CLUSTER ANALYSIS II

Andrea Giansanti
Dipartimento di Fisica, Sapienza Università di Roma

Andrea.Giansanti@roma1.infn.it

CB_20_21_L16 , Rome 19th Nov 2020

DIPARTIMENTO DI FISICA

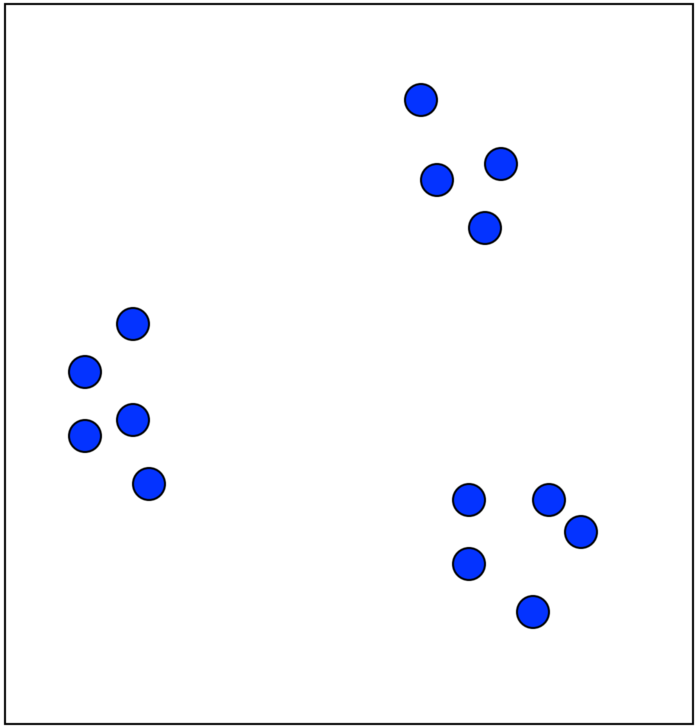
SAPIENZA
UNIVERSITÀ DI ROMA

OUTLINE

- K-Mean (recap) (partitioning)
- An information based metric in the space of clusterings
- DBSCAN (density based)
- Superparamagnetic (couplings, interactions)

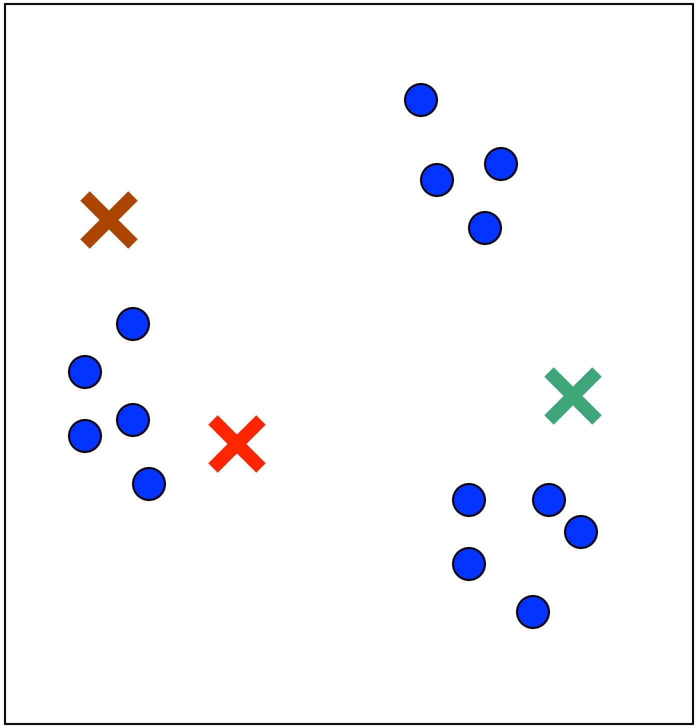
K-means (recap 1)

“GUESS” K=3



K-means (recap 2)

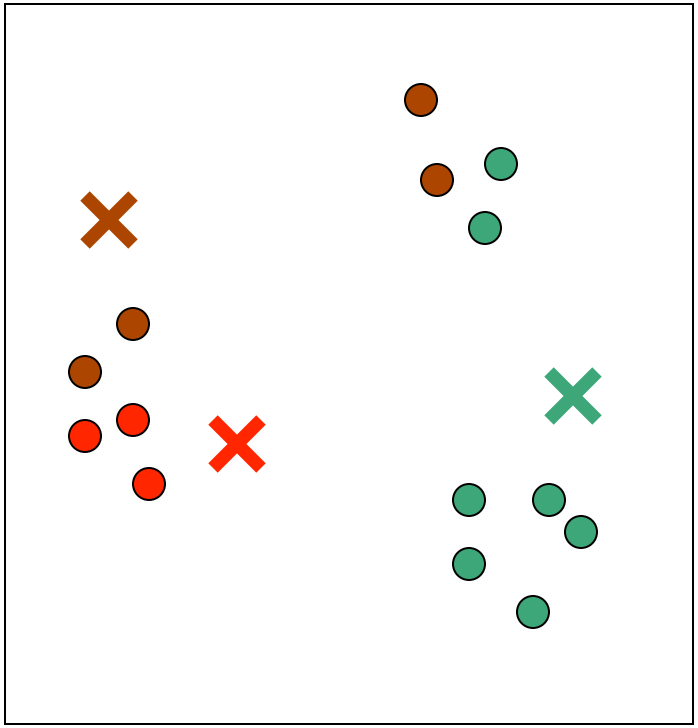
- Start with random positions of centroids. ✕



Iteration = 0

K-means (recap 3)

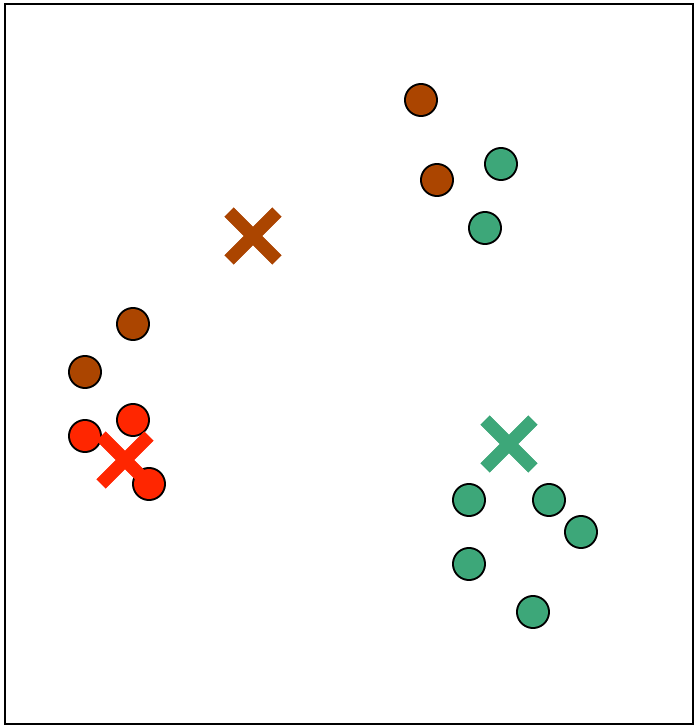
- Start with random positions of centroids.
- Assign each data point to closest centroid



Iteration = 1

K-means (recap4)

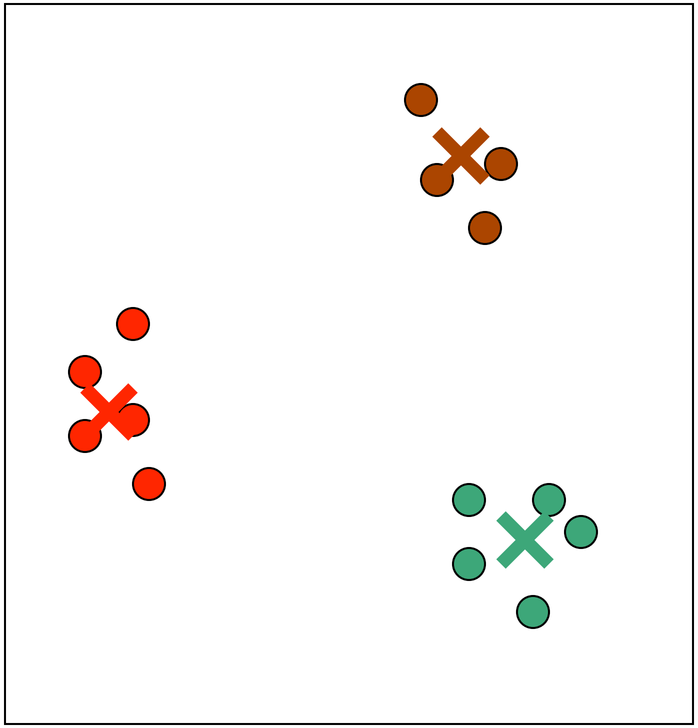
- Start with random positions of centroids.
- Assign each data point to closest centroid
- Move centroids to center of assigned points



Iteration = 1

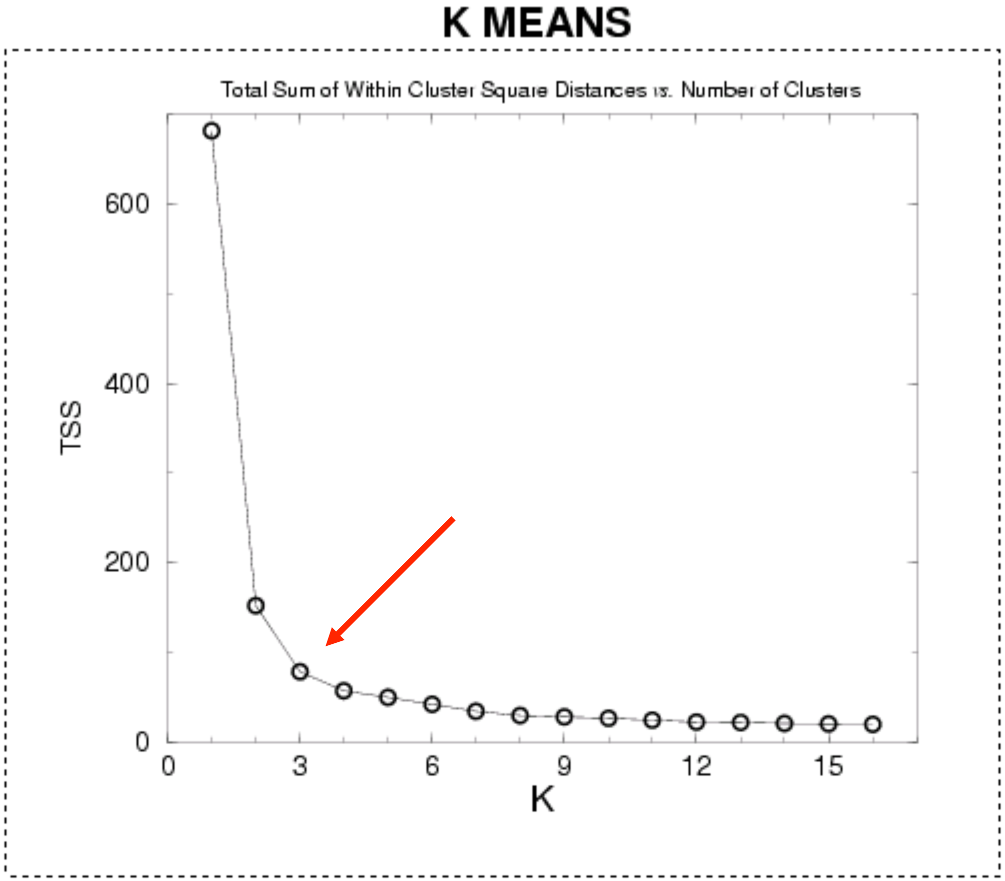
K-means; algorithm to find minima

- Start with random positions of centroids.
- Assign each data point to closest centroid
- Move centroids to center of assigned points
- Iterate till minimal cost



Iteration = 3

E=Total Sum of Squares vs K



K-means - Summary

- Result depends on initial centroids' position
- Fast algorithm: compute distances from data points to centroids
- O(N) operations (vs O(N^2))

- Must preset K
- Fails for non-spherical distributions

Entropies, Mutual Information Between Clusterings

relative proportions of the clusters.

We now define the *mutual information* between two clusterings, i.e. the information that one clustering has about the other. Denote by $P(k)$, $k = 1, \ldots, K$ and $P'(k')$, $k' = 1, \ldots, K'$ the random variables associated with the clusterings $\mathcal{C}$, $\mathcal{C}'$. Let $P(k, k')$ represent the probability that a point belongs to C_k in clustering $\mathcal{C}$ and to $C'_{k'}$ in $\mathcal{C}'$, namely the joint distribution of the random variables associated with the two clusterings:

$$P(k, k') = \frac{|C_k \bigcap C'_{k'}|}{n}. \tag{15}$$

We define $I(\mathcal{C}, \mathcal{C}')$ the mutual information between the clusterings $\mathcal{C}$, $\mathcal{C}'$ to be equal to the mutual information between the associated random variables

$$I(\mathcal{C}, \mathcal{C}') = \sum_{k=1}^{K} \sum_{k'=1}^{K'} P(k, k') \log \frac{P(k, k')}{P(k)P'(k')}. \tag{16}$$

Intuitively, we can think of $I(\mathcal{C}, \mathcal{C}')$ in the following way: we are given a random point in D. The uncertainty about its cluster in $\mathcal{C}'$ is measured by $H(\mathcal{C}')$. Suppose now that we are told which cluster the point belongs to in $\mathcal{C}$. How much does this knowledge reduce the uncertainty about $\mathcal{C}'$? This reduction in uncertainty, averaged over all points, is equal to $I(\mathcal{C}, \mathcal{C}')$.

$$I(\mathcal{C}, \mathcal{C}') = I(\mathcal{C}', \mathcal{C}) \geqslant 0. \tag{17}$$

Also, the mutual information can never exceed the total uncertainty in a clustering, so

$$I(\mathcal{C}, \mathcal{C}') \leqslant \min(H(\mathcal{C}),\ H(\mathcal{C}')). \tag{18}$$

Equality in the above formula occurs when one clustering completely determines the other. For example, if $\mathcal{C}'$ is obtained from $\mathcal{C}$ by merging two or more clusters, then

$$I(\mathcal{C}, \mathcal{C}') = H(\mathcal{C}') < H(\mathcal{C}).$$

When the two clusterings are equal, and only then, we have

$$I(\mathcal{C}, \mathcal{C}') = H(\mathcal{C}') = H(\mathcal{C}).$$

We propose to use as a comparison criterion for two clusterings $\mathcal{C}, \mathcal{C}'$ the quantity

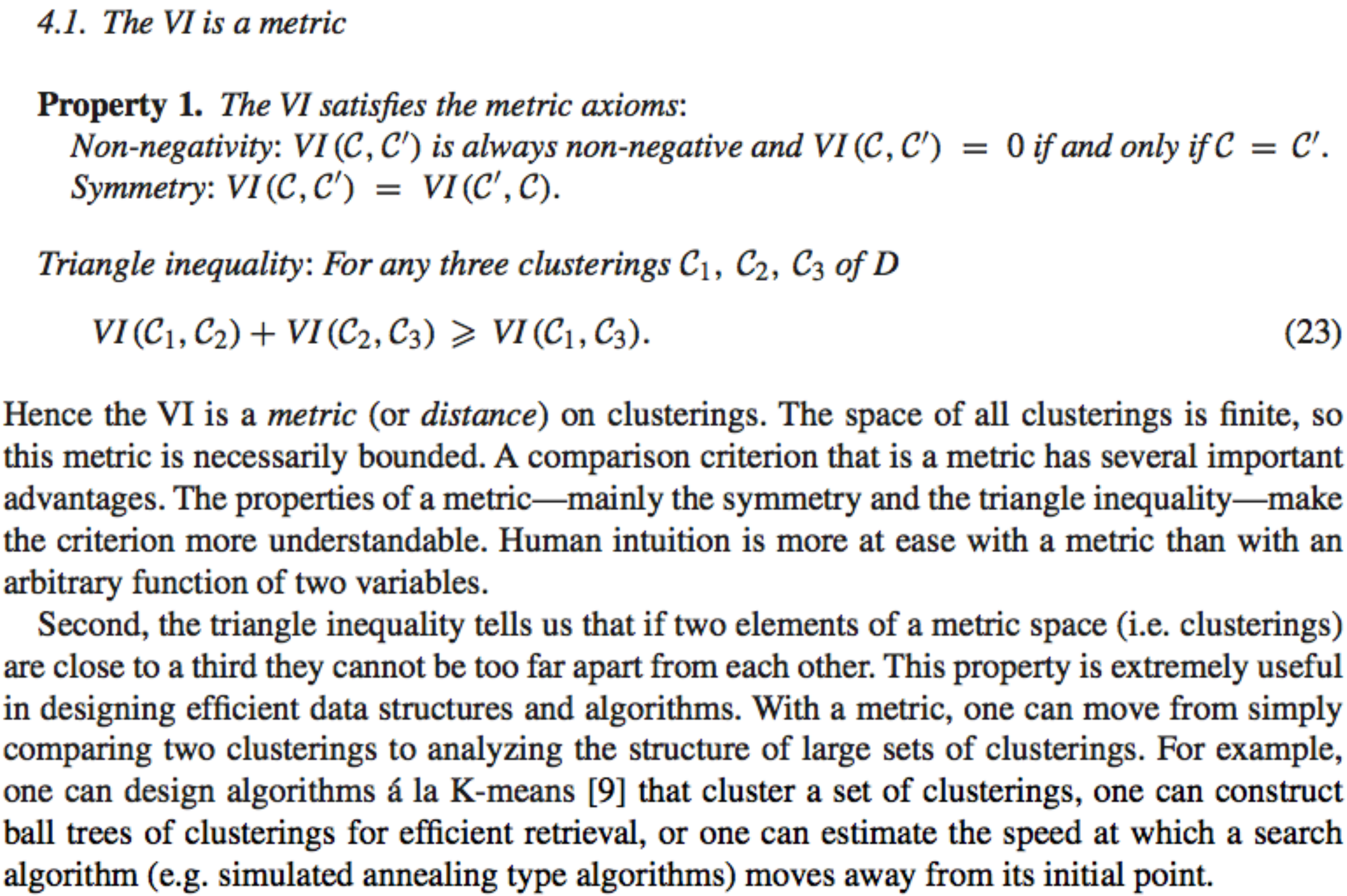
$$VI(\mathcal{C}, \mathcal{C}') = H(\mathcal{C}) + H(\mathcal{C}') - 2I(\mathcal{C}, \mathcal{C}'). \tag{19}$$

At a closer examination, this is the sum of two positive terms

$$VI(\mathcal{C}, \mathcal{C}') = [H(\mathcal{C}) - I(\mathcal{C}, \mathcal{C}')] + [H(\mathcal{C}') - I(\mathcal{C}, \mathcal{C}')]. \tag{20}$$

By analogy with the total variation of a function, we call it VI between the two clusterings. The two terms represent the conditional entropies $H(\mathcal{C}|\mathcal{C}')$, $H(\mathcal{C}'|\mathcal{C})$. The first term measures the amount

VI induces a metric in the space of clusterings (Meila2007)



4.1. The VI is a metric

Property 1. *The VI satisfies the metric axioms*:

Non-negativity: $VI(\mathcal{C}, \mathcal{C}')$ *is always non-negative and* $VI(\mathcal{C}, \mathcal{C}') = 0$ *if and only if* $\mathcal{C} = \mathcal{C}'$.

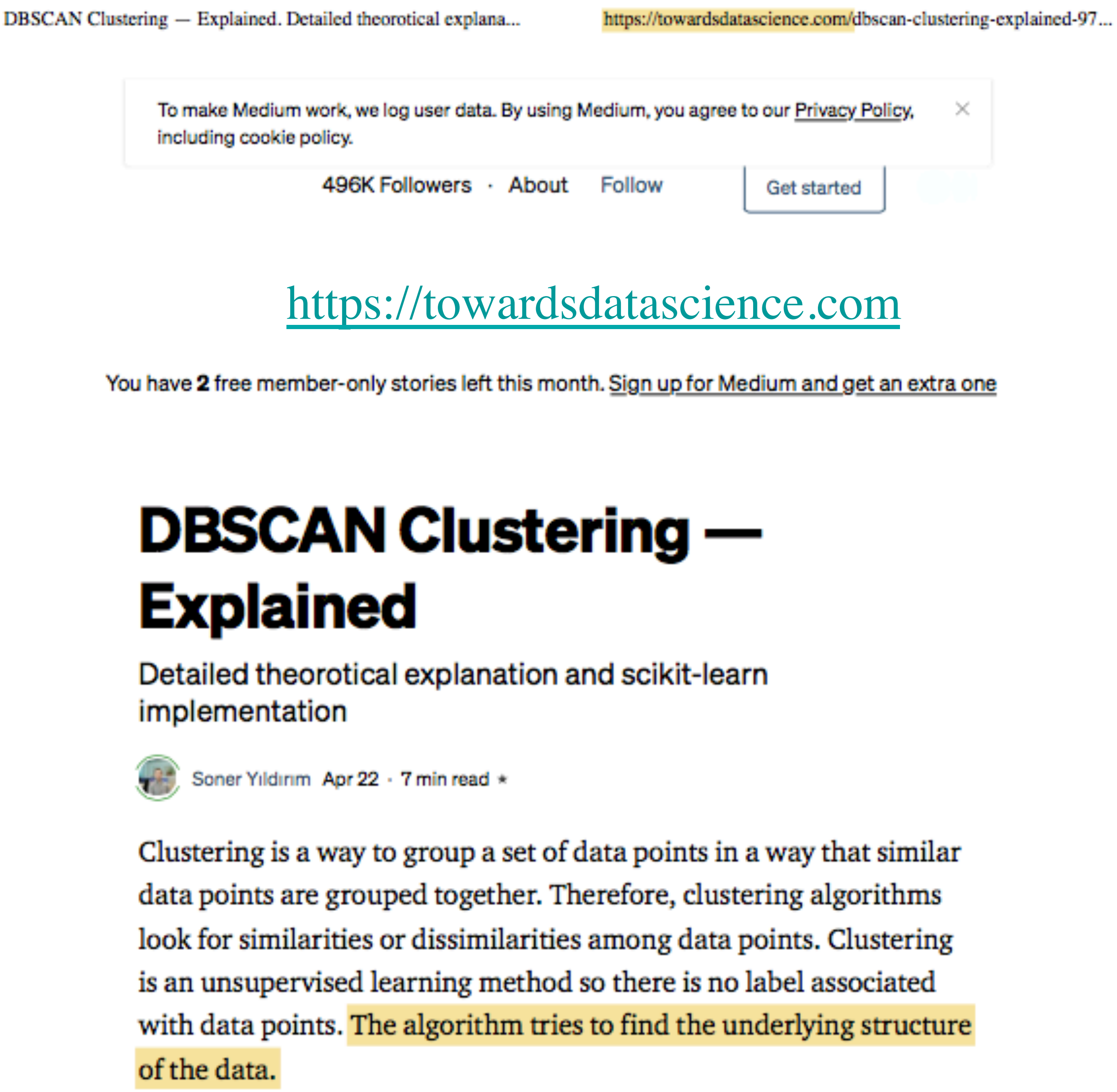
Symmetry: $VI(\mathcal{C}, \mathcal{C}') = VI(\mathcal{C}', \mathcal{C})$.

Triangle inequality: *For any three clusterings* $\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3$ *of* D

$$VI(\mathcal{C}_1, \mathcal{C}_2) + VI(\mathcal{C}_2, \mathcal{C}_3) \geqslant VI(\mathcal{C}_1, \mathcal{C}_3). \tag{23}$$

Hence the VI is a *metric* (or *distance*) on clusterings. The space of all clusterings is finite, so this metric is necessarily bounded. A comparison criterion that is a metric has several important advantages. The properties of a metric—mainly the symmetry and the triangle inequality—make the criterion more understandable. Human intuition is more at ease with a metric than with an arbitrary function of two variables.

Second, the triangle inequality tells us that if two elements of a metric space (i.e. clusterings) are close to a third they cannot be too far apart from each other. This property is extremely useful in designing efficient data structures and algorithms. With a metric, one can move from simply comparing two clusterings to analyzing the structure of large sets of clusterings. For example, one can design algorithms á la K-means [9] that cluster a set of clusterings, one can construct ball trees of clusterings for efficient retrieval, or one can estimate the speed at which a search algorithm (e.g. simulated annealing type algorithms) moves away from its initial point.



DBSCAN Clustering — Explained. Detailed theorotical explana... https://towardsdatascience.com/dbscan-clustering-explained-97...

To make Medium work, we log user data. By using Medium, you agree to our Privacy Policy, including cookie policy.

https://towardsdatascience.com

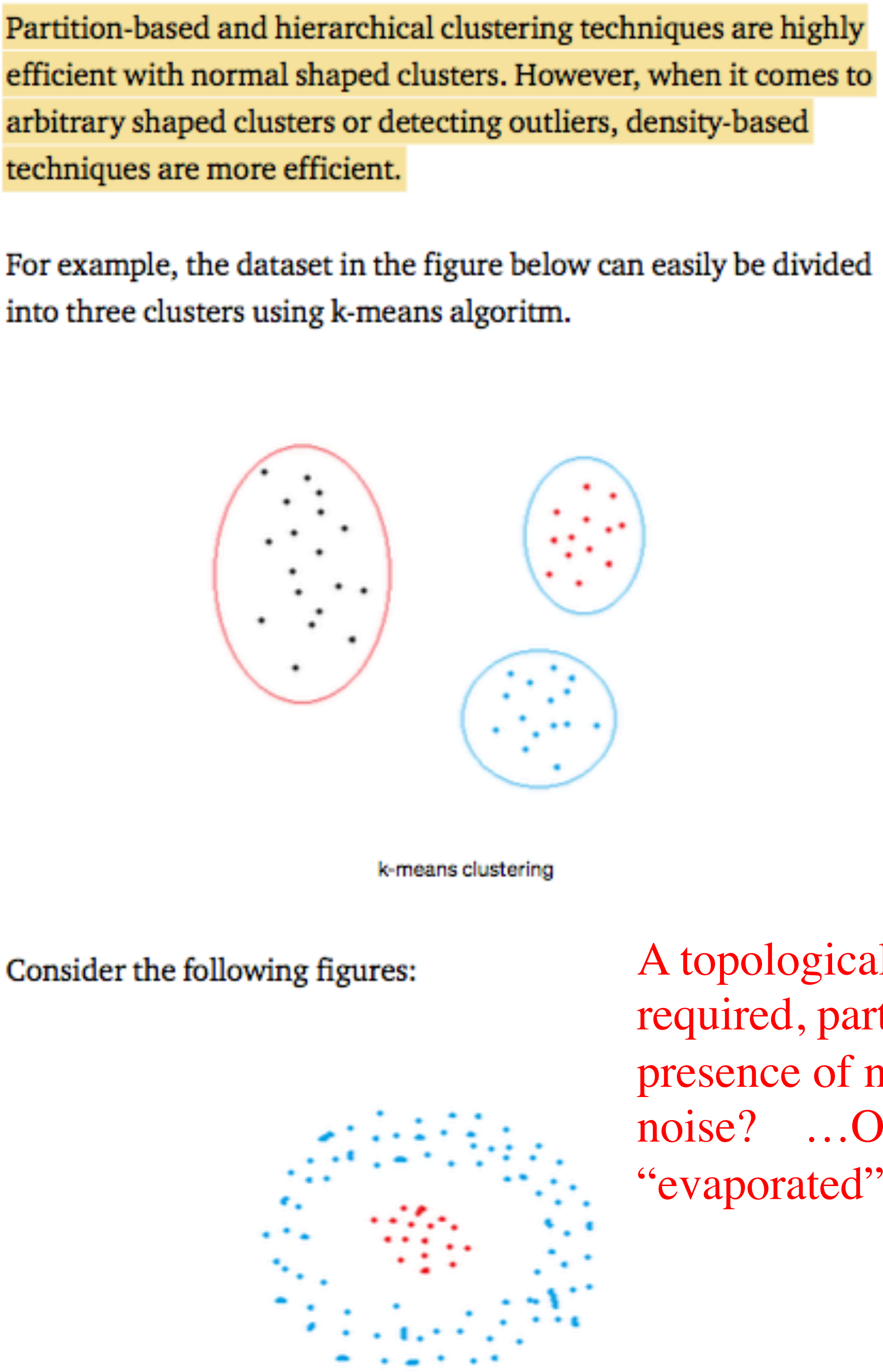
You have **2** free member-only stories left this month. Sign up for Medium and get an extra one

DBSCAN Clustering — Explained

Detailed theorotical explanation and scikit-learn implementation

Soner Yıldırım Apr 22 · 7 min read ★

Clustering is a way to group a set of data points in a way that similar data points are grouped together. Therefore, clustering algorithms look for similarities or dissimilarities among data points. Clustering is an unsupervised learning method so there is no label associated with data points. The algorithm tries to find the underlying structure of the data.



Partition-based and hierarchical clustering techniques are highly efficient with normal shaped clusters. However, when it comes to arbitrary shaped clusters or detecting outliers, density-based techniques are more efficient.

For example, the dataset in the figure below can easily be divided into three clusters using k-means algoritm.

k-means clustering

Consider the following figures:

A topological data analysis is required, particularly in the presence of noise: what is noise? …Outliers, “evaporated” data

DBSCAN algorithm

DBSCAN stands for density-based spatial clustering of applications with noise. It is able to find arbitrary shaped clusters and clusters with noise (i.e. outliers).

There are two key parameters of DBSCAN:

- **eps**: The distance that specifies the neighborhoods. Two points are considered to be neighbors if the distance between them are less than or equal to eps.
- **minPts:** Minimum number of data points to define a cluster.

Based on these two parameters, points are classified as core point, border point, or outlier:

- **Core point:** A point is a core point if there are at least minPts number of points (including the point itself) in its surrounding area with radius eps.
- **Border point:** A point is a border point if it is reachable from a core point and there are less than minPts number of points within its surrounding area.
- **Outlier:** A point is an outlier if it is not a core point and not reachable from any core points.

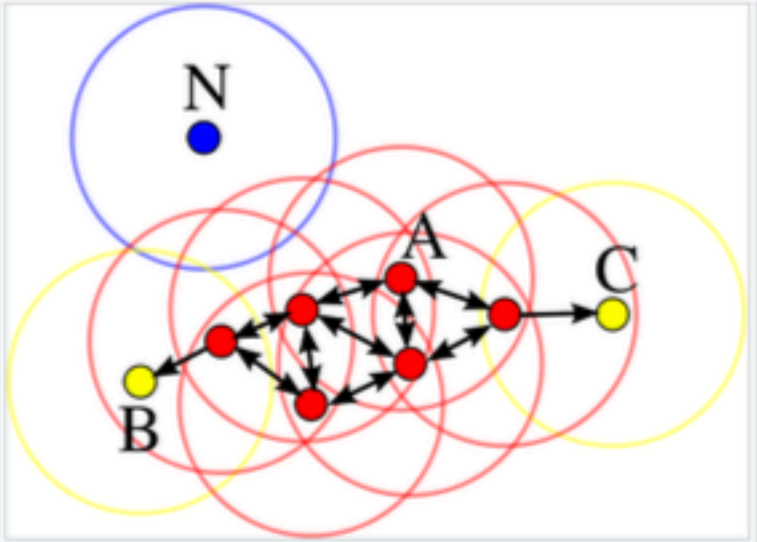


Figure source

In this case, minPts is 4. Red points are core points because there are **at least** 4 points within their surrounding area with radius eps. This area is shown with the circles in the figure. The yellow points are border points because they are reachable from a core point and have less than 4 points within their neighborhood. Reachable means being in the surrounding area of a core point. The points B and C have two points (including the point itself) within their neigborhood (i.e. the surrounding area with a radius of eps). Finally N is an outlier because it is not a core point and cannot be reached from a core point.

Dbscan LINKS fro towards datascience.com
with practical applications in python

Soner Yildirim
https://towardsdatascience.com/dbscan-clustering-explained-97556a2ad556

Kamil Mysiak
https://towardsdatascience.com/explaining-dbscan-clustering-18eaf5c83b31

DBScan in Wikipedia
https://en.wikipedia.org/wiki/DBSCAN

Original paper Ester1996 and DBSCAN revisited by the same authors

ELSEVIER

Physica A 263 (1999) 158–169

PHYSICA A

Superparamagnetic Clustering of Data - The Definitive Solution of an Ill-Posed Problem

Eytan Domany

Department of Physics of Complex Systems, Weizmann Inst. of Science, Rehovot 76100, Israel

Abstract

Clustering is an important technique in exploratory data analysis, with applications in image processing, object classification, target recognition, data mining etc. The aim is to partition data according to natural classes present in it, assigning data points that are "more similar" to the same "cluster". We solved this ill-posed problem without making any assumptions about the structure of the data, by using a physical system as an *analog computer*. The physical system we use is a disordered (granular) magnet. The method was tested successfully on a variety of artificial and real-life problems, such as classification of flowers, processing of satellite images, speech recognition and identification of textures and images. We are currently involved in several collaborations, applying the method to image classification, fMRI data analysis and classification of protein structures.

I review here work done using[1] a novel clustering technique, *Super Paramagnetic Clustering (SPC)*[5,6]. The motivation for the method originates in the physics of disordered granular magnets. In Sec 2.1 I introduce the cost function used by SPC; this cost function has the form of the Hamiltonian of a disordered Potts ferromagnet. The connection to Equilibrium Statistical Mechanics is natural and is explained in Sec 2.2. As we will see, the *temperature T controls the resolution* at which the data are clustered. Various equilibrium properties of the system are measured by Monte Carlo; in particular, the correlations of neighboring pairs is measured and serves to determine the assignment of data points to clusters, as explained in Sec 2.3. In Sec 3 we apply the method to a variety of problems.

2 Superparamagnetic Clustering of Data

2.1 The Cost Function

The basic premise of our approach is the following; data points i, j that are highly similar to one another, i.e. with small d_{ij}, are likely to belong to the same clusters; the closer two points are, the more unlikely they are to belong to different clusters. To put this statement on a formal ground, we assign to every data point i a Potts spin variable [2] $S_i = 1, 2, ...q$. Any particular clustering assignment is represented as a configuration $\{S\} = \{S_1, S_2, ...S_N\}$ of all the Potts spin variables. Losely speaking, $S_i = S_j$ indicates that i and j belong to the same cluster. An assignment with $S_i \neq S_j$ means that the two points are in different clusters, and such an assignment draws a penalty J_{ij}. A cost function that reflects these statements has the form

$$\mathcal{H}(\{S\}) = \sum_{<i,j>} J_{ij} \left(1 - \delta_{S_i,S_j}\right) \tag{1}$$

with J_{ij} a decreasing function of the "distance" d_{ij} between the data points i, j. In most applications we used a Gaussian decay of the interaction strength with distance, cut off beyond some distance or some number of neighbors; we expect, however, that neither the kind of spins used, nor the precise functional form of $J_{ij}(d_{ij})$ has a qualitative effect on the results. In particular, the number of Potts components q has nothig to do with the number of clusters.

At temperature $T = 0$ such a disordered ferromagnet is in its ground state, in which all spins are aligned. At high temperatures the system is completely disordered, with vanishing correlation between any pair of spins. The manner in which the system changes as T varies between these extremes depends on the struture in the data. If we have one

PRACTICAL APPLICATIONS

Super paramagnetic unsupervised clustering (**P**)(Blatt1996, Tetko2005 see also: http://www.vcclab.org/lab/spc/and also the very useful link to Rudy Stoop's computational biology clustering page

http://stoop.ini.uzh.ch/research/clustering).