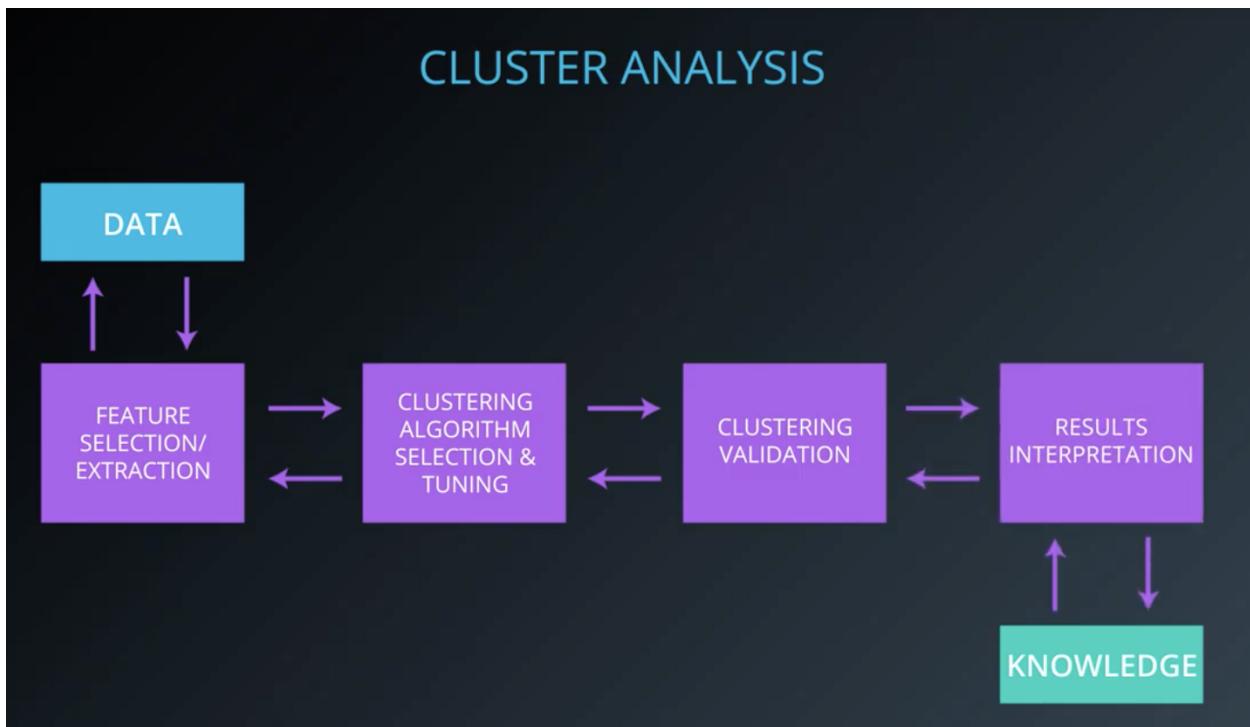


Udacity Nanodegree — Clustering

Process for clustering



K-means Clustering

Hierarchical Clustering

HIERARCHICAL CLUSTERING

ADVANTAGES:

- Resulting hierarchical representation can be very informative
- Provides an additional ability to visualize
- Especially potent when the dataset contains real hierarchical relationships (e.g. Evolutionary biology)

DISADVANTAGES:

- Sensitive to noise and outliers
- Computationally intensive $O(N^2)$

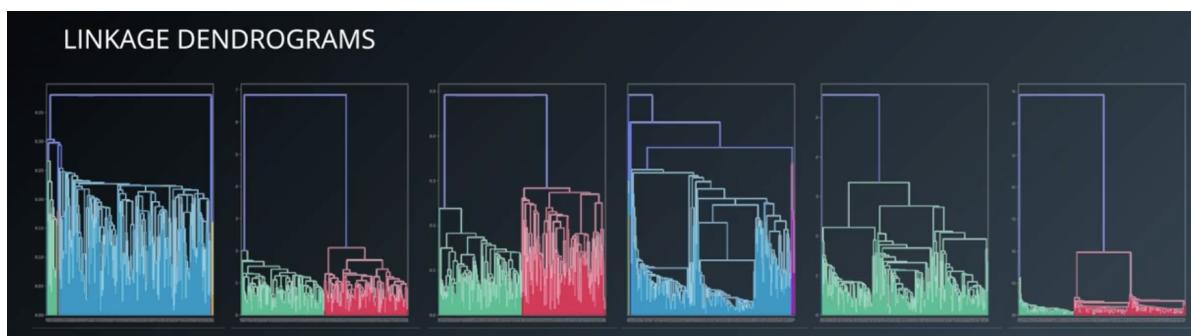
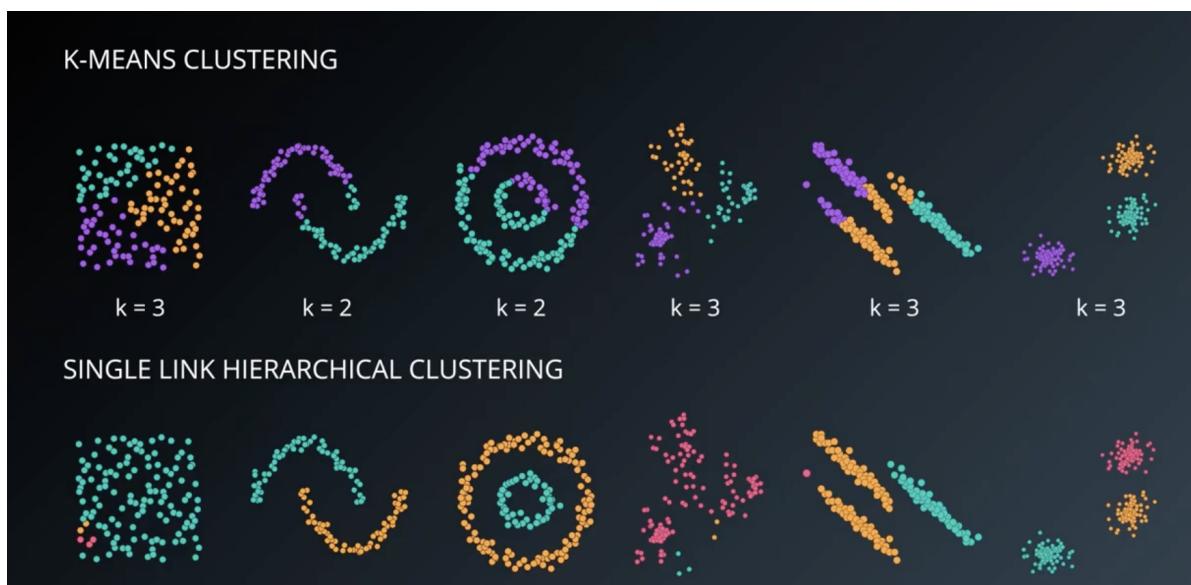
Application

Paper: [Using Hierarchical Clustering of Secreted Protein Families to Classify and Rank Candidate Effectors of Rust Fungi](#)

Paper: [Association between composition of the human gastrointestinal microbiome and development of fatty liver with choline deficiency](#)

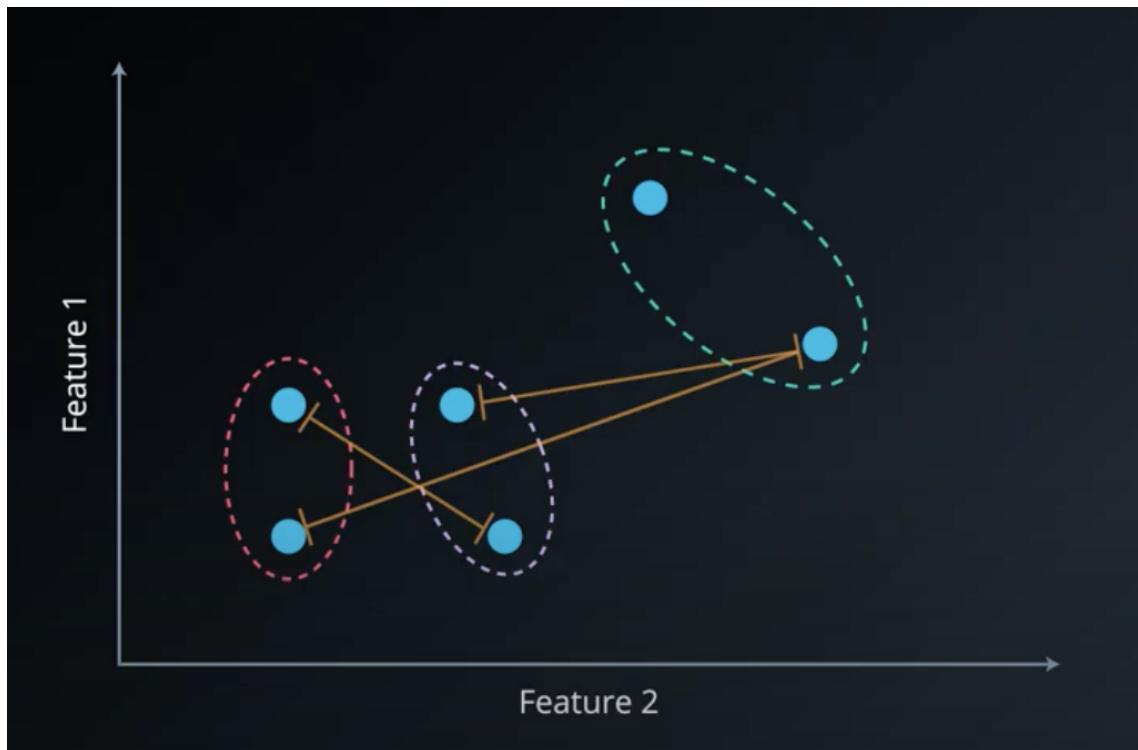
▼ Hierarchical Clustering - Single Link Clustering

1. Look at the closest point for a point(1-2, 4-5)
2. If there is already clusters, find the closest point between two clusters to determine which cluster to join. e.g. 7 & 6+8, closest to 6, belong to 6+8



▼ Hierarchical Clustering - Complete Link Clustering

1. Look at the closest point for a point(1-2, 4-5)
2. If there is already clusters, find the farthest point between two clusters to determine which cluster to join. e.g. 7 & 6+8, the farthest is 8, belong to 6+8

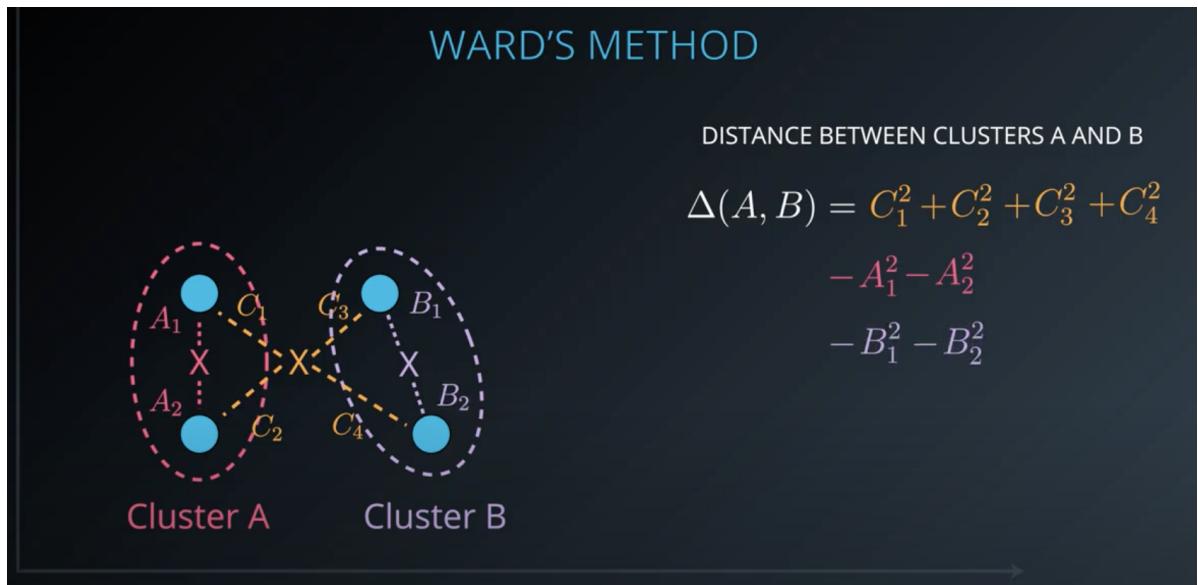


▼ Hierarchical Clustering - Average Link Clustering

1. Look at the closest point for a point(1-2, 4-5)
2. If there is already clusters, find the average distance for every point between two clusters to determine which cluster to join. e.g. 7 & 6+8, the farthest is 8, belong to 6+8



▼ Hierarchical Clustering - Ward's method



Density-based clustering

DBSCAN clusters points based on density(stamps) and in that process it identifies noise.

DENSITY-BASED CLUSTERING | DBSCAN

ADVANTAGES:

- We don't need to specify the number of clusters
- Flexibility in the shapes & sizes of clusters
- Able to deal with noise
- Able to deal with outliers

DISADVANTAGES:

- Border points that are reachable from two clusters
- Faces difficulty finding clusters of varying densities

- For border problem, if a point os reachable from two clusters, the clustering become arbitrary
- For varying densities, we can use HDBSCAN as a alternative

Application

Paper: [Traffic Classification Using Clustering Algorithms](#)

Paper: [Anomaly detection in temperature data using dbSCAN algorithm](#)

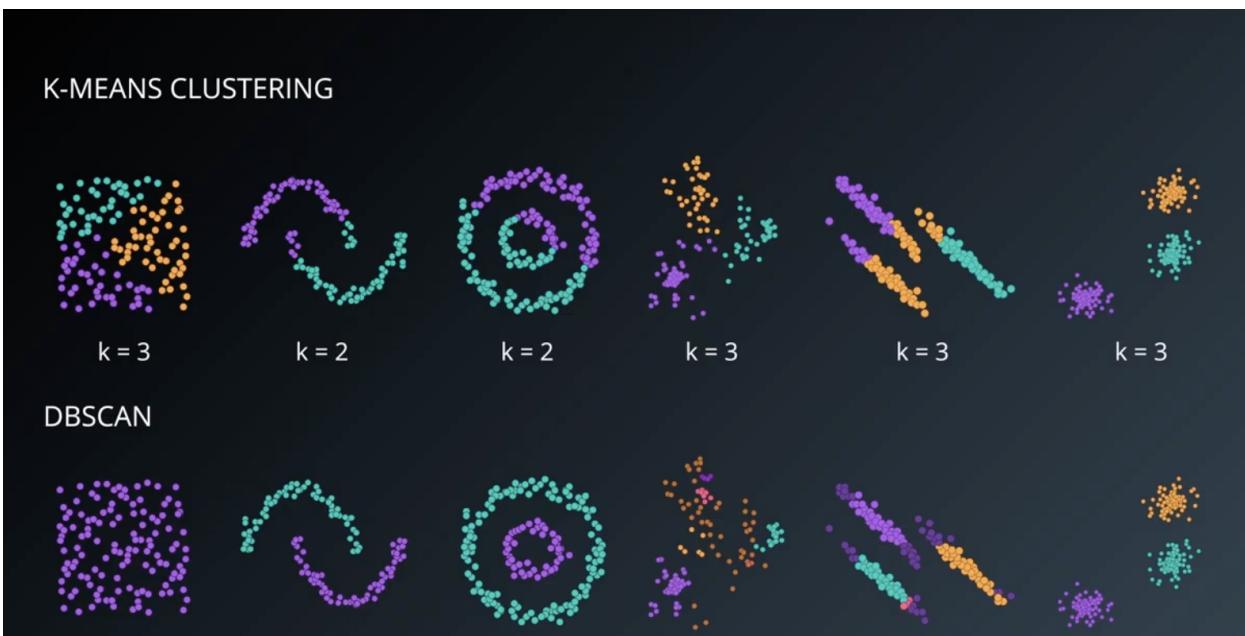
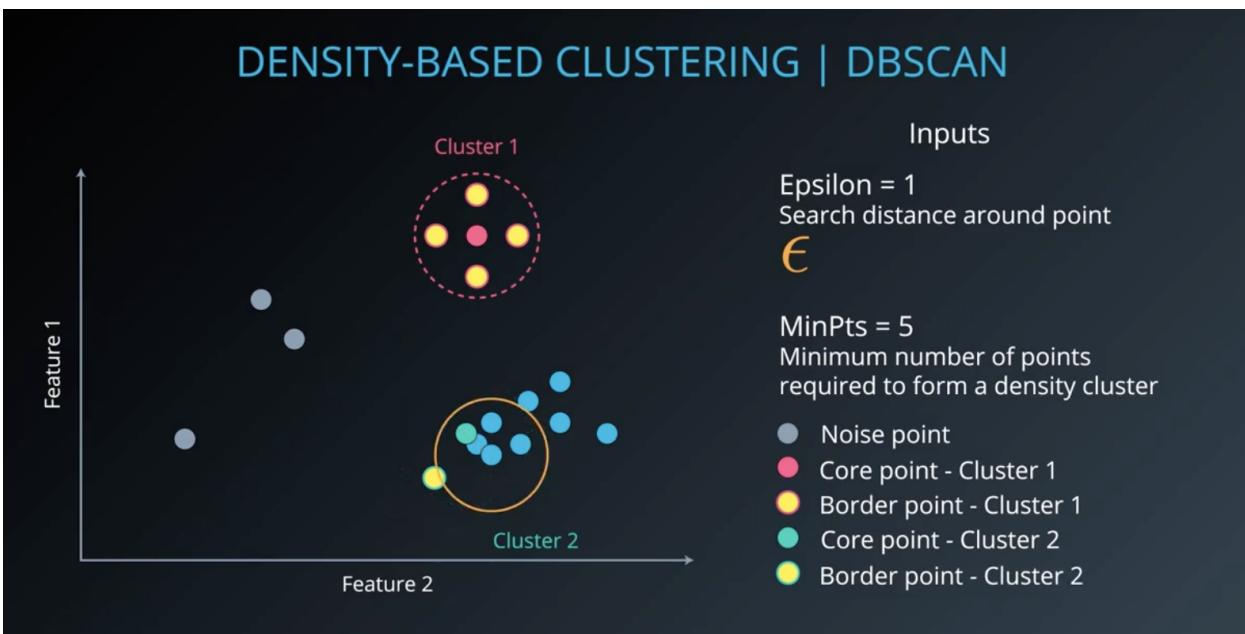
Paper: [Hierarchical density based clustering](#)

Variables

eps: distance for a cluster(radius from the center)

min_samples: min samples need to include to form a cluster

DENSITY-BASED CLUSTERING | DBSCAN



Gaussian Mixture Model Clustering (GMM)

Assume each cluster following a normal distribution

GAUSSIAN MIXTURE MODEL CLUSTERING

Advantages:

- Soft-clustering (sample membership of multiple clusters)
- Cluster shape flexibility

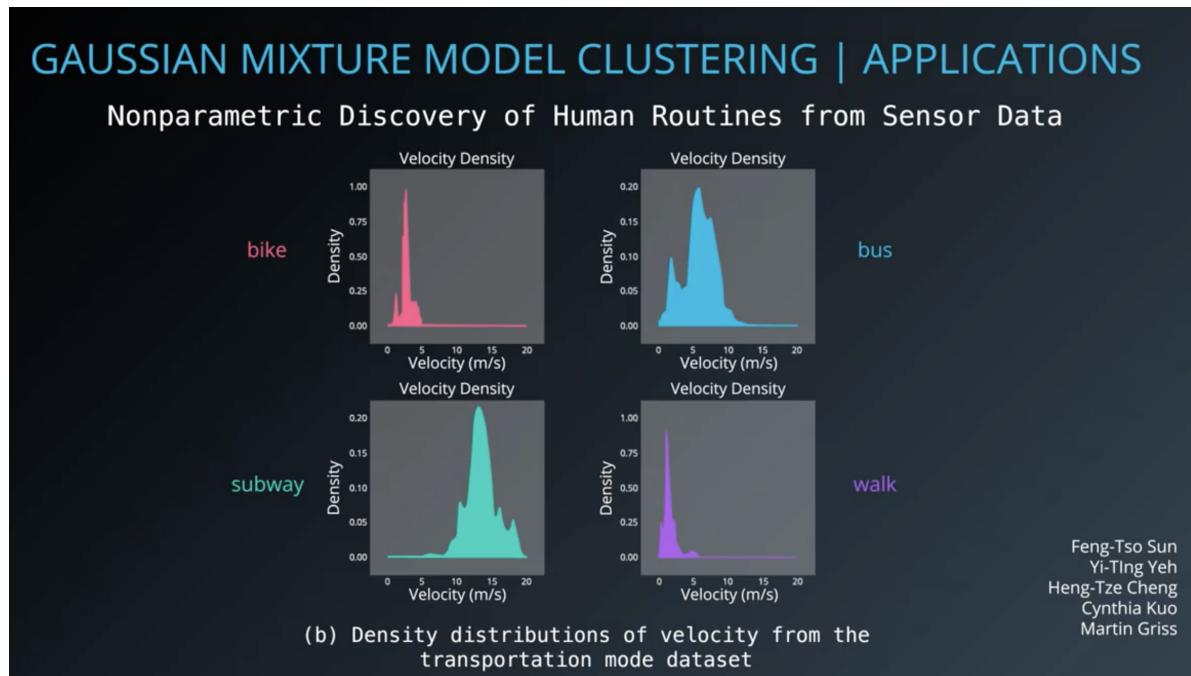
Disadvantages:

- Sensitive to initialization values
- Possible to converge to a local optimum
- Slow convergence rate

soft clustering: Fuzzy clustering (also referred to as soft clustering or soft k-means) is a form of clustering in which each data point can belong to more than one cluster.

Application

- ▼ Paper: [Nonparametric discovery of human routines from sensor data](#) [PDF]



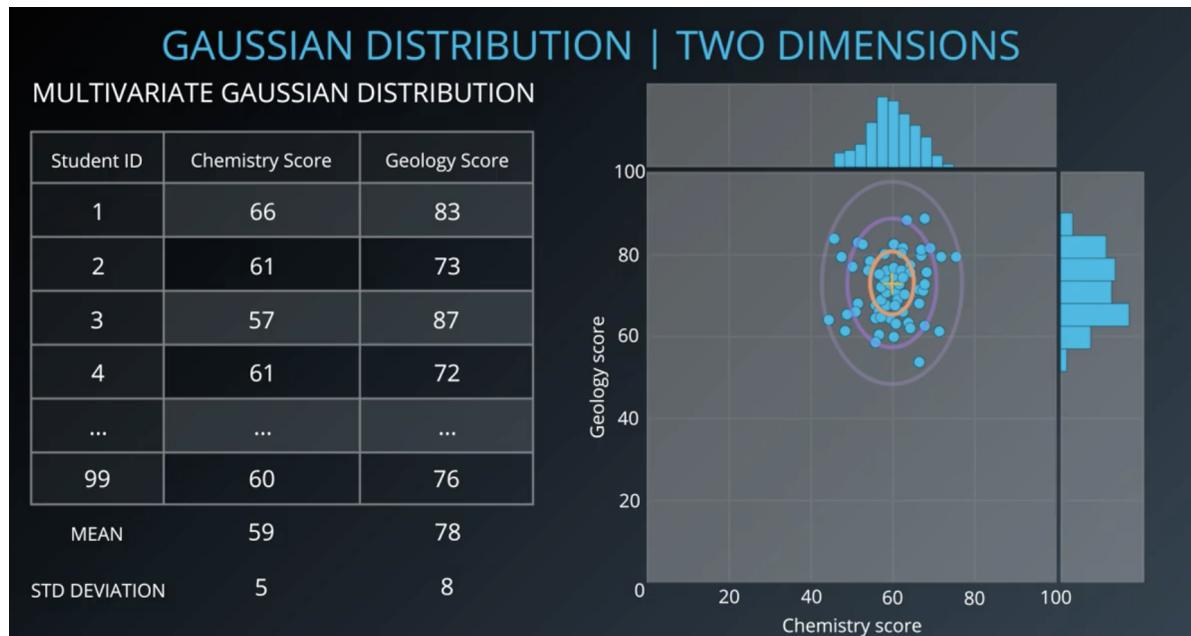
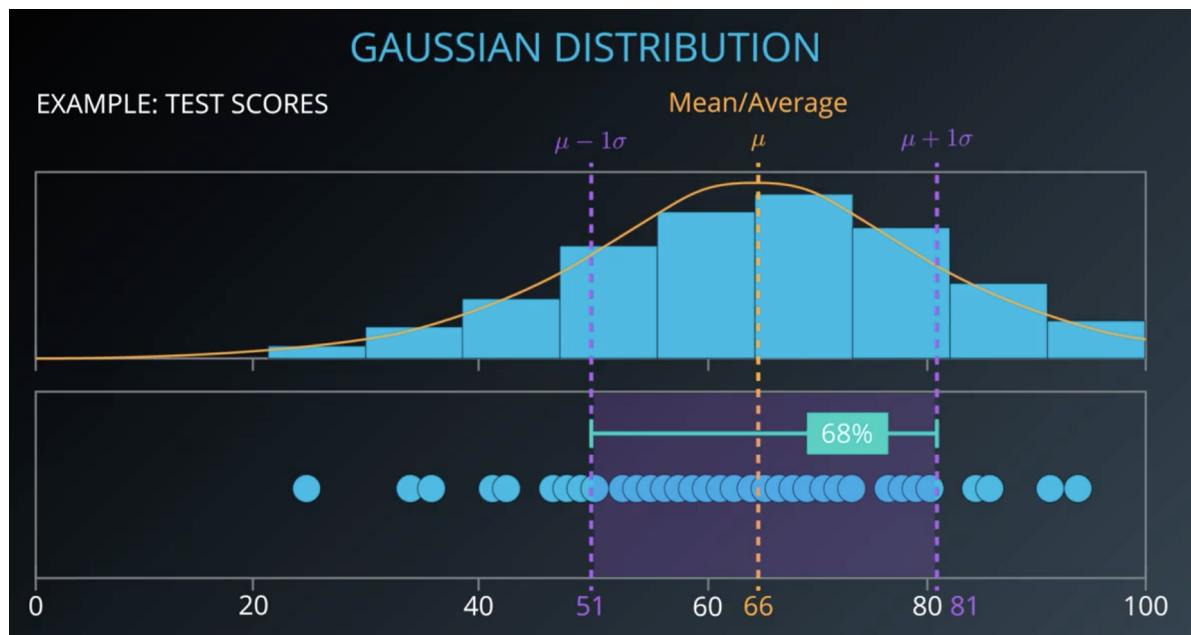
Paper: [Application of the Gaussian mixture model in pulsar astronomy](#) [PDF]

Paper: [Speaker Verification Using Adapted Gaussian Mixture Models](#) [PDF]

Paper: [Adaptive background mixture models for real-time tracking](#) [PDF]

Background subtraction Video: <https://www.youtube.com/watch?v=ILt9H6RFO6A>

- ▼ Gaussian distribution(Normal distribution)



▼ Steps for GMM

GAUSSIAN MIXTURE MODEL CLUSTERING | EXAMPLE

Expectation - Maximization For Gaussian Mixtures:



Dataset to cluster into two clusters

STEP #1: INITIALIZE K GAUSSIAN DISTRIBUTIONS

STEP #2: SOFT-CLUSTER DATA - "EXPECTATION"

STEP #3: RE-ESTIMATE THE GAUSSIANS - "MAXIMIZATION"

STEP #4: EVALUATE LOG-LIKELIHOOD TO CHECK FOR CONVERGENCE

REPEAT FROM STEP #2 UNTIL CONVERGED

1. Step one: two random points for gaussian distribution
2. Using expectation formula to calculate the prob of being in cluster A and B

GAUSSIAN MIXTURE MODEL CLUSTERING | EXAMPLE

Step 2 - Soft-cluster the data points - "Expectation" step

SOFT CLUSTERING ("RESPONSIBILITIES")

Point #	feature 1	feature 2	Cluster A	Cluster B
1	62	71	0.99976	1 - 0.99976
2	58	81		
3	52	74		
...		
N	52	78		

$$E[Z_{1A}] = \frac{N(\mathbf{x}_i|\mu_A, \sigma_A^2)}{N(\mathbf{x}_i|\mu_A, \sigma_A^2) + N(\mathbf{x}_i|\mu_B, \sigma_B^2)} = \frac{0.001288}{0.001288 + 0.0000038}$$

$$N(\mathbf{x}|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^2} e^{-\frac{1}{2\sigma^2}(\mathbf{x}-\mu)^2}$$


Cluster	μ	σ^2
A	(64.63, 76.30)	100
B	(46.02, 51.30)	57

3. Calculate Cluster A's μ by calculate the weighted average of Cluster A

GAUSSIAN MIXTURE MODEL CLUSTERING | EXAMPLE

Step 3 - Re-estimate parameters of Gaussians - "Maximization" step



NEW GAUSSIAN PARAMETERS

Cluster	new μ	new σ^2
A		
B		

$$\text{new } \mu_A = \frac{\sum_{i=1}^N E[Z_{ij}]X_i}{\sum_{i=1}^N E[Z_{ij}]} = \frac{0.99937x(62 \ 71) + 0.9998x(58 \ 81) + \dots}{0.99937 + 0.9998 + 0.55818 + \dots}$$

Point #	feature 1	feature 2	Cluster A	Cluster B
1	62	71	0.99937	0.00063
2	58	81	0.9998	0.0002
3	52	74	0.55818	0.44182
...
N	52	78	0.99133	0.00867

Calculate Cluster A's σ^2

GAUSSIAN MIXTURE MODEL CLUSTERING | EXAMPLE

Step 3 - Re-estimate parameters of Gaussians - "Maximization" step



NEW GAUSSIAN PARAMETERS

Cluster	new μ	new σ^2
A	(64.4872457, 76.3074590)	103.92494596
B	(46.0271498, 51.3087720)	67.10773268

$$\text{new } \sigma_A^2 = \frac{\sum_{i=1}^N E[Z_{iA}](X_i - \mu_A^{new})(X_i - \mu_A^{new})^T}{\sum_{i=1}^N E[Z_{iA}]} \\ = 103.92494596$$

Point #	feature 1	feature 2	Cluster A	Cluster B
1	62	71	0.99937	0.00063
2	58	81	0.9998	0.0002
3	52	74	0.55818	0.44182
...
N	52	78	0.99133	0.00867

Move the center and variance

GAUSSIAN MIXTURE MODEL CLUSTERING | EXAMPLE

Step 3 - Re-estimate parameters of Gaussians - "Maximization" step



NEW GAUSSIAN PARAMETERS

Cluster	new μ	new σ^2
A	(64.4872457, 76.3074590)	103.92494596
B	(46.0271498, 51.3087720)	67.10773268

$$\text{new } \sigma_A^2 = \frac{\sum_{i=1}^N E[Z_{iA}] (X_i - \mu_A^{new}) (X_i - \mu_A^{new})^T}{\sum_{i=1}^N E[Z_{iA}]}$$

$$= 103.92494596$$

Point #	feature 1	feature 2	Cluster A	Cluster B
1	62	71	0.99937	0.00063
2	58	81	0.9998	0.0002
3	52	74	0.55818	0.44182
...
N	52	78	0.99133	0.00867

4. Step 4: Evaluate log-likelihood: max this likelihood formula

Choose good mixing coefficient(π_k), μ_k , σ_k^2

$$\ln p(X|\mu, \sigma^2) = \sum_{i=1}^N \ln \left(\sum_{k=1}^K \pi_k N(X_i | \mu_k, \sigma_k^2) \right)$$

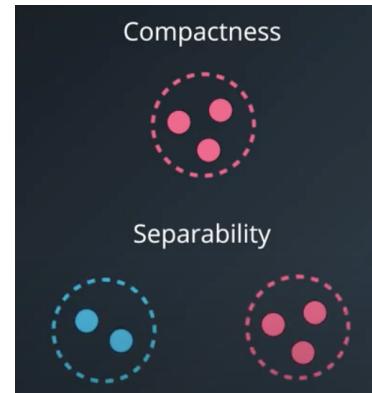
Clustering Validation

- external indices: we have labels
- internal indices: no labels
- relative indices:

Validation indices include

- compactness: how close the points in each cluster are to each other

- separability: how far distinct clusters are from each other



silhouette: for circular data

DBSCAN validation:

<http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=83C3BD5E078B1444CB26E243975507E1?doi=10.1.1.707.9034&rep=rep1&type=pdf>

▼ Questions

what is the difference between standardized and normalized?