Introduction

- Data analysis:
  - Exploratory
    - Hypothesis creation
  - Confirmatory
    - Decision-making

Data Analysis

- Grouping of data:
  - Is there a correlation between data patterns?
  - Which data patterns are similar?
    - Which words are similar?
    - What kind of constructions are similar?

Cluster Analysis

- Tryon [3]
  - Unsupervised classification of observed data into groups (clusters).
  - Use:
    * No a priori hypothesis.
    * Grouping of Objects or Individuals.
    * Grouping of Variables.

Application of Clustering

- Wide area e. g.:
  - medicine
  - chemistry
  - psychiatry
  - linguistics
- Development of taxonomies.
- Dissection of a population.

Clustering Objectives

- Everitt [1, 3-4]
  - Typology detection or identification.
  - Model Fitting.
  - Prediction based on groups.
  - Hypothesis testing.
  - Data exploration.
  - Hypothesis generating.
  - Data reduction.

Names for Cluster Analysis

- Different names used in the literature:
  - Q-analysis
  - Typology
  - Grouping
  - Clumping
  - Numerical taxonomy
  - Unsupervised pattern recognition

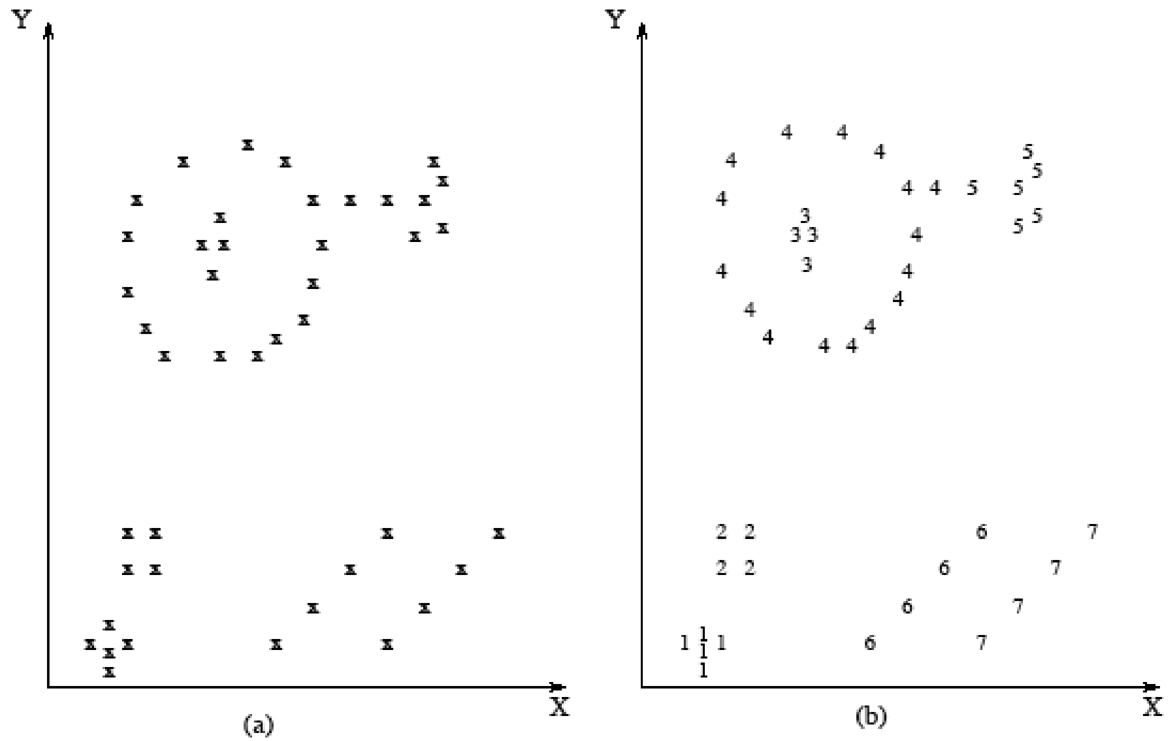
Clustering vs. Classification

- Classification:
  - Grouping on the basis of a priori labels
  - Discriminant analysis = supervised classification
  - Given a set of labeled patterns, label an unlabeled pattern

Clustering

- Labeling of unlabeled data sets or patterns
  - *Data-driven*, not taxonomy driven = unsupervised
  - Labels are related to clusters
  - Cluster labels are obtained solely from data

Clustering



Jain et al. [2]

Prerequisites for Clustering

- Representation of data (pattern and features)
- Data or pattern proximity measure (domain dependent)
- Clustering algorithm

Data Representation for Clustering

- Representation of data: pattern and features
  - Number of classes
  - Available and expected patterns
  - Features: number, type, scale
- May partially be opaque or unknown

Features for Clustering

- Feature selection
  - Identification of the subset of features that is most efficient for clustering.
- Feature extraction
  - Transformation of input features and creation of new salient features.

Clustering Process

- Input: Data selection and preparation
- Input: Feature selection and/or extraction
- Evaluation: Proximity measures
- Evaluation: Clustering algorithm
- Output: Taxonomy, Grouping, Clusters

Proximity Measures for Clustering

- The choice of pattern proximity measures is:
  - Domain or data dependent
  - Distance function defined on pairs of patterns
    - e. g. Euclidean distance etc.

Clustering

- Grouping
  - Hierarchical algorithms with nested groups
  - Overlapping groups
  - etc.

Data Abstraction for Clustering

- Extraction of data sets that are:
  - simple
  - compact
- Machine oriented: efficiency
- Human oriented: intuitive and comprehensible

Clustering Evaluation

- Pre-clustering evaluation: Cluster tendency
- Post-clustering evaluation: Cluster validity
  - Rather subjective
  - Valid: if clusters are not the result of an artifact or randomly chosen.

Clustering

- Evaluation: Cluster validity
  - External assessment:
    * Compare recovered structure to some a priori structure
    * Automatically compare taxonomies,
      hierarchical trees,
      distance of centroids etc.

Clustering

- Evaluation: Cluster validity
  - Internal assessment:
    * Are resulting clusters intrinsically appropriate for the data.

Clustering

- Evaluation: Cluster validity
  - Relative test:
    * Compare two resulting clusters and measure relative merit.

Clustering

- Clustering algorithms
  - Vast number
  - Selection on the basis of:
    * Way in forming clusters
    * Data-structure
    * Robustness (changes, data types)

Clustering

- Further criteria
  - Data normalization
  - Choice of similarity measure
  - Data amount (small, large)
  - Use of domain knowledge or heuristics

Clustering

- Types of algorithms and techniques:
  - Hierarchical
  - Optimization
  - Density or mode-seeking
  - Clumping
  - K-means Clustering
  - Expectation Maximization (EM)

Clustering

- Formalization:
  - Feature Vector, Datum, Pattern:
    With d measurements: $\mathbf{x} = (x_1, x_2, ..., x_d)$
  - $x_1, x_2, \ldots,$ in general: x_i is a *feature* or *attribute* of $\mathbf{x}$
  - $d =$ *dimension* of pattern or pattern space

Clustering

- Formalization:
  - Pattern set: $\mathscr{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$
  - The i^{th} pattern in $\mathscr{X}$: $\mathbf{x}_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,d})$
  - or

Feature Matrix for Clustering

$$\mathscr{X} = \begin{bmatrix} \mathbf{x}_{1,1} & \mathbf{x}_{1,2} & \cdots & \mathbf{x}_{1,d} \\ \mathbf{x}_{2,1} & \mathbf{x}_{2,2} & \cdots & \mathbf{x}_{2,d} \\ \vdots & & & \\ \mathbf{x}_{k,1} & \mathbf{x}_{k,2} & \cdots & \mathbf{x}_{k,d} \end{bmatrix}$$

Clustering

- Class:
  - Refers to the state of nature that governs the pattern generation process.
  - Clustering techniques group patterns to classes.

Clustering

- Hard clustering techniques:
  - Assign a label l_i to each pattern $\mathbf{x}_i$ identifying its class.
  - For a set of patterns $\mathscr{X}$ the set of labels is $\mathscr{L} = \{l_1, l_2, \ldots, l_n\}$ with $l_i \in \{1, \ldots, k\}$, with k the number of clusters

Clustering

- Fuzzy clustering:
  - Assign each pattern $\mathbf{x}_i$ a fractional degree of membership f_{ij} in each output cluster j.

Clustering

- Distance measure:
  - Specialization of a proximity measure
  - Metric on the feature space for quantifying the similarity of patterns.

Clustering

- Pattern and feature selection:
  - No theoretical guidelines
  - Depending on experiment, data, user
  - Deep understanding of features and possible transformations can lead to better results in clustering.

Clustering

- Objects:
  - Physical object (e. g. door)
  - Abstract notion (e. g. language style)
  - Representation:
    - Multidimensional vectors
    - Each dimension is a feature
    - Features are: quantitative or qualitative

Clustering

- Quantitative features:
  - Continuous values (e. g. length)
  - Discrete values (e. g. number of vowels)
  - Interval values (e. g. duration of vowels)

Clustering

- Qualitative features:
  - Nominal or unordered (e. g. lexical or morpho-syntactic category)
  - Ordinal (e. g. sound intensity "quiet"–"loud"; speed "slow"–"fast")

Clustering

- Structured features:
  - Tree structure (e. g. ontology or thesaurus)
  - Mapping structured features to linked values and features

$\rightarrow$ *symbolic objects*

Clustering

- Strategy:
  - Isolate most descriptive and discriminatory features
  - → Feature selection
  - → Feature extraction
  - Goals
    - improve classification performance
    - improve computational efficiency

Clustering

- Similarity measures:
  - Essential to most clustering techniques
  - Most common calculation:
    - *Dissimilarity*
    - For continuous features: *Euclidean distance*

$$d_2(\mathbf{x}_i, \mathbf{x}_j) = \left(\sum_{k=1}^{d} (x_{i,k} - x_{j,k})^2\right)^{1/2} = \|\mathbf{x}_i - \mathbf{x}_j\|_2$$

Clustering

- *Euclidean distance*
  - Proximity evaluation in 2D or 3D space
  - Good for compact or isolated clusters
  - Tendency of largest-scaled feature to dominate others
    - Solution: normalization

Clustering

- *Mahalanobis Metric*

$$d_M(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j) C_x^{-1} (\mathbf{x}_i - \mathbf{x}_j)^T$$

Clustering

- Covariance:
  - variance = average of the squared deviation of a feature from its mean
  - covariance = average of the products of the deviations of feature values from their means

Clustering

- Covariance of two features
  - Measures their tendency to vary together, i. e. co-vary.
  - Variance is the average of the squared deviation of a feature from its mean.
  - Covariance is the average of the products of the deviations of feature values from their means.

Clustering

- Covariance of two features
  - Feature i and Feature j:
    - Let $\{x_{1,i}, x_{2,i}, \ldots, x_{n,i}\}$ be a set of n examples of Feature i,
    - Let $\{x_{1,j}, x_{2,j}, \ldots, x_{n,j}\}$ be a corresponding set of n examples of Feature j
    - $x_{k,i}$ and $x_{k,j}$ are features of the same pattern k

Clustering

- Covariance of two features
  - Let m_i be the mean of Feature i, and m_j be the mean of Feature j
  - Then the covariance $c_{i,j}$ of Feature i and Feature j is:

$$\{[x_{1,i}-m_i][x_{1,j}-m_j]+\ldots+[x_{n,i}-m_i][x_{n,j}-m_j]\}/(n-1)$$

Clustering

- Covariance matrix
  - Collection of all covariances in covariance matrix C:

$$C = \begin{bmatrix} \mathbf{c}_{1,1} & \mathbf{c}_{1,2} & \cdots & \mathbf{c}_{1,d} \\ \mathbf{c}_{2,1} & \mathbf{c}_{2,2} & \cdots & \mathbf{c}_{2,d} \\ \vdots & & & \\ \mathbf{c}_{d,1} & \mathbf{c}_{d,2} & \cdots & \mathbf{c}_{d,d} \end{bmatrix}$$

Clustering

- Covariance properties
  - If Feature i and Feature j tend to increase together, then $c_{i,j} > 0$
  - If Feature i tends to decrease when Feature j increases, then $c_{i,j} < 0$
  - If Feature i and Feature j are independent, then $c_{i,j} = 0$

Clustering

- Covariance properties
  - $|c_{i,j}| <= s_i s_j$, where s_i is the standard deviation of Feature i
  - $c_{i,i} = s_i^2 = v_i$

Clustering

- Covariance properties
  - Covariance $c_{i,j}$ is a number between $-s_i s_j$ and $+s_i s_j$ that measures the dependence between Feature i and Feature j, with $c_{i,j} = 0$ if there is no dependence.

Clustering

- *Mahalanobis Metric*
  - With uncorrelated features and same variance in all directions this corresponds to *Euclidean distance*.
  - Automatically accounts for scaling of the coordinate axes.
  - Corrects for correlation between different features.

Clustering

- *Mahalanobis Metric*
  - Problems:
    - Potentially hard to determine covariance matrices accurately
    - Memory and time requirements grow quadratically rather than linearly with the number of features, significant when the number of features becomes large.

Distance Matrix for Clustering

- Store distance between all vectors in a matrix
- Distance is commutative: $D(x_i, x_j) = D(x_j, x_i)$
- Distance $D(x_i, x_i) = 0$

Distance Matrix for Clustering

x1					
x2	4				
x3	3	6			
x4	5	7	9		
x5	1	2	8	11	
	x1	x2	x3	x4	x5

Clustering Methods

- Agglomerative or divisive clustering
  - Agglomerative
    - Merge least distant elements to one cluster
  - Divisive
    - Split cluster in sub-cluster

Agglomerative Hierarchical Clustering

- Agglomerative Hierarchical Clustering
  - Nearest Neighbor or Single Link Method
    * The distance between groups is the distance between their nearest neighbors
  - Furthest Neighbor or Complete Link Method
    * The distance between groups is the distance between their most remote pair of individuals

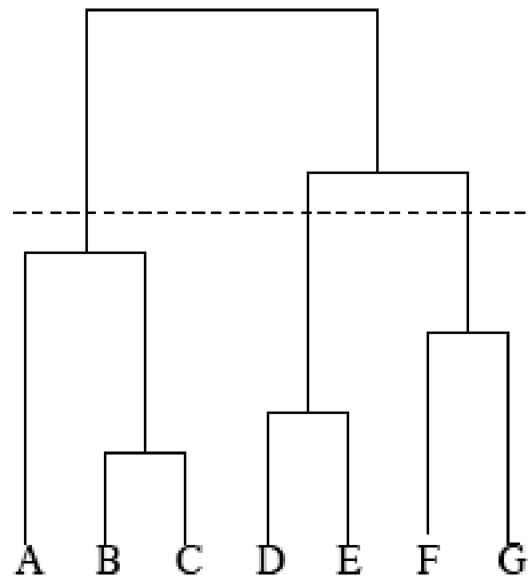
Agglomerative Hierarchical Clustering

- Agglomerative Hierarchical Clustering: Single Link Method
  - Given features and statistics, calculate distance matrix
  - Single link: search for minimal value and merge the corresponding two elements together (new cluster)
  - Recalculate the distance matrix, min(new cluster, other cluster) distance
  - Repeat until only one cell remains

Agglomerative Hierarchical Clustering

- Agglomerative Hierarchical Clustering: Complete Link Method
  - Given features and statistics, calculate distance matrix
  - Single link: search for minimal value and merge the corresponding two elements together (new cluster)
  - Recalculate the distance matrix, max(new cluster, other cluster) distance
  - Repeat until only one cell remains

Agglomerative Hierarchical Clustering



Agglomerative Hierarchical Clustering

- Example:
  - Cluster 1 & 5

x1					
x2	4				
x3	3	6			
x4	5	7	9		
x5	1	2	8	11	
	x1	x2	x3	x4	x5

Agglomerative Hierarchical Clustering

- Example:
  - Single link: cluster x1+5 & 2
  - Complete link: cluster x1+5 & 2

x1+5				
x2	2|4			
x3	3|8	6		
x4	5|11	7	9	
	x1+5	x2	x3	x4

Agglomerative Hierarchical Clustering

- Example:
  - Single link: cluster x(1+5)+2 & 3
  - Complete link: cluster x(1+5)+2 & 3

x(1+5)+2			
x3	3|8		
x4	5|11	9	
	x(1+5)+2	x3	x4

Clustering Example

- Example: Elghamry (2003)
  - Clustering of words
    - Hypothesis:
      Bifurcation of lexicon (open vs. closed class) can be accomplished using simple features of words available in the input.

Clustering Example

- Observations:
  - Frequency difference
  - Predictability from context
  - Learning patterns
  - Information load or semantic properties
  - Size and shape

Clustering Example

- Roberts (2002)
  - Clustering for tagger development
  - Clustering of content words given a set of function words

Clustering Example

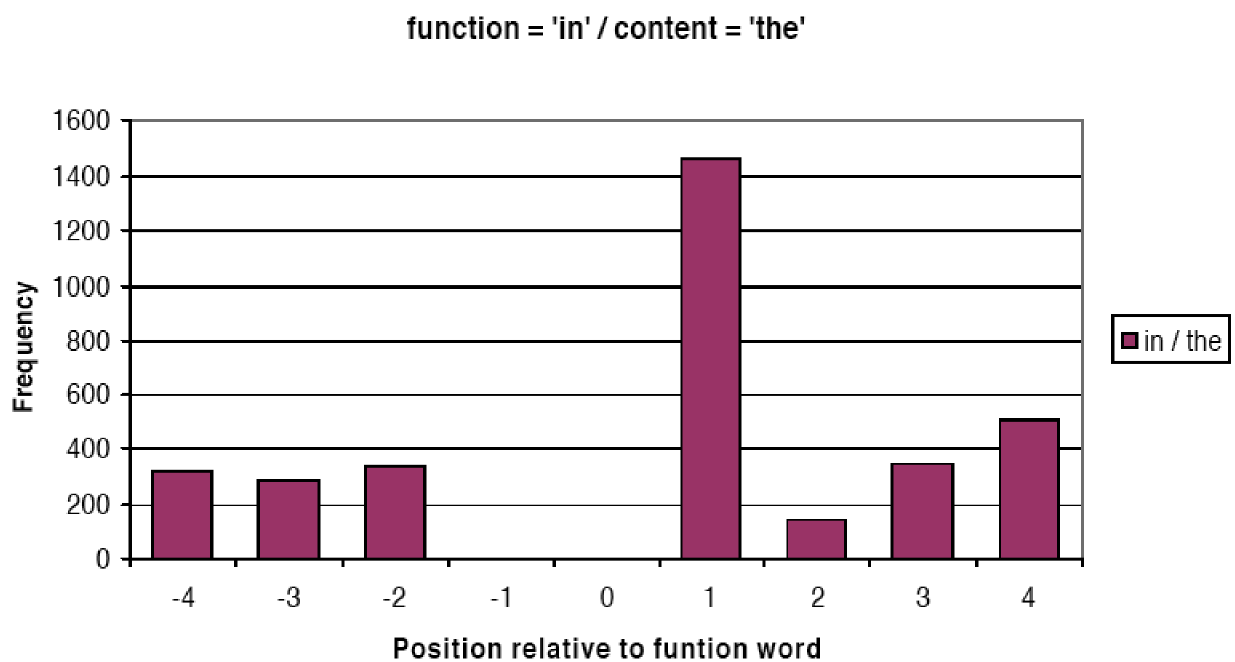
- Parameters (Roberts, 2002):
  - Contextual patterns (pos. in clause, bigram size)
  - Distance metric (distance in vector space)
  - Clustering method
  - Corpus size

Clustering Example

- Procedure (Roberts, 2002):
  - Selected set of 50 most frequent function words
  - Specification of a window (left and right of function word)
  - For all function words and all other words measure the position of other word in the window

Clustering Example

- Roberts (2002)



Clustering Example

- Translation into a vector (Roberts, 2002):

Relative Position	*Frequency*
-4	317
-3	288
-2	341
-1	0
0	0
1	1457
2	146
3	350
4	510

$$\begin{bmatrix} 317 \\ 288 \\ 341 \\ 0 \\ 1457 \\ 146 \\ 350 \\ 510 \end{bmatrix}$$

Clustering Example

- Vector concatenation for every substantive in all its functional contexts (Roberts, 2002) to create single vectors
- Vector normalization
- Clustering on the basis of vectors

Clustering Example

- Manipulation of parameters: distance metric, clustering algorithm, number of function words (vector length)

Clustering Example

- Results (Roberts, 2002):
  - Number of function words correlates with clustering accuracy
  - For some word classes 100% accuracy was reached (nouns, name prefixes etc.)

References

[1] Brian Everitt. *Cluster analysis*. Heinemann Educational for the Social Science Research Council, London, 1974.

[2] Anil K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: A review. *ACM Computing Surveys*, 31(3):264–323, 1999.

[3] Robert Choate Tryon. *Cluster analysis*. Edward Brothers, Ann Arbor, 1939.

Optimization Clustering

- Given a clustering criterion
  - How to find a partition into n groups that optimizes the criterion?
- Find all possible partitions and calculate their value of the given criterion.
- Choose the partition with the optimal value.

Optimization Clustering

- Complexity:
  - Number of possible partitions given n objects into g groups (Liu, 1968):

$$N(n,g) = \frac{1}{g!}\sum_{m=1}^{g}(-1)^{g-m}\binom{g}{m}m^n$$

- Example:
  - $N(50,4) = 5.3 \times 10^{28}$ or $N(100,5) = 6.6 \times 10^{67}$

Optimization Clustering

- Complexity solution
  - Programming strategies
    - Dynamic programming
    - Branch and bound algorithms
- Hill-climbing algorithms
  - Iterative search for optimum value of clustering criteria via rearrangement of existing partitions

Optimization Clustering

- K-means generates
  - k number of disjoint clusters (non-hierarchical)
  - globular clusters (spherical, elliptical, convex)
- properties:
  - numerical
  - unsupervised
  - iterative

Optimization Clustering

- K-means
  - k clusters
  - At least one element per cluster
  - No overlapping clusters
  - Non-hierarchical

Optimization Clustering

- K-means
  - Every member of a cluster is closer to its cluster than to any other cluster
  - Procedure

Optimization Clustering

- K-means
  - Initial partitioning of data set into k clusters
  - For each data point: calculate distance to each cluster
  - If one data point is closer to another cluster, relocate it
  - Repeat until no further relocations possible

Optimization Clustering

- K-means advantages
  - For large number of variables it is faster than hierarchical algorithms (for small k's)
  - Tighter clusters than hierarchical clustering, if cluster are globular

Optimization Clustering

- K-means disadvantages
  - Initial set of k clusters can affect the result
  - Does not work well with non-globular clusters

Optimization Clustering

- K-means example

Individual	Variable 1	Variable 2
1	1.0	1.0
2	1.5	2.0
3	3.0	4.0
4	5.0	7.0
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5

Optimization Clustering

- Initial 2 clusters on the basis of the most distant individuals:

	Individual	Mean Vector
Group 1	1	(1.0, 1.0)
Group 2	4	(5.0, 7.0)

Optimization Clustering

- Initial clustering of all remaining individuals:
  - For every other individual:
    - Calculate Euclidean distance to the centroid of every cluster
    - Assign individual to cluster
    - Recalculate centroid for every cluster

Optimization Clustering

- Mean vector or centroid (with coordinate x_i) with equal weight coordinates:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Optimization Clustering

- Initial clustering of all remaining individuals:

	Group 1		Group 2	
	Individual	Mean Vector	Individual	Mean Vector
Step 1	1	(1.0, 1.0)	4	(5.0, 7.0)
Step 2	1, 2	(1.3, 1.5)	4	(5.0, 7.0)
Step 3	1, 2, 3	(1.8, 2.3)	4	(5.0, 7.0)
Step 4	1, 2, 3	(1.8, 2.3)	4, 5	(4.3, 6.0)
Step 5	1, 2, 3	(1.8, 2.3)	4, 5, 6	(4.3, 5.7)
Step 6	1, 2, 3	(1.8, 2.3)	4, 5, 6, 7	(4.1, 5.4)

Optimization Clustering

- Initial partitions and clustering criterion:

	Individual	Mean Vector	Sum of SQR error
Group 1	1, 2, 3	(1.8, 2.3)	6.84
Group 2	4, 5, 6, 7	(4.1, 5.4)	5.38
total			12.22

- Error = for every point distance to centroid
  – Criterion: the smaller the sum of square errors, the better the cluster

Optimization Clustering

- Optimization Iteration:
  - Compare each individual's distance to its own mean with distance to the opposite group mean.
  - If distance to the mean in opposite group is smaller, relocate the individual.
  - Calculate the sum of square errors, if smaller than before, this is an improvement.

Optimization Clustering

- Distance to means:

Individual	distance to mean 1	distance to mean 2
1	1.5	5.4
2	0.4	4.3
3	2.1	1.8
4	5.7	1.8
5	3.2	0.7
6	3.8	0.8
7	2.8	1.1

Optimization Clustering

- Subsequent partitions and new clustering criterion:

	Individual	Mean Vector	Sum of SQR error
Group 1	1, 2	(1.3, 1.5)	0.63
Group 2	3, 4, 5, 6, 7	(3.9, 5.1)	7.9
total			8.53

- Decrease of clustering criterion (from 12.22 to 8.53).

Genetic algorithms

- Search and optimization techniques
  - Randomized!
- Components:
  - *Objective* or *fitness function*
  - Search space parameters encoded in a string = *chromosomes*
  - A collection of such strings = *population*

Genetic algorithms

- Survival of the fittest:
  - *Selection* of a set of strings (*mating pool*)
  - Subject to operations:
    - *Crossover*
    - *Mutation*

Genetic algorithms

- Iteration:
  - Selection & Crossover & Mutation
- Termination:
  - Fixed number of iterations
  - Specific termination condition

Genetic algorithms

- Given:
  - Fixed number K of clusters (cluster centres)
  - Set of n unlabeled points
- Clustering metric $\mathscr{M}$
  - Sum of Euclidean distance of the points from their respective cluster center
  - $\mathscr{M}(C_1, C_2, \ldots, C_K) = \sum_{i=1}^{K} \sum_{x_j \in C_i} \|X_j - Z_i\|$

Genetic algorithms

- Task:
  - Search cluster centres $Z_1, Z_2, \ldots, Z_K$ such that $\mathscr{M}$ is minimized
- String representation:
  - Sequence of real numbers representing K cluster centres
  - Length for N-dimensions $= N * K = X_1\ Y_1\ X_2\ Y_2 \ldots$

Genetic algorithms

- *Population initialization*:
  - Random choice of K points from the population
- *Crossover*:
  - Generate crossover point randomly ($[1, l-1]$)
  - Exchange right of crossover point
  - From two parents produce two offsprings

Genetic algorithms

- Mutation:
  - Flipping value of a binary gene
  - Real numbers:
    Generate δ in the range of $[0, 1]$
    For gene value v:
    - $v \pm 2 * \delta * v, v \neq 0$
    - $v \pm 2 * \delta, v = 0$

Genetic algorithms

- Fitness computation:
  - Initialization: random choice of centroids
  - Subsequent assignment of the points to centroids, where distance to centroid is the minimal
  - Recalculation of centroid
  - Replacement of chromosome by new centroids
- Fitness function: $1/\mathscr{M}$

Next Readings

- Déjean [1] (French!)
- Nakov [3]
- Maulik [2] (Everybody for next week!)

Term Projects

- Proposals
- Possibilities

References

[1] Hervé Déjean. *Concepts et algorithmes pour la découverte des structures formelles des langues*. doctoral dissertation, Université de Caen Basse Normandie, 1998.

[2] Ujjwal Maulik and Sanghamitra Bandyopadhyay. Genetic algorithm-based clustering techniques. *Pattern Recognition*, 33:1455–1465, 2000.

[3] Preslav Nakov. *Recognition and Morphological Clas-*

sification of Unknown Words for German. diploma, Sofia University, 2001.

[4] Andrew Roberts. Automatic acquisition of word classification using distributional analysis of content words with respect to function words. Technical report, University of Leeds, School of Computing, 2002.

[5] Marcin Olof Szummer. *Learning from Partially Labeled Data*. doctoral dissertation, MIT, 2002.