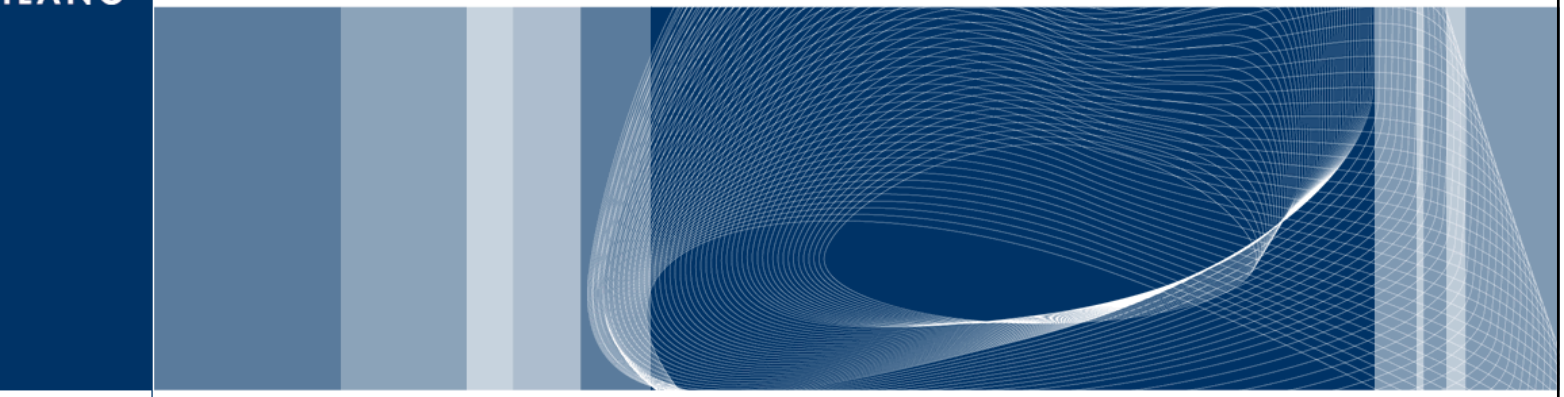
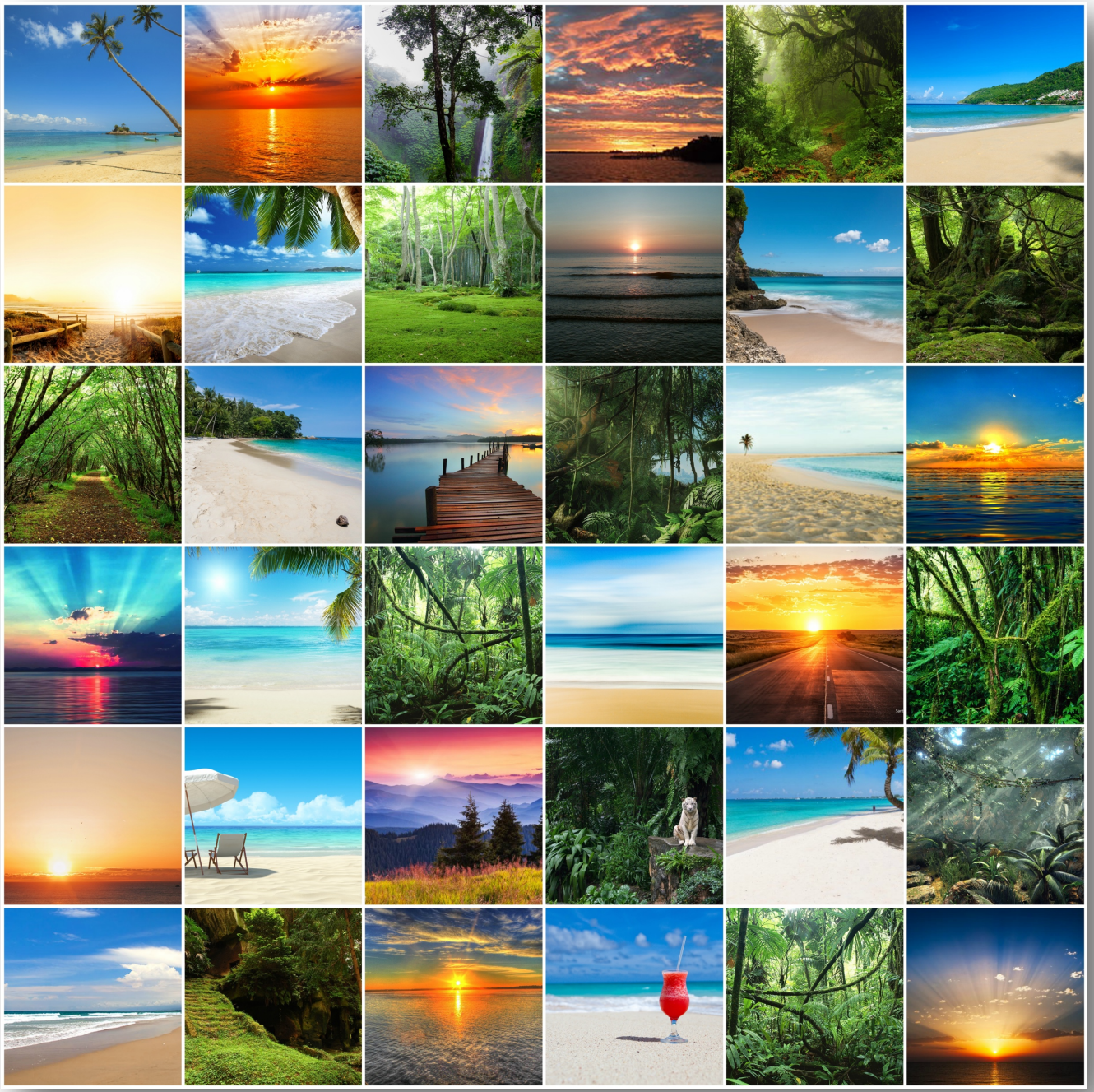
POLITECNICO DI MILANO



Clustering

Data Science for Mobility

Prof. Pier Luca Lanzi



Clustering algorithms group a collection of data points into "clusters" according to some distance measure

Data points in the same cluster should have a small distance from one another

Data points in different clusters should be at a large distance from one another

Clustering searches for “natural” grouping/structure in un-labeled data

What is Cluster Analysis?

- A cluster is a collection of data objects
  - Similar to one another within the same cluster
  - Dissimilar to the objects in other clusters

- Cluster analysis
  - Given a set data points try to understand their structure
  - Finds similarities between data according to the characteristics found in the data
  - Groups similar data objects into clusters
  - It is unsupervised learning since there is no predefined classes

Clustering Applications

- Marketing
  - Help marketers discover distinct groups in their customer bases, and then use this knowledge to develop targeted marketing programs
- Land use
  - Identification of areas of similar land use in an earth observation database
- Insurance
  - Identifying groups of motor insurance policy holders with a high average claim cost
- City-planning
  - Identifying groups of houses according to their house type, value, and geographical location
- Earth-quake studies
  - Observed earth quake epicenters should be clustered along continent faults

What Is Good Clustering?

- A good clustering consists of high quality clusters with
  - High intra-class similarity
  - Low inter-class similarity
- The quality of a clustering result depends on both the similarity measure used by the method and its implementation
- The quality of a clustering method is also measured by its ability to discover some or all of the hidden patterns
- Evaluation
  - Various measure of intra/inter cluster similarity
  - Manual inspection
  - Benchmarking on existing labels

Measure the Quality of Clustering

- Dissimilarity/Similarity metric: Similarity is expressed in terms of a distance function, typically metric d(i, j)
- There is a separate "quality" function that measures the "goodness" of a cluster
- The definitions of distance functions are usually very different for interval-scaled, Boolean, categorical, ordinal ratio, and vector variables
- Weights should be associated with different variables based on applications and data semantics
- It is hard to define "similar enough" or "good enough" as the answer is typically highly subjective

Clustering Methods

- Hierarchical vs point assignment
- Numeric and/or symbolic data
- Deterministic vs. probabilistic
- Exclusive vs. overlapping
- Hierarchical vs. flat
- Top-down vs. bottom-up

Data Structures

Outlook	Temp	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
…	…	…	…	…

Data Matrix

$$\begin{bmatrix} x_{11} & \dots & x_{1f} & \dots & x_{1p} \\ \dots & \dots & \dots & \dots & \dots \\ x_{i1} & \dots & x_{if} & \dots & x_{ip} \\ \dots & \dots & \dots & \dots & \dots \\ x_{n1} & \dots & x_{nf} & \dots & x_{np} \end{bmatrix}$$

Dis/Similarity Matrix

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ : & : & : & & \\ d(n,1) & d(n,2) & \dots & \dots & 0 \end{bmatrix}$$

Distance/Similarity Measures

Distance Measures

- Given a space and a set of points on this space, a distance measure d(x,y) maps two points x and y to a real number, and satisfies three axioms
- $d(x,y) \geq 0$
- $d(x,y) = 0$ if and only $x=y$
- $d(x,y) = d(y,x)$
- $d(x,y) \leq d(x,z) + d(z,y)$

Euclidean Distances

- L_r-norm

$$d([x_1, x_2, \ldots, x_n],\ [y_1, y_2, \ldots, y_n]) = (\sum_{i=1}^{n} |x_i - y_i|^r)^{1/r}$$

- Euclidean distance (r=2)

$$d([x_1, x_2, \ldots, x_n],\ [y_1, y_2, \ldots, y_n]) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

- Manhattan distance (r=1)

$$d([x_1, x_2, \ldots, x_n], [y_1, y_2, \ldots, y_n]) = \sum_{i=1}^{n} |x_i - y_i|$$

- L_∞-norm

$$d([x_1, x_2, \ldots, x_n], [y_1, y_2, \ldots, y_n]) = \max_{i=1}^{n} |x_i - y_i|$$

Jaccard Distance

- Jaccard distance is defined as

$$d(x,y) = 1 - SIM(x,y)$$

- SIM is the Jaccard similarity,

$$SIM(x,y) = \frac{x \cap y}{x \cup y}$$

- Which can also be interpreted as the percentage of identical attributes

Cosine Distance

- The cosine distance between x, y is the angle that the vectors to those points make

$$d(x,y) = \arccos \frac{\sum_1^n x_i y_i}{\sqrt{\sum_i^n x_i^2}\sqrt{\sum_i^n y_i^2}}$$

- This angle will be in the range 0 to 180 degrees, regardless of how many dimensions the space has.

- Example: given x = (1,2,-1) and y = (2,1,1) the angle between the two vectors is 60

Edit Distance

- The distance between a string x=x1x2…xn and y=y1y2…ym is the smallest number of insertions and deletions of single characters that will transform x into y

- Alternatively, the edit distance d(x, y) can be compute as the longest common subsequence (LCS) of x and y and then,

$$d(x,y) = |x| + |y| - 2|LCS|$$

- Example
  - The edit distance between **x=abcde** and **y=acfdeg** is 3 (delete b, insert f, insert g), the LCS is acde which is coherent with the previous result

Hamming Distance

- Hamming distance between two vectors is the number of components in which they differ
- Or equivalently, given the number of variables p, and the number m of matching components, we define

$$d(x, y) = \frac{p - m}{p}$$

- Example: the Hamming distance between the vectors 10101 and 11110 is 3/5.

Requisites for Clustering Algorithms

- Scalability
- Ability to deal with different types of attributes
- Ability to handle dynamic data
- Discovery of clusters with arbitrary shape
- Minimal requirements for domain knowledge to determine input parameters
- Able to deal with noise and outliers
- Insensitive to order of input records
- High dimensionality
- Incorporation of user-specified constraints
- Interpretability and usability

Curse of Dimensionality

in high dimensions, almost all pairs of points are equally far away from one another

almost any two vectors are almost orthogonal