_{i\cdot}}{2} + \sum_j \binom{n_{\cdot j}}{2} - 2\sum_{ij}\binom{n_{ij}}{2}}/\sqrt{\sum_j \binom{n_{\cdot j}}{2}}$	$[0, +\infty)$
12	classification error	ε	$1 - \frac{1}{n}\max_\sigma \sum_j n_{\sigma(j),j}$	[0,1)
13	van Dongen criterion	VD	$(2n - \sum_i \max_j n_{ij} - \sum_j \max_i n_{ij})/2n$	[0, 1)
14	micro-average precision	MAP	$\sum_i p_i(\max_j \frac{p_{ij}}{p_i})$	(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	$\sum_i n_{i\cdot}^2 + \sum_j n_{\cdot j}^2 - 2\sum_i\sum_j n_{ij}^2$	$[0, 2\binom{n}{2})$

Note: $p_{ij} = n_{ij}/n,\ p_i = n_{i\cdot}/n,\ p_j = n_{\cdot j}/n.$

UNIVERSITAT DE BARCELONA

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.

UNIVERSITAT DE BARCELONA

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.

UNIVERSITAT DE BARCELONA

References on cluster validation

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.

References on cluster validation

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)

Conclusions

- Unsupervised learning induces categories from unlabeled data.
- There are a variety of approaches, including:
  - HAC
  - k-means, FCM
  - EM
- Semi-supervised learning uses both labeled and unlabeled data to improve results.

Reading proposal

- Rui Xu and D. Wunsch, II. 2005. "**Survey of clustering algorithms**", IEEE TRANSACTIONS ON NEURAL NETWORKS, VOL. 16, NO. 3, MAY 2005, *pages 645-678.*

DOI=10.1109/TNN.2005.845141
http://dx.doi.org/10.1109/TNN.2005.845141