So humans got the SARS virus from palm civets... right?

- With the data at hand, we see how the virus used different hosts, moving from **bat to human to civet**, in that order. So the civets actually got SARS **from humans**.

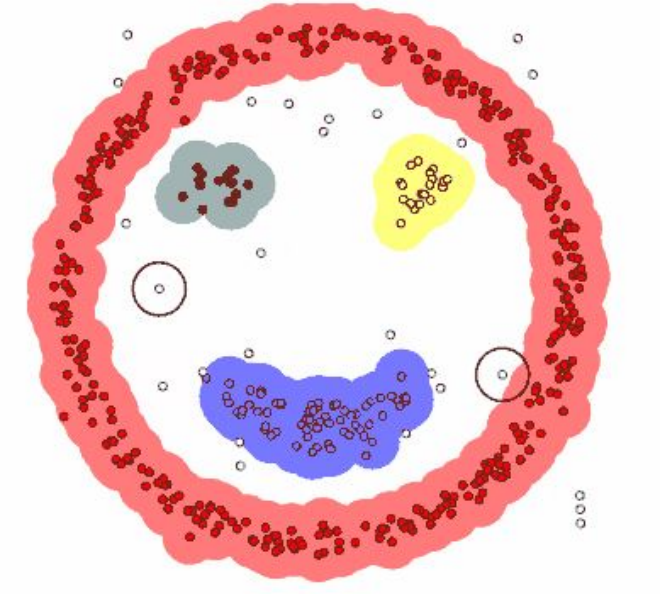
AGENDA

- What is Clustering?
- Unsupervised Learning
- Why Clustering?
- Types of Clustering
  - Partitioning Clustering
- K Means Clustering
- Challenges in K Means Clustering
- Elbow Method
- Euclidean Distance
- Illustration of K Means algorithm
- Applications of K Means

References

- Hierarchical Clustering
  - Agglomerative Clustering
  - Divisive Clustering
- Applications
- **Density Based Clustering**
- Distance metrics
  - Manhattan
  - Minkowski
  - Mahalanobis

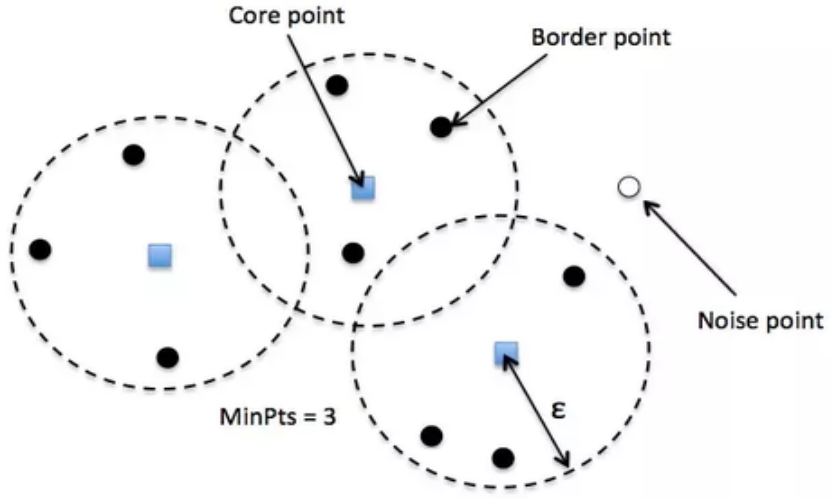
Density Based Clustering



Density Based Clustering

- DBCSAN stands **for Density-Based Spatial Clustering of Applications with Noise**
- It divides a dataset into subgroups of **high density regions.**
- There are two parameters required for DBSCAN:
  - **epsilon (ε)** and **minimum amount of points** required to form a cluster (**minPts**).
- **ε** is a distance parameter **that defines the radius** to search for nearby neighbors.

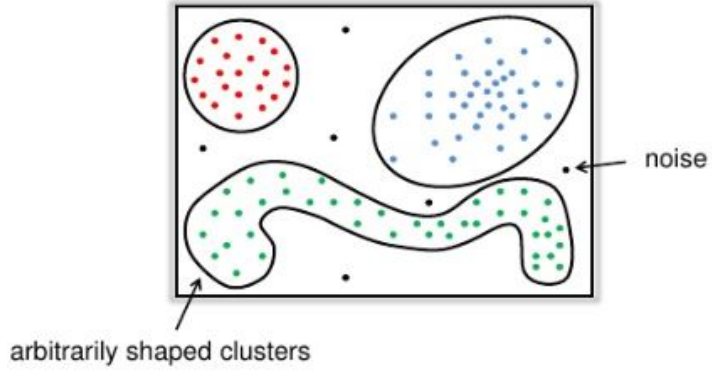
- **Core point** – a point that has at least a minimum number of other points (**minPts)** within its **ε** radius.
- **Border point** – a point is within the **ε** radius of a core point but has less than the minimum number of other points (**minPts)** within its own **ε** radius
- **Noise point** – a point that is neither a core point or a border point.



So much noise...

When to use Density based clustering?

- DBSCAN groups together points that are **close to each other**.
- Based on a **distance measurement** (usually Euclidean distance) and a **minimum number of points.**
- It also marks as **outliers the points** that are in low-density regions.
- It **removes noise**.



AGENDA

- What is Clustering?
- Unsupervised Learning
- Why Clustering?
- Types of Clustering
  - Partitioning Clustering
- K Means Clustering
- Challenges in K Means Clustering
- Elbow Method
- Euclidean Distance
- Illustration of K Means algorithm
- Applications of K Means

References

- Hierarchical Clustering
  - Agglomerative Clustering
  - Divisive Clustering
- Applications
- Density Based Clustering
- **Distance metrics**
  - **Manhattan**
  - Minkowski
  - Mahalanobis

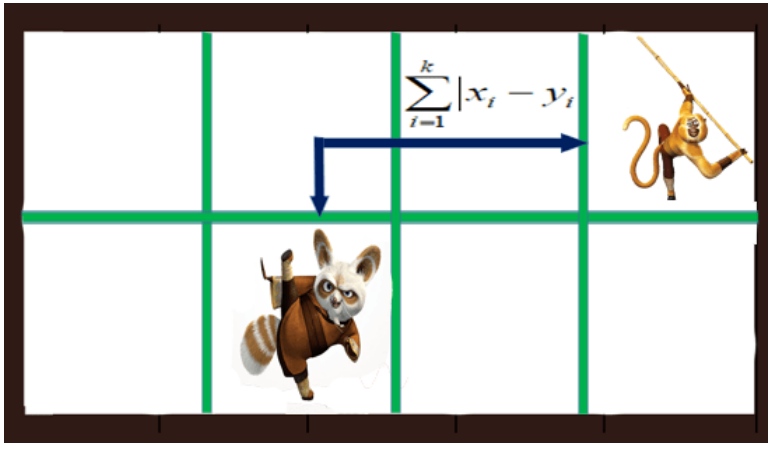
Distance Metrics

Manhattan Distance

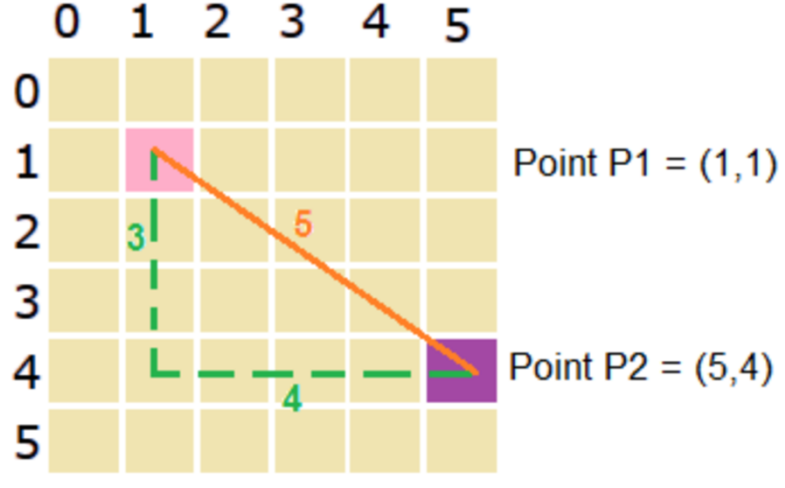
Manhattan Distance

- The **distance between two points** measured along axes at **right angles.**
- It is based on **absolute value distance**.
- Absolute value distance should give **more robust results**.

$$d_1(\mathbf{p}, \mathbf{q}) = \|\mathbf{p} - \mathbf{q}\|_1 = \sum_{i=1}^{n} |p_i - q_i|,$$



Manhattan Distance



Euclidean distance = $\sqrt{(5-1)^2 + (4-1)^2} = 5$

Manhattan distance = $|5-1| + |4-1| = 7$

AGENDA

- What is Clustering?
- Unsupervised Learning
- Why Clustering?
- Types of Clustering
  - Partitioning Clustering
- K Means Clustering
- Challenges in K Means Clustering
- Elbow Method
- Euclidean Distance
- Illustration of K Means algorithm
- Applications of K Means

References

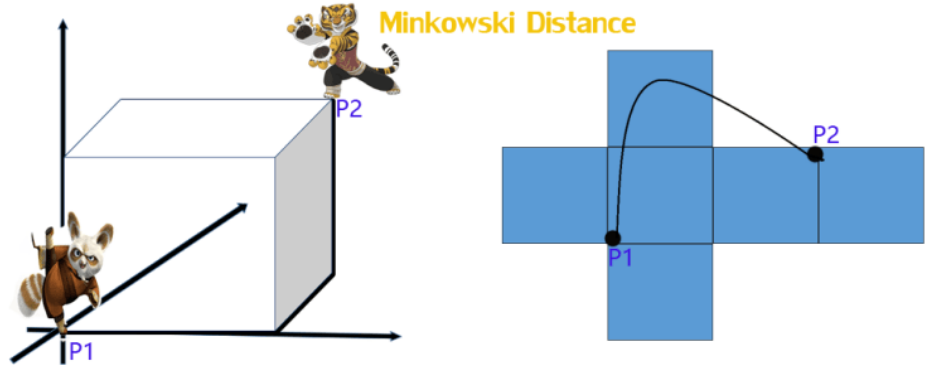
- Hierarchical Clustering
  - Agglomerative Clustering
  - Divisive Clustering
- Applications
- Density Based Clustering
- **Distance metrics**
  - Manhattan
  - **Minkowski**
  - Mahalanobis

Minkowski Distance

Minkowski Distance

- It is a generalised metric form **of Euclidean distance** and **Manhattan distance.**
- **d^MKD** - Minkowski distance between the data record.
- **i and j, k** - the index of a variable
- **n** - total number of variables y
- **λ -** order of the Minkowski metric.

$$d^{MKD}(i,j) = \sqrt[\lambda]{\sum_{k=0}^{n-1} \left| y_{i,k} - y_{j,k} \right|^{\lambda}}$$



Use Case : This distance can be used for both **ordinal** and **quantitative variables**.


I DON'T ALWAYS
LOOK FOR OUTLIERS
BUT WHEN I DO I USE
MAHALANOBIS DISTANCE

AGENDA

- What is Clustering?
- Unsupervised Learning
- Why Clustering?
- Types of Clustering
  - Partitioning Clustering
- K Means Clustering
- Challenges in K Means Clustering
- Elbow Method
- Euclidean Distance
- Illustration of K Means algorithm
- Applications of K Means

References

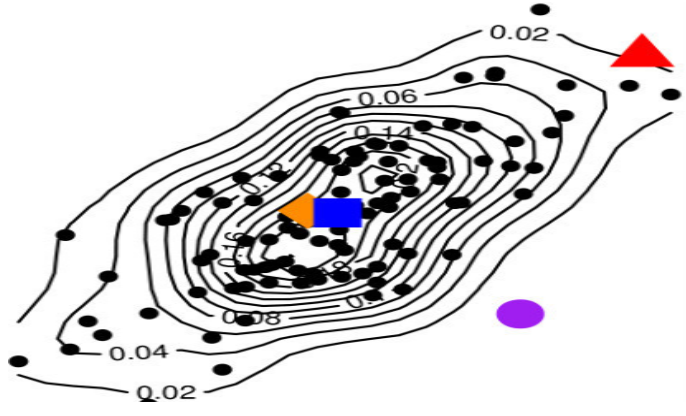
- Hierarchical Clustering
  - Agglomerative Clustering
  - Divisive Clustering
- Applications
- Density Based Clustering
- **Distance metrics**
  - Manhattan
  - Minkowski
  - **Mahalanobis**

Mahalanobis Distance

Mahalanobis Distance

- The Mahalanobis distance (MD) is **the distance between two points in multivariate space**.
- **For example**: it's fairly common to find a 6' tall woman weighing 185 lbs, but it's rare to find a 4' tall woman who weighs that much.
- In the equation below , S is the covariance matrix and x , y are vectors.

$$d_M(x,y) = \sqrt{(x-y)^T S^{-1}(x-y)}$$



Use Case : To **find multivariate outliers**, which indicates **unusual combinations** of two or more variables.

Geometric Interpretation of distance metrics