Clustering

Statistics for HCI

SUPERVISED AND UNSUPERVISED LEARNING

Supervisor: PhD.Vera Schmitt

Mohamed Mesto

Quality & Usability Lab

CONTENTS OF THE PRESENTATION

1. Introduction to Clustering Algorithms

2. Motivation of the method
    1. When use it
    2. What kind of research questions can be answered by the method

3. Theoretical background and assumptions

4. Explanation of the method

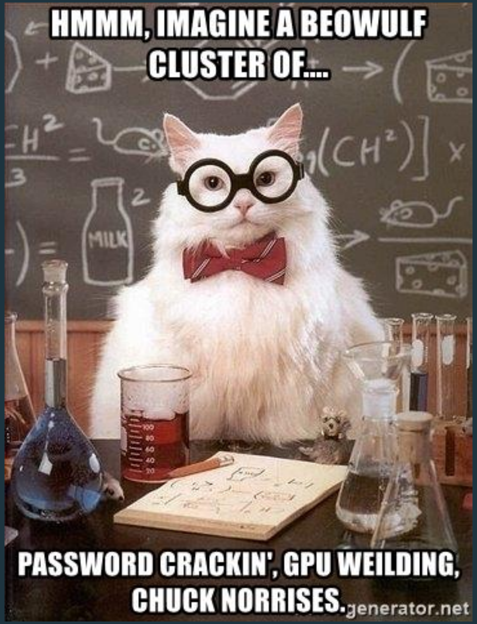
5. Exceptions and extensions

6. Implementation (Python)

INTRODUCTION TO THE CLUSTERING

01

- K-means
- Gaussian Mixture Models
- Agglomerative Clustering


HMMM, IMAGINE A BEOWULF CLUSTER OF....
PASSWORD CRACKIN', GPU WEILDING, CHUCK NORRISES.
generator.net

HOW CAN WE MAKE OUR LIFE BETTER AND LESS TIME-CONSUMING?

Towards the desire to improve human life and in conjunction with the growing requirements and needs of consumers over time, the demand has become urgent to develop and employ artificial intelligence and machine learning algorithms to achieve the aspirations of customers in intelligent life.

WHAT SUPERVISED AND UNSUPERVISED LEARNING IS ABOUT...

UNSUPERVISED LEARNING

is used to discover patterns from a provided unlabeled dataset. In this method, the algorithms are implemented without human interposition.

Clustering, Association, and Dimensionality reduction

SUPERVISED LEARNING

is an algorithm category that specifies a predictive model utilizing data points with known outcomes.

Classification, Regression

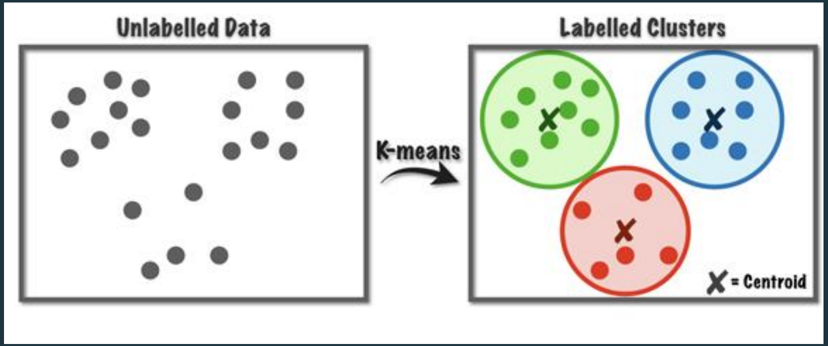
WHAT DOES THE CLUSTERING MEAN…

CLUSTERING

Clustering is an unsupervised approach applied on unlabeled datasets. It aims to collect them into combinations depending on their relationships such as
K-means, Gaussian Mixture Models, and Agglomerative Clustering

MOTIVATION OF K-MEANS

02



* https://medium.com/@luigi.fiori.lf0303/k-means-clustering-using-python-db57415d26e6

WHEN TO USE K-MEANS?

- Unsupervised method
- Primarily, used in data mining and statistics
- performs the gathering/clustering of unlabeled datasets into groups

WHICH RQS CAN BE ANSWERED?
WHICH FIELDS CAN BE SUPPORTED?

Use cases:

- Marketing/customer segmentation
- Document clustering
- Image segmentation

THEORETICAL BACKGROUND AND ASSUMPTIONS

03

ASSUMPTIONS

THE NUMBER OF THE CLUSTERS (K)

The number of the Clusters

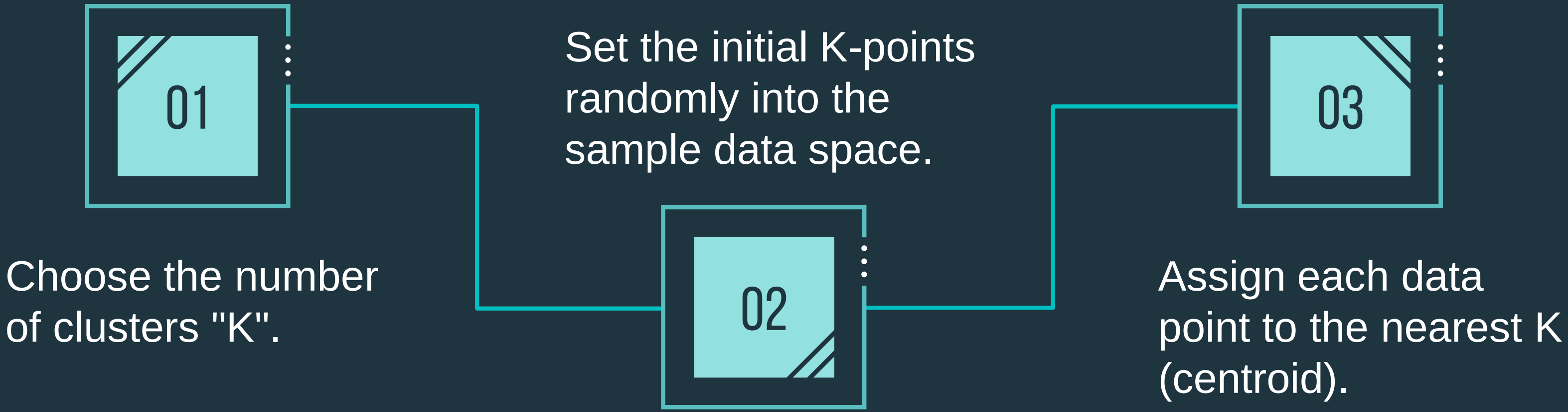
INITIAL K-POINTS (CENTROIDS)

randomly into the sample data space

EXPLANATION OF THE K-MEANS METHOD

04

METHODOLOGY



01

Choose the number of clusters "K".

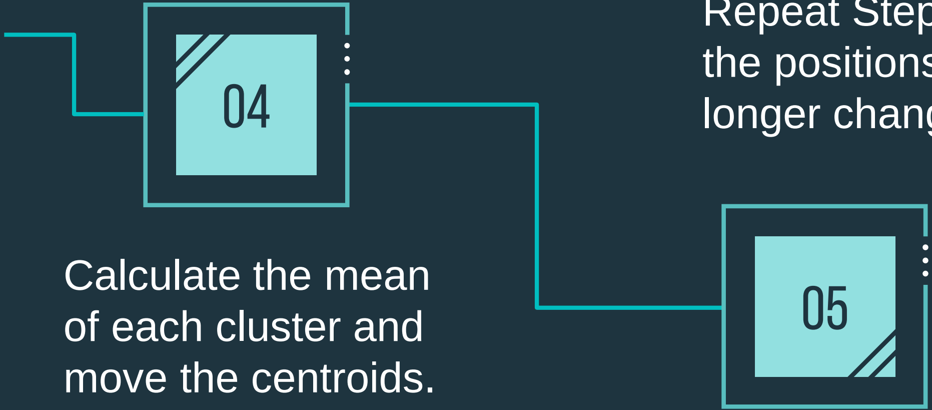
02

Set the initial K-points randomly into the sample data space.

03

Assign each data point to the nearest K (centroid).

METHODOLOGY



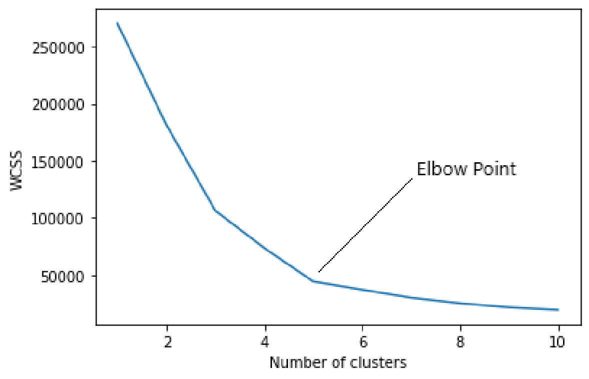
04

Calculate the mean of each cluster and move the centroids.

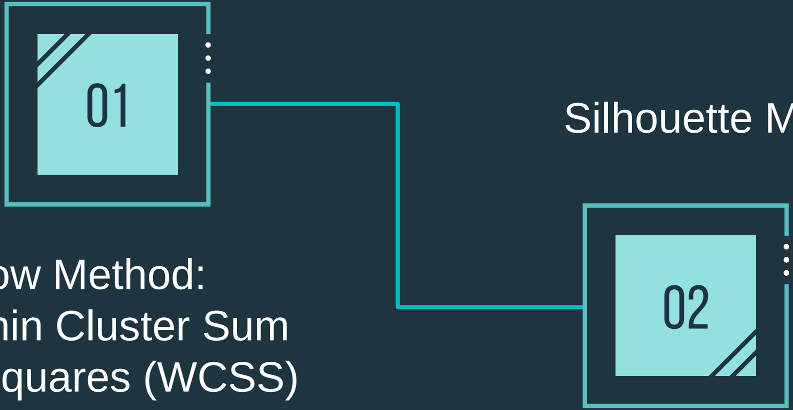
05

Repeat Steps 3 & 4 continuously until the positions of the centroids are no longer changed.

HYPERPARAMETER TUNING: CHOOSING OPTIMAL "K"



- https://www.analyticsvidhya.com/blog/2021/01/in-depth-intuition-of-k-means-clustering-algorithm-in-machine-learning/



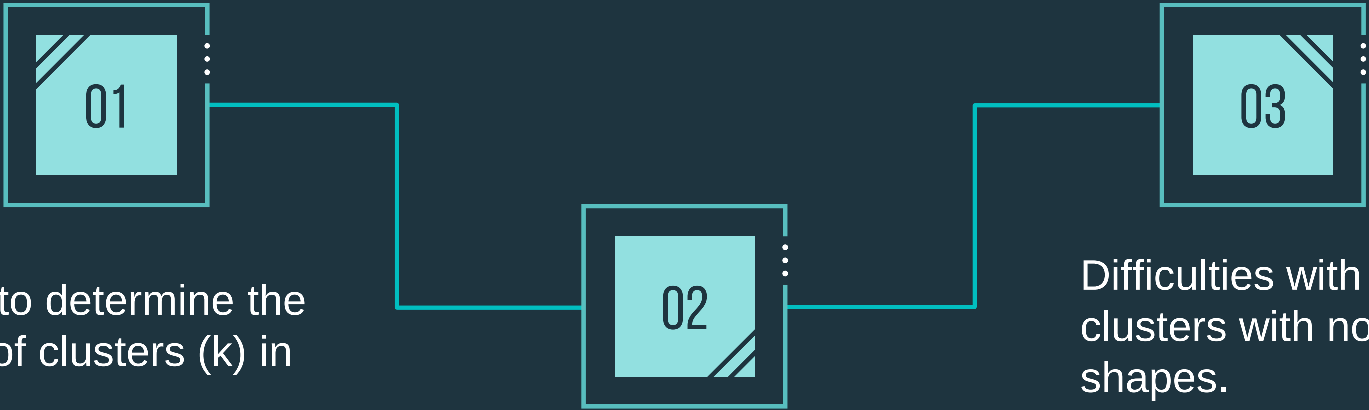
Elbow Method:
Within Cluster Sum
of Squares (WCSS)

Silhouette Method

EXCEPTIONS AND EXTENSIONS

05

K-MEANS DISADVANTAGES:



It needs to determine the number of clusters (k) in advance.

Unable to deal with noisy data and outliers.

Difficulties with specifying clusters with non-convex shapes.

IMPLEMENTATION (PYTHON)

06

K-MEANS IMPLEMENTATION PART

1. FUNDAMENTALS

Packages and setup

2. DATA PREPARATION

Data cleaning steps and required data preparation

3. IMPLEMENTATION OF METHOD

Describing all implementation steps

4. INTERPRETATION

Interpreting the results and all output elements

K-MEANS IMPLEMENTATION PART

1. FUNDAMENTALS

Packages and setup

REQUIRED PYTHON LIBRARIES

- NumPy: for scientific computing.
- Matplotlib: a plotting library for Python.
- Matplotlib.pyplot: functions that allow matplotlib to work like MATLAB.
- Pandas: used for data science/data analysis.
- Sklearn.mixture:

K-MEANS IMPLEMENTATION PART

2. DATA PREPARATION

Data cleaning steps and required data preparation

- Irregular column name
- Imbalanced data sets
- Missing data
- Duplicate rows
- Overlapping
- Untidy
- Density: Shortage of data
- Noise

DIAGNOSE DATA FOR CLEANING

- Preprocessing the Dataset
- Correction/Clarification of the Dataset Columns' name
- Correction of the data values of the Dataset

K-MEANS IMPLEMENTATION PART

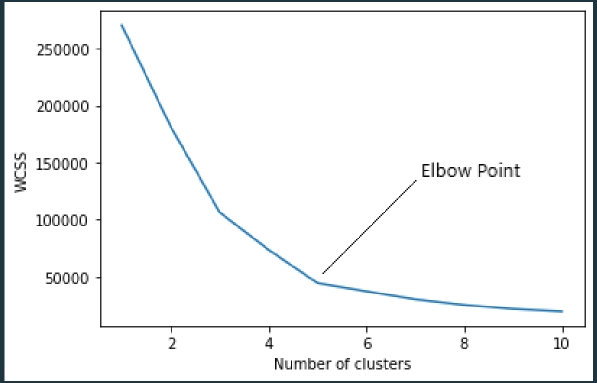
3. IMPLEMENTATION OF METHOD

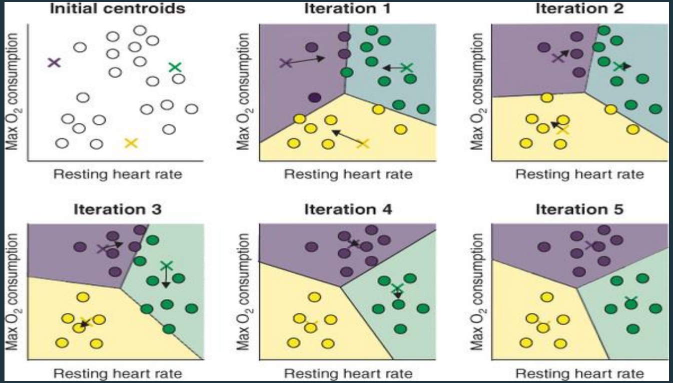
- Reading the Dataset
- Using Dependent variables
- Correction the of data values of the Dataset
- Splitting the dataset into the Training set and Test set
- Using the elbow method to find the optimal number of clusters
- Training the K-Means model on the dataset
- Visualising the clusters

K-MEANS IMPLEMENTATION PART

4. INTERPRETATION

Interpreting the results and all output elements





- https://www.analyticsvidhya.com/blog/2021/01/in-depth-intuition-of-k-means-clustering-algorithm-in-machine-learning/
- https://livebook.manning.com/book/machine-learning-for-mortals-mere-and-otherwise/chapter-16/21

INTRODUCTION TO THE GMMS:

GAUSSIAN MIXTURE MODELS

01

WHAT ARE THE GAUSSIAN MIXTURE MODELS GMMS?

- What makes GMMs a better candidate than K-means?

WHAT ARE THE GAUSSIAN MIXTURE MODELS IS ABOUT…

GAUSSIAN DISTRIBUTION (GD)

- What does it relate to the Gaussian Distribution?

EXPECTATION-MAXIMIZATION (EM)

- What is the Expectation-Maximization Algorithm (EM)?

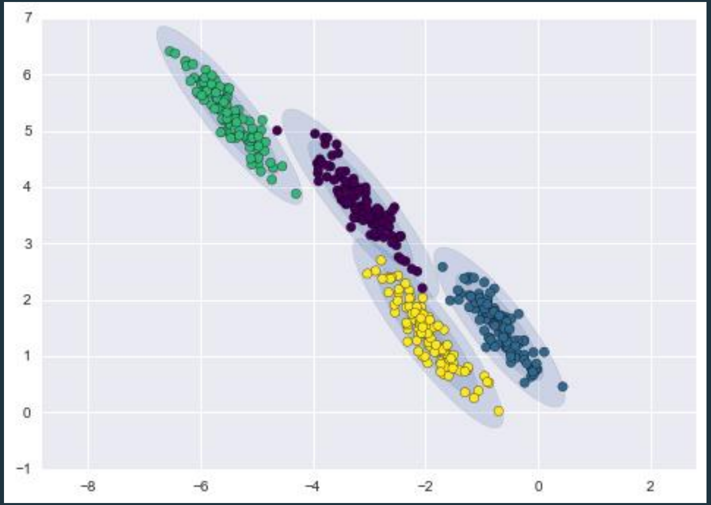
WHAT DOES GMMS MEAN…

THE GAUSSIAN MIXTURE MODELS GMMS

Gaussian Mixture Models (GMMs) is one of the most famous clustering algorithms. It uses the Gaussian, which is a method for plotting data. However, it differs from the K-mean algorithm because it considers variance.

MOTIVATION OF GMMS

02



https://jakevdp.github.io/PythonDataScienceHandbook/05.12-gaussian-mixtures.html

WHEN TO USE GMMS?

- Unsupervised method
- Considering Mean and Variance
- performs the gathering/clustering of unlabeled datasets into groups

WHICH RQS CAN BE ANSWERED? WHICH FIELDS CAN BE SUPPORTED?

Use cases of GMMs:

- Clustering and density estimation in physics
- Modeling weather observations in geoscience (Zi, 2011) clustering
- Certain autoregressive models
- Noise from some time series.

* https://www.statisticshowto.com/gaussian-mixture-model/

THEORETICAL BACKGROUND AND ASSUMPTIONS

03

ASSUMPTIONS

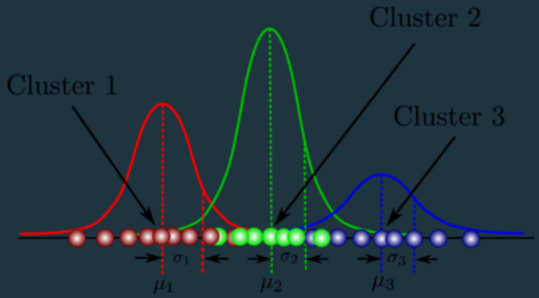
GAUSSIAN DISTRIBUTION

EXPECTATION-MAXIMIZATION (EM)

Before diving into Gaussian Mixture Models, let us look at the "Gaussian Distribution“ and Expectation-Maximization (EM)

GAUSSIAN DISTRIBUTION

It is also known as Normal Distribution. Mean (μ),variance (σ2).



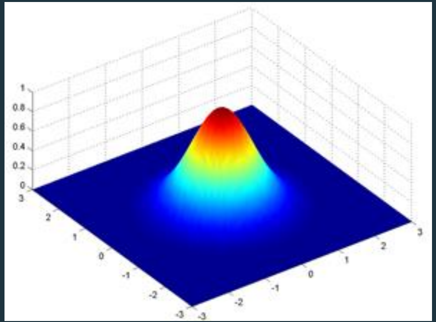
- https://towardsdatascience.com/gaussian-mixture-models-explained-6986aaf5a95

01

It is a bell-shaped curve with the data points harmoniously dispersed around the mean value.

02

The curve's shape will be a 3D bell curve as displayed below:



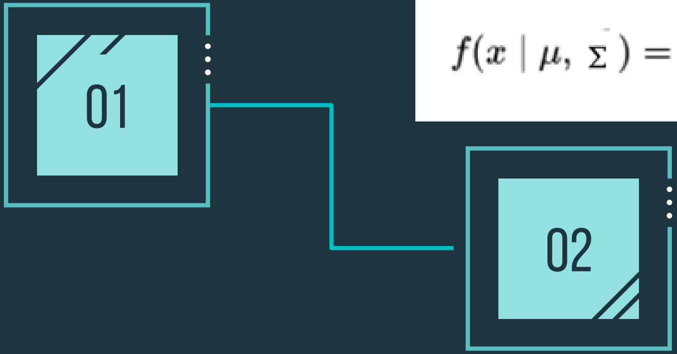
- https://towardsdatascience.com/gaussian-mixture-models-explained-6986aaf5a95

GAUSSIAN DISTRIBUTION

For the Gaussian distribution's probability density function, we distinguish the following cases: Mean (μ),variance (σ2).

In one-dimensional space:

$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



In two-dimensional space:

$$f(x \mid \mu, \Sigma) = \frac{1}{\sqrt{2\pi|\Sigma|}} \exp\left[-\frac{1}{2} (\boldsymbol{x}-\boldsymbol{\mu})' \Sigma^{-1} (\boldsymbol{x}-\boldsymbol{\mu}) \right]$$

Where:
x: describes the input vector.
μ: represents the 2D mean vector.
Σ: defines the 2×2 covariance matrix.

- https://towardsdatascience.com/gaussian-mixture-models-explained-6986aaf5a95

GAUSSIAN DISTRIBUTION

In a d-dimensional space (multivariate Gaussian model):

The general rule: The method result will be a combination or mixture of k Gaussian distributions if the input is a dataset of d features

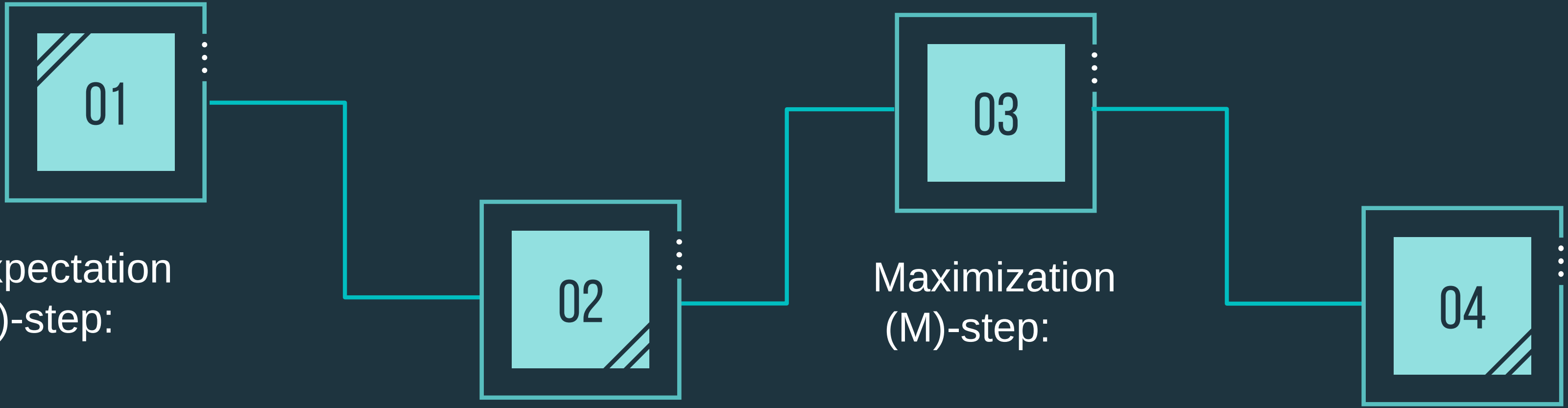
03

Where:
x: describes the input vector.
μ: represents the 2D mean vector.
Σ: defines the 2×2 covariance matrix.
Where:
x , μ as vectors of length d.
Σ: defines the d×d covariance matrix. it is also possible to generate the equation!
k is equal to that cluster number.

- https://www.analyticsvidhya.com/blog/2019/10/gaussian-mixture-models-clustering/

EXPECTATION-MAXIMIZATION ALGORITHM (EM)

an iterative method to find the suitable model parameters by accomplishing maximum likelihood estimation.



Expectation (E)-step:

Optimizing the model:

Maximization (M)-step:

Repeating Steps 2, 3, and 4 until steadiness is achieved

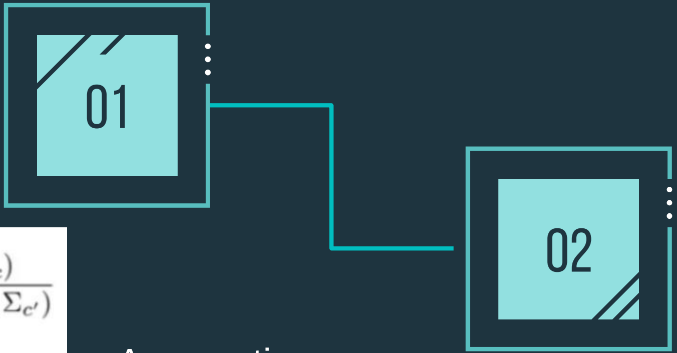
- https://www.analyticsvidhya.com/blog/2019/10/gaussian-mixture-models-clustering/
- https://en.wikipedia.org/wiki/Expectation%E2%80%93maximization_algorithm

EXPECTATION-MAXIMIZATION (EM) IN GAUSSIAN MIXTURE MODELS (GMMS)

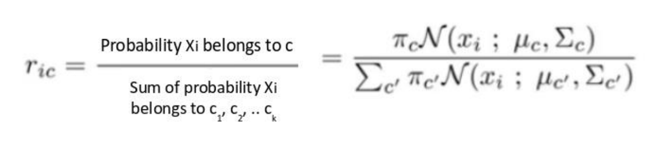
After understanding the EM algorithm, let us use it in GMMs.
To compute the GMMs, we need to find the values of the variables μ, Σ, and Π.

EXPECTATION-MAXIMIZATION (EM) IN GAUSSIAN MIXTURE MODELS (GMMS)

Expectation (E)-step:



Optimizing the model
update the μ, Σ ,and Π values using the following formulas in next slide



$$r_{ic} = \frac{\text{Probability } x_i \text{ belongs to c}}{\text{Sum of probability } x_i \text{ belongs to } c_1, c_2, .. c_k} = \frac{\pi_c \mathcal{N}(x_i \; ; \; \mu_c, \Sigma_c)}{\sum_{c'} \pi_{c'} \mathcal{N}(x_i \; ; \; \mu_{c'}, \Sigma_{c'})}$$

Assumptions
k: is the number of clusters => k Gaussian distributions.
Mean Values: μ1, μ2, .. μk
Covariance values : Σ1, Σ2, .. Σk
Πi: is the density of the distribution.

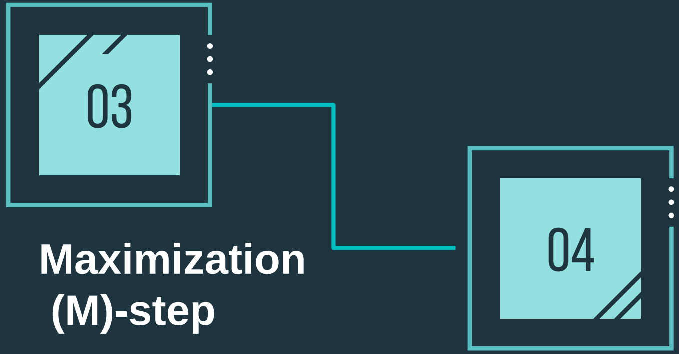
- https://towardsdatascience.com/gaussian-mixture-models-explained-6986aaf5a95

EXPECTATION-MAXIMIZATION (EM) IN GAUSSIAN MIXTURE MODELS (GMMS)

$$\Pi = \frac{\text{Number of points assigned to cluster}}{\text{Total number of points}}$$

$$\mu = \frac{1}{\text{Number of points assigned to cluster}} \Sigma_i \, r_{ic} x_i$$

$$\Sigma_c = \frac{1}{\text{Number of points assigned to cluster}} \Sigma_i \, r_{ic} \, (x_i - \mu_c)^T \, (x_i - \mu_c)$$

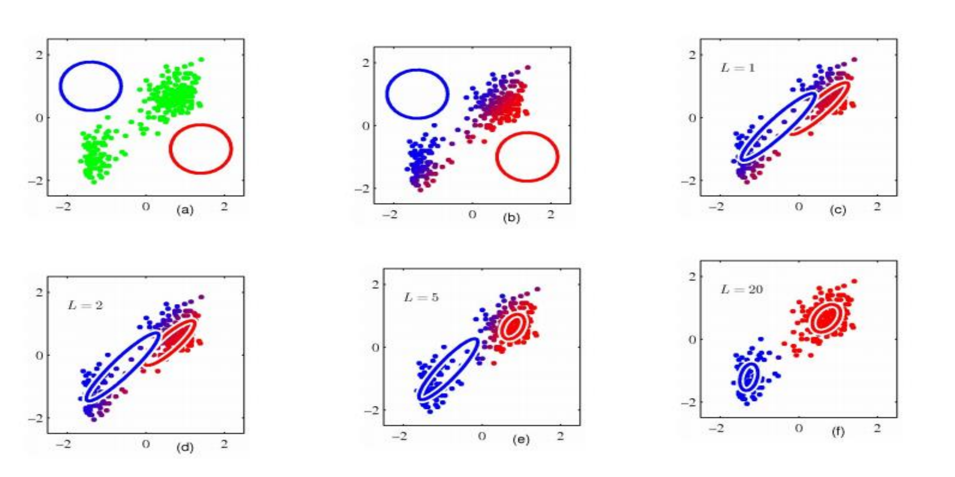


Maximization (M)-step

Repeating Steps 2, 3, and 4: Iterate until steadiness is achieved

- Text Clustering, K-Means, Gaussian Mixture Models, Expectation- Maximization, Hierarchical Clustering Sameer Maskey Week 3, Sept 19, 2012

AN ILLUSTRATION OF EXPECTATION MAXIMIZATION ALGORITHM (EM)



1 Text Clustering, K-Means, Gaussian Mixture Models, Expectation-Maximization, Hierarchical Clustering Sameer Maskey Week 3, Sept 19, 2012

EXPLANATION OF THE GMMS METHOD

04

METHODOLOGY

When apply the k-mean algorithm to a dataset, and we have 3 clusters, then each point belongs to a particular cluster. So the following cases are considered a challenge to the k-average algorithm:

01

1. If data points belong to the first cluster with a certain probability and to the second cluster with another probability.

2. If we have an overlap between two clusters within the data space, because the data points belong to two different types of data (a mixture)

02

METHODOLOGY

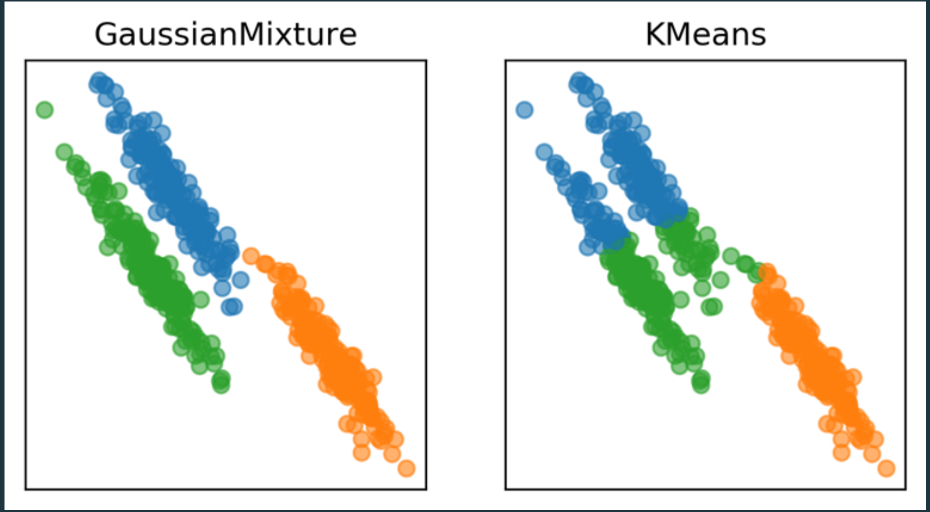
The applicable solution for them is GMMs. because GMMs are probabilistic models and use the soft clustering approach for distributing the points in different clusters.

EXCEPTIONS AND EXTENSIONS

05

GAUSSIAN MIXTURE MODELS (GMMS) VS K-MEANS

k-means considers only the mean to update the centroid while GMMs takes into account the mean as well as the variance of the data. Therefore, the k-means is called a hard assignment.



- https://jakevdp.github.io/PythonDataScienceHandbook/05.12-gaussian-mixtures.html

IMPLEMENTATION (PYTHON)

06

GMMS IMPLEMENTATION PART

1. FUNDAMENTALS

Packages and setup

3. IMPLEMENTATION OF METHOD

Describing all implementation steps

2. DATA PREPARATION

Data cleaning steps and required data preparation

4. INTERPRETATION

Interpreting the results and all output elements

GMMS IMPLEMENTATION PART

1. FUNDAMENTALS

Packages and setup

REQUIRED PYTHON LIBRARIES

- NumPy: for scientific computing.
- Matplotlib: a plotting library for Python.
- Matplotlib.pyplot: functions that allow matplotlib to work like MATLAB.
- Pandas: used for data science/data analysis.
- Sklearn.mixture:

GMMS IMPLEMENTATION PART

2. DATA PREPARATION

Data cleaning steps and required data preparation

- Imbalanced data sets
- Missing data
- Duplicate rows
- Overlapping
- Untidy
- Density: Shortage of data
- Noise

DIAGNOSE DATA FOR CLEANING

- Preprocessing the Dataset
- Correction/Clarification of the Dataset Columns' name
- Correction of the data values of the Dataset

GMMS IMPLEMENTATION PART

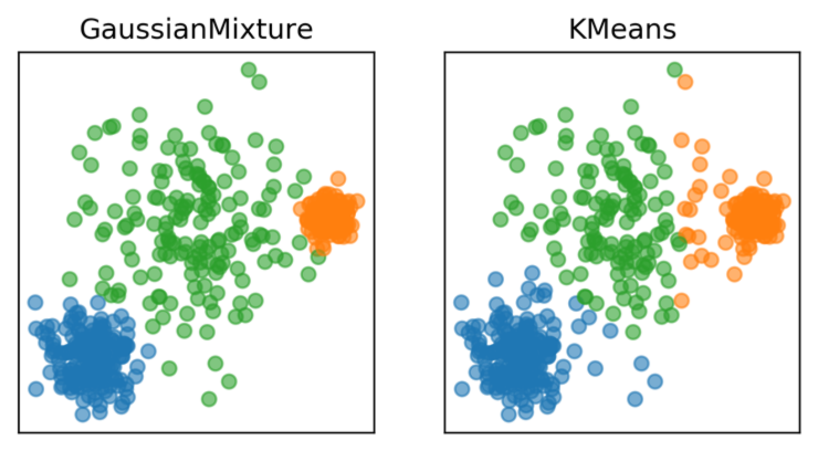
3. IMPLEMENTATION OF METHOD

- Reading the Dataset
- Using Dependent variables
- Correction the of data values of the Dataset
- Test for Normal Distribution
- Training the GMMs model on the dataset
- E- step
- M-step
- Visualising the clusters

GMMS IMPLEMENTATION PART

4. INTERPRETATION

Interpreting the results and all output elements



https://jakevdp.github.io/PythonDataScienceHandbook/05.12-gaussian-mixtures.html

INTRODUCTION TO HIERARCHICAL CLUSTERING:

AGGLOMERATIVE CLUSTERING

01

WHAT IS THE AGGLOMERATIVE CLUSTERING ALGORITHM?

- What are the similarities and differences between it and the K-means?

WHAT DOES HC OR HCA MEAN...

HIERARCHICAL CLUSTERING (HC)

Hierarchical Clustering (HC) or Hierarchical Clustering Analysis (HCA) is a clustering algorithm used in statistical analysis. It aims to analyze and plot the studying data and present the clusters in a hierarchy diagram.

Dendrogram Diagram

WHAT ARE THE HIERARCHICAL CLUSTERING (HC) IS ABOUT…

AGGLOMERATIVE CLUSTERING

- A "bottom-up" method
- Initially, each data point is a cluster of its own

DIVISIVE CLUSTERING

- A "top-down" method
- initially, all the data points in the dataset belong to one cluster.

- https://en.wikipedia.org/wiki/Hierarchical_clustering
- https://towardsdatascience.com/hierarchical-clustering-agglomerative-and-divisive-explained-342e6b20d710

MOTIVATION OF HC

02

WHEN TO USE HC?

Advantages of AHC:

- Easy to implement, object ordering, , and informative for the display.
- No need for pre-specify the number of clusters.
- Easy to decide the number of clusters by cutting the Dendrogram at the specific level.
- The ability of AHC approach to create smaller clusters , which may uncover similarities in data.

WHICH RQS CAN BE ANSWERED?
WHICH FIELDS CAN BE SUPPORTED?

Use cases of *Agglomerative* HC:

- **Analyze social network data**

* https://www.sciencedirect.com/topics/computer-science/hierarchical-clustering

03

THEORETICAL BACKGROUND AND ASSUMPTIONS

04

EXPLANATION OF THE HC METHOD

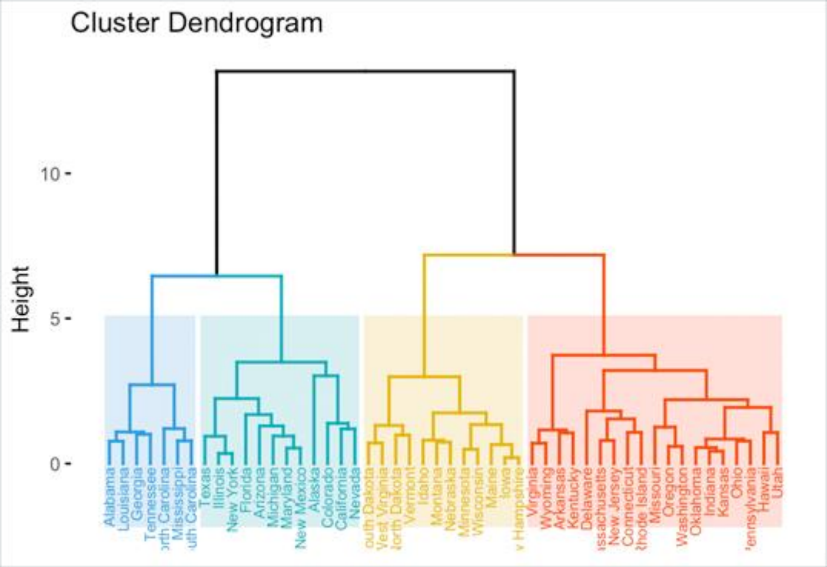
AGGLOMERATIVE CLUSTERING ALGORITHM METHODOLOGY:

Initially, We should consider that every data point represents a cluster on its own.

Every two nearest clusters could be joined with each other to form one single cluster.

Repeat step 2 until the whole number of clusters are included.

AN EXAMPLES FOR AGGLOMERATIVE CLUSTERING



* https://www.datanovia.com/en/lessons/examples-of-dendrograms-visualization/

EXCEPTIONS AND EXTENSIONS

05

AGGLOMERATIVE CLUSTERING VS K-MEANS

AGGLOMERATIVE CLUSTERING

K-MEANS

HC and K-means are both clustering algorithms in the statistical analysis field.

- no need to appoint the number of clusters in advance.
- The determines are based on previous opinions. **HC** should be used to know the number of clusters.

- using the centroid
- calculating the distances between the data points.

- https://en.wikipedia.org/wiki/Hierarchical_clustering
- https://towardsdatascience.com/hierarchical-clustering-agglomerative-and-divisive-explained-342e6b20d710

AGGLOMERATIVE CLUSTERING VS K-MEANS

AGGLOMERATIVE CLUSTERING

- The demand is high to determine the number of clusters more easily. HC's **dendrogram** is the right decision.
- It is more enlightening and interpretable.

K-MEANS

- The provided dataset has a particular number of clusters, but they belong to an unknown group.
- For fast computing , When the provided dataset has a large number of variables.

- https://en.wikipedia.org/wiki/Hierarchical_clustering
- https://towardsdatascience.com/hierarchical-clustering-agglomerative-and-divisive-explained-342e6b20d710

IMPLEMENTATION (PYTHON)

06

AGGLOMERATIVE CLUSTERING IMPLEMENTATION PART

1. FUNDAMENTALS

Packages and setup

2. DATA PREPARATION

Data cleaning steps and required data preparation

3. IMPLEMENTATION OF METHOD

Describing all implementation steps

4. INTERPRETATION

Interpreting the results and all output elements

AGGLOMERATIVE CLUSTERING IMPLEMENTATION PART

1. FUNDAMENTALS

Packages and setup

REQUIRED PYTHON LIBRARIES

- NumPy: for scientific computing.
- Matplotlib: a plotting library for Python.
- Matplotlib.pyplot: functions that allow matplotlib to work like MATLAB.
- Pandas: used for data science/data analysis.
- Sklearn.mixture:

AGGLOMERATIVE CLUSTERING IMPLEMENTATION PART

2. DATA PREPARATION

Data cleaning steps and
required data preparation

DIAGNOSE DATA FOR CLEANING

- Preprocessing the Dataset
- Correction/Clarification of the Dataset Columns' name
- Correction of the data values of the Dataset

AGGLOMERATIVE CLUSTERING IMPLEMENTATION PART

3. IMPLEMENTATION OF METHOD

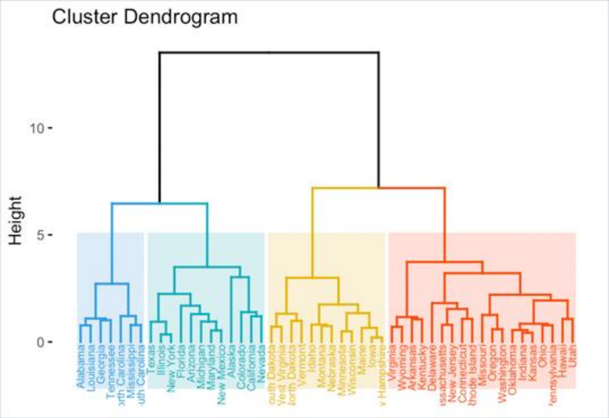
- Reading the Dataset
- Using Dependent variables
- Correction the of data values of the Dataset
- Using the Dendrogram to find the optimal number of clusters
- Training the Hierarchical Clustering model on the dataset
- Visualising the clusters

* Machine Learning A-Z™: Hands-On Python & R In Data Science

AGGLOMERATIVE CLUSTERING IMPLEMENTATION PART

4. INTERPRETATION

Interpreting the results and all output elements



* https://www.datanovia.com/en/lessons/examples-of-dendrograms-visualization/

SOME ADDITIONAL HINTS

00

The references are as APA Style are included in the notebook Colab Theory and app parts and the report able to exported as HTML and PDF file in Latex template