THINKFUL

Gaussian Mixture Models

DATA SCIENCE

Warm Up

Consider the following question:

- Thus far, all of the clustering models we have discussed are “hard clustering” models – that is, each point either belongs fully to a cluster or not at all. What are the drawbacks of hard assignment?

Agenda

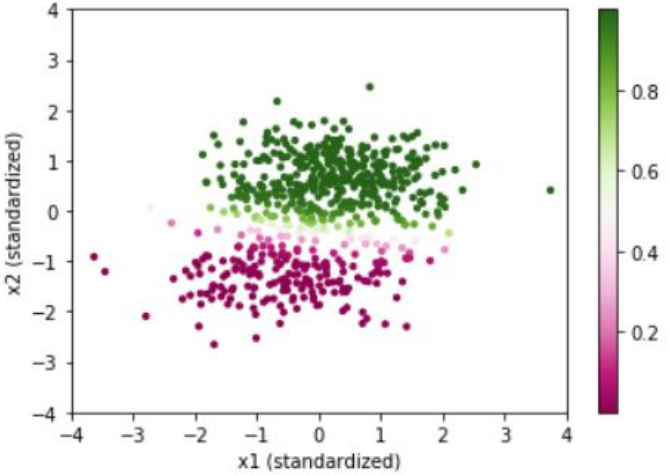
- Overview
- E-M Algorithm
- Gaussian Mixtures Modeling
- Advantages and Disadvantages

Overview

Gaussian Mixtures Models provide a “soft clustering” alternative to previously discussed techniques

- Assumes that the observations are generated by a mixture of processes governed by multivariate Gaussian density functions
- Model calculates a mean, covariance matrix, and mixing coefficient for each cluster (number of clusters must be prespecified) using an Expectation Maximization algorithm
- Output: assigns a probability of membership to each cluster (rather than the hard assignment of k-means, k-medoids, etc).

Overview



2-component Gaussian mixtures on a synthetic data set. Color scale represents posterior probability of membership to the cluster with mean corresponding to top cluster. Data was generated by multivariate Gaussian functions with the following parameters:

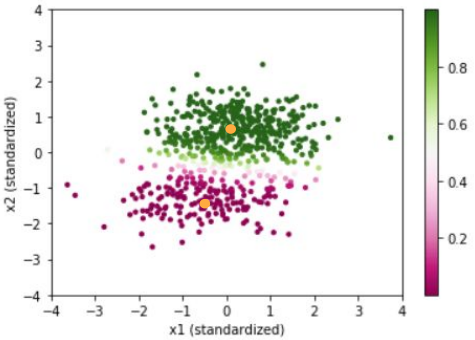
Cluster 1: $\mu = (23,23), \sum = \begin{bmatrix} 5 & 0 \\ 0 & 1 \end{bmatrix}, \tau = 2/3$

Clusters: $\mu = (20,20), \sum = \begin{bmatrix} 5 & 0 \\ 0 & 1 \end{bmatrix}, \tau = 1/3$

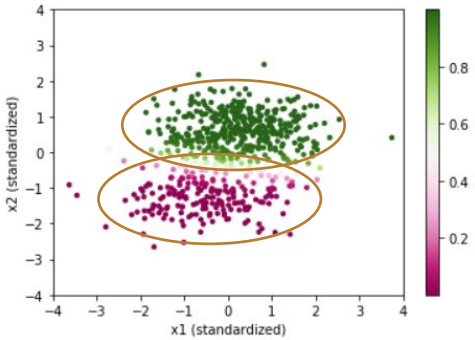
GMM Terminology

The Gaussian density functions estimated by GMM models are characterized by three parameters:

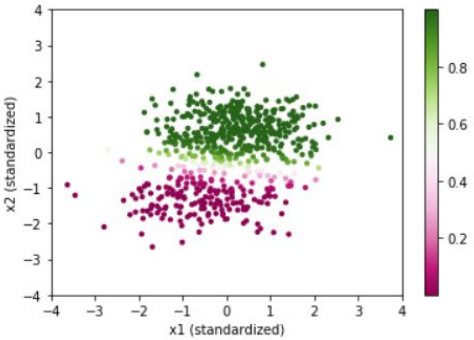
μ: Mean



∑: Covariance matrix



τ: Mixing proportion



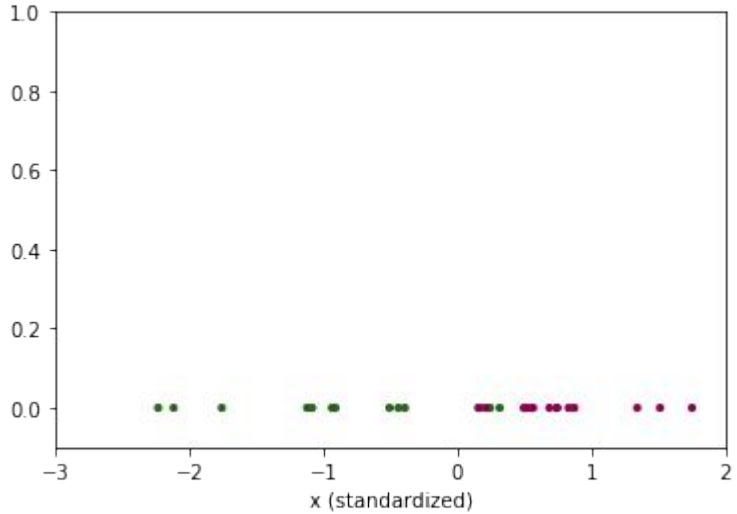
UP NEXT

Expectation Maximization Algorithm

E-M Algorithm

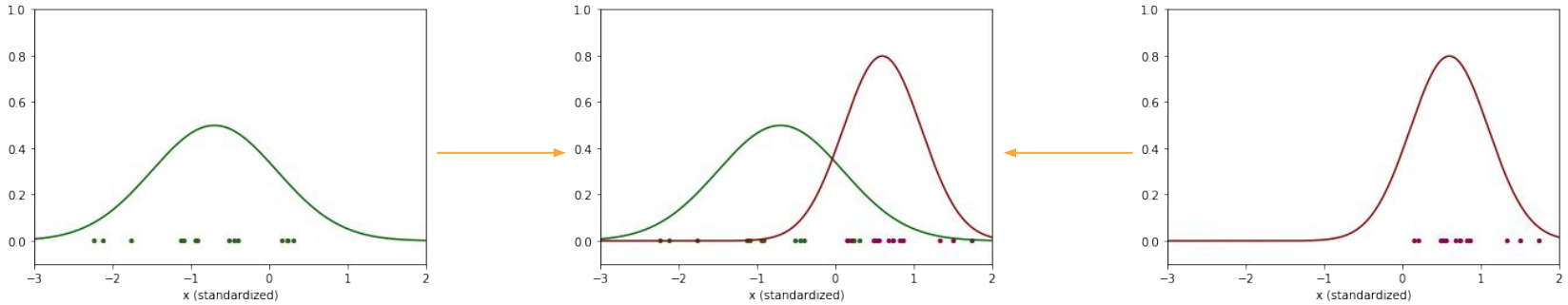
- This data was generated using `np.random.normal` to sample two different normal populations.
- This could also be said as the observations were generated by 2 different Gaussian (normal) processes.
- The last 2 columns show which process made which observations

Observation # (i)	Value (x_i)	Generated by process 1? (z_{i1})	Generated by process 2? (z_{i2})
1	-1.2	1	0
2	0.7	0	1
3	-1.4	1	0
…	…	…	…



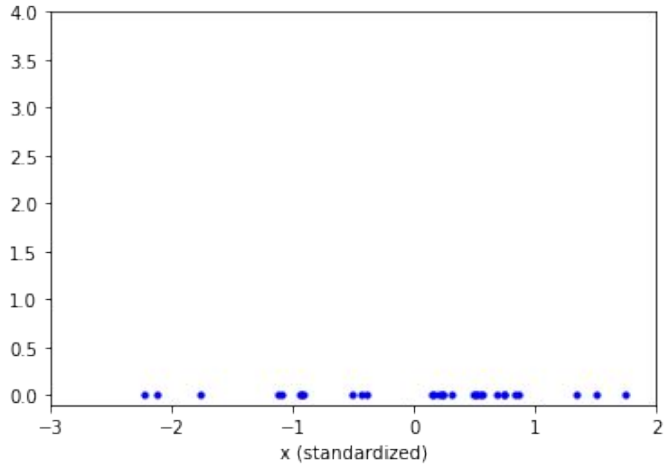
E-M Algorithm

If we knew which point was generated by which process (i.e. if we knew z_{ij}), we could use MLE to estimate the parameters of both underlying Gaussian density functions, which would characterize our clusters



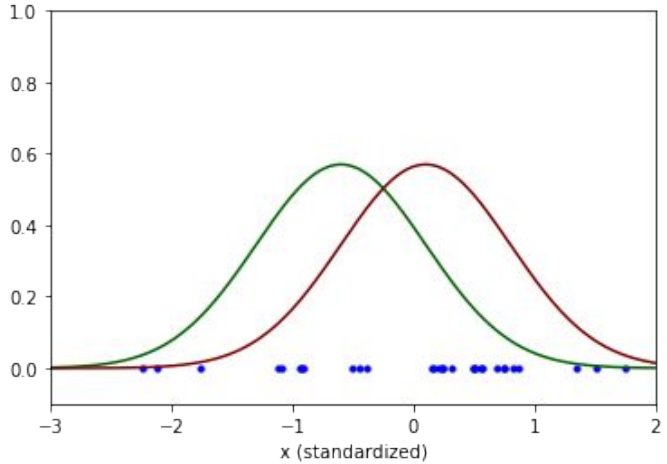
E-M Algorithm

However, in unsupervised learning, we do not know which process generated which point. Z_{ij} is a **hidden** or **latent** variable, and the E-M algorithm is one of the most popular algorithms for parameter estimation in contexts involving latent variables.



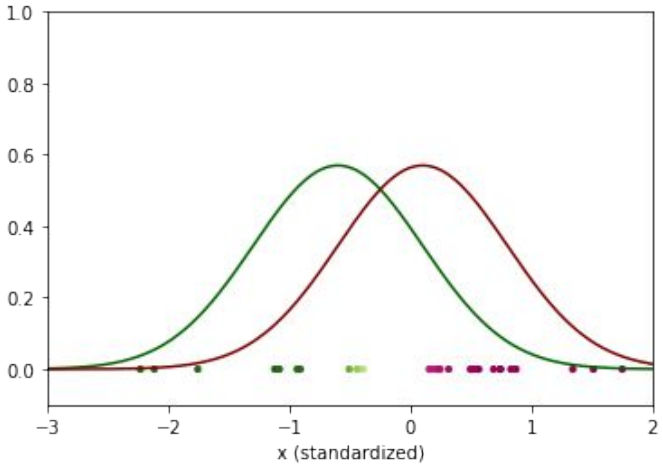
E-M Algorithm

Begin by making an initial guess for the parameters of our two Gaussian proportions, $\mu_1, \mu_2, \sigma_1,$ and σ_2, as well as the mixing proportions τ_1 and τ_2.



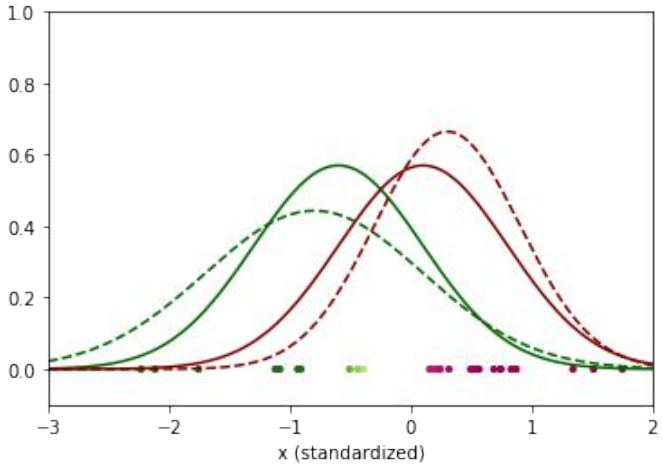
E-M Algorithm

E (Expectation) step: given our existing estimates of the parameters θ and mixing proportions τ, guess the values of the missing data (i.e. the missing Z_{ij}) by calculating the posterior probabilities for each point to be generated by process 1 and process 2 given the feature value(s). Theses are sometimes called the *responsibilities* and are mathematically equal to $\mathrm{E}[Z_{ij}|X, \theta, \tau]$.



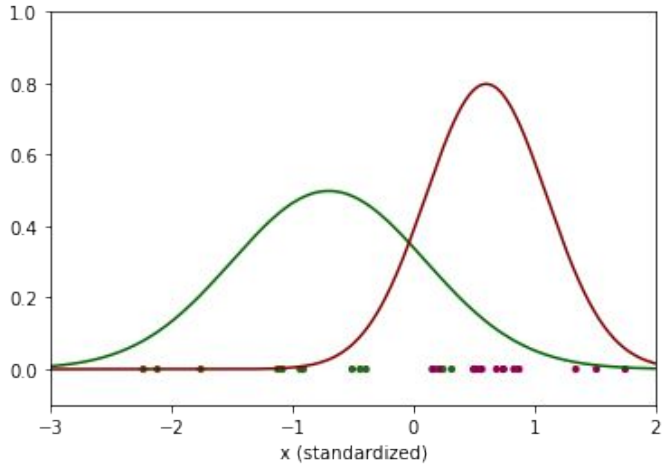
E-M Algorithm

M (Maximization) step: Update our estimates of $\mu_1, \mu_2, \sigma_1, \sigma_2, \tau_1,$ and τ_2 using our newly completed data ($\mathrm{E}[Z_{ij}|X, \mu_1, \dots]$) as weights; these parameters are estimated using MLE on the newly completed data



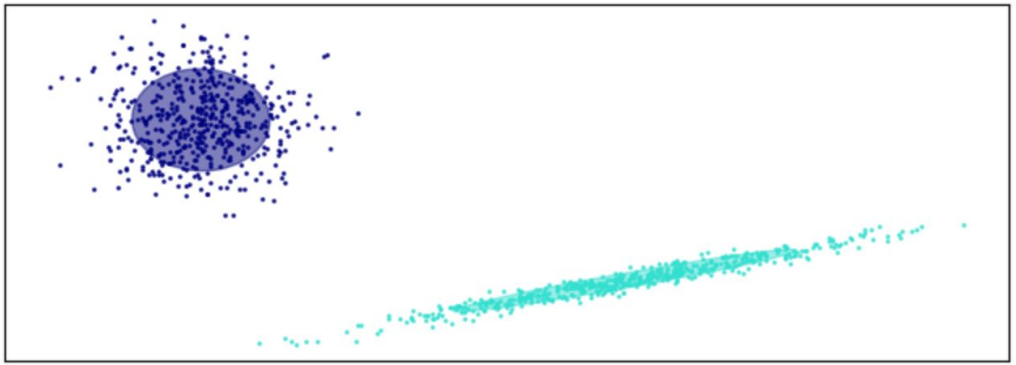
E-M Algorithm

Iteratively repeat the E and M steps until convergence



UP NEXT

Gaussian Mixture Modeling



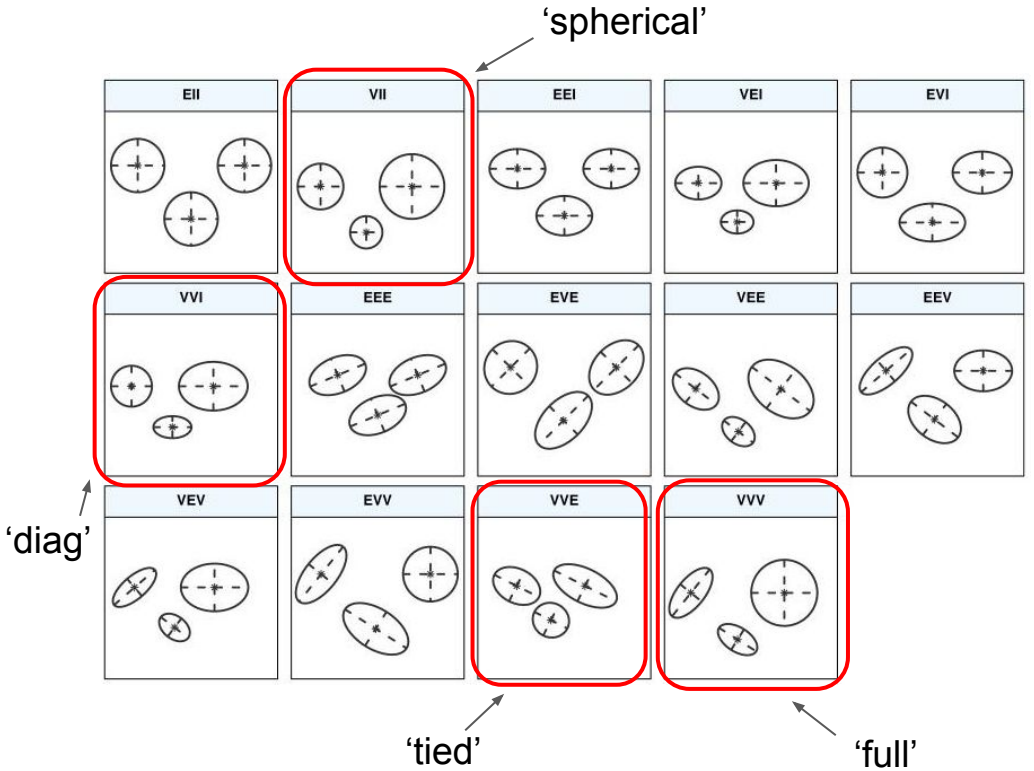
Gaussian Mixture Models

Gaussian Mixture Models extend the E-M algorithm to the multivariate case

- Must specify the number of clusters upfront
- Sensitive to initialization; a common approach is to initial means using k-means
- In addition to the parameters we estimated in the univariate case, we also must estimate covariances between pairs of features
- We can make assumptions which restrict the shapes of the clusters to reduce computation time, or allow each cluster to have independent, unconstrained covariance matrices for greater flexibility

Gaussian Mixture Models

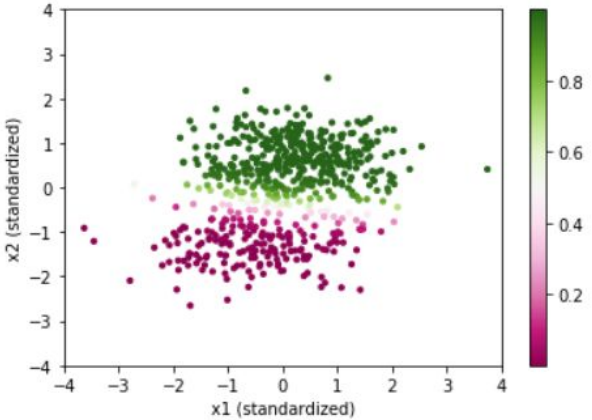
Model	Σ_k	Distribution	Volume	Shape	Orientation
EII	$\lambda \boldsymbol{I}$	Spherical	Equal	Equal	—
VII	$\lambda_k \boldsymbol{I}$	Spherical	Variable	Equal	—
EEI	$\lambda \boldsymbol{A}$	Diagonal	Equal	Equal	Coordinate axes
VEI	$\lambda_k \boldsymbol{A}$	Diagonal	Variable	Equal	Coordinate axes
EVI	$\lambda \boldsymbol{A}_k$	Diagonal	Equal	Variable	Coordinate axes
VVI	$\lambda_k \boldsymbol{A}_k$	Diagonal	Variable	Variable	Coordinate axes
EEE	$\lambda \boldsymbol{D}\boldsymbol{A}\boldsymbol{D}^\top$	Ellipsoidal	Equal	Equal	Equal
EVE	$\lambda \boldsymbol{D}\boldsymbol{A}_k\boldsymbol{D}^\top$	Ellipsoidal	Equal	Variable	Equal
VEE	$\lambda_k \boldsymbol{D}\boldsymbol{A}\boldsymbol{D}^\top$	Ellipsoidal	Variable	Equal	Equal
VVE	$\lambda_k \boldsymbol{D}\boldsymbol{A}_k\boldsymbol{D}^\top$	Ellipsoidal	Variable	Variable	Equal
EEV	$\lambda \boldsymbol{D}_k\boldsymbol{A}\boldsymbol{D}_k^\top$	Ellipsoidal	Equal	Equal	Variable
VEV	$\lambda_k \boldsymbol{D}_k\boldsymbol{A}\boldsymbol{D}_k^\top$	Ellipsoidal	Variable	Equal	Variable
EVV	$\lambda \boldsymbol{D}_k\boldsymbol{A}_k\boldsymbol{D}_k^\top$	Ellipsoidal	Equal	Variable	Variable
VVV	$\lambda_k \boldsymbol{D}_k\boldsymbol{A}_k\boldsymbol{D}_k^\top$	Ellipsoidal	Variable	Variable	Variable



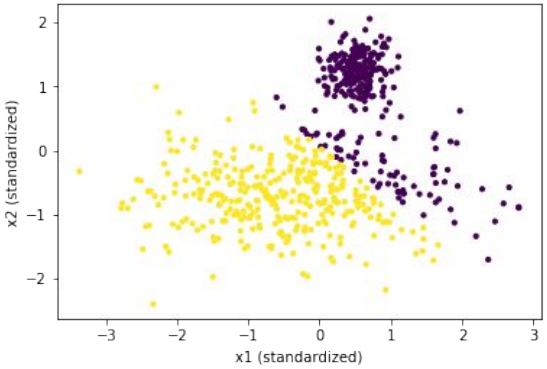
Gaussian Mixture Models

The output of a Gaussian model includes:

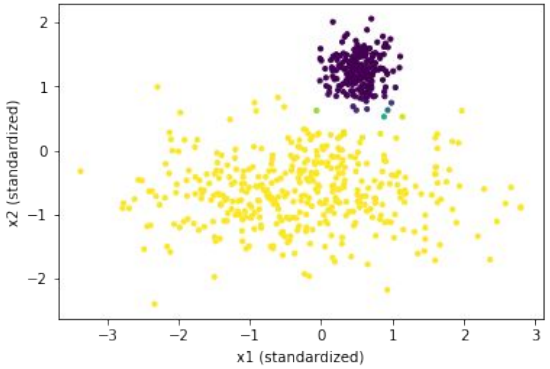
- Cluster means, which provide a centroid-like single point representation of each cluster
- Cluster covariance matrices, which provide greater detail about the spread of the observations generated by the underlying process
- Cluster mixing proportions, which indicate what proportion of the data is assumed to be generated by each underlying process
- The posterior probability that each point belongs to each cluster; alternatively, the responsibility of each cluster for each observation



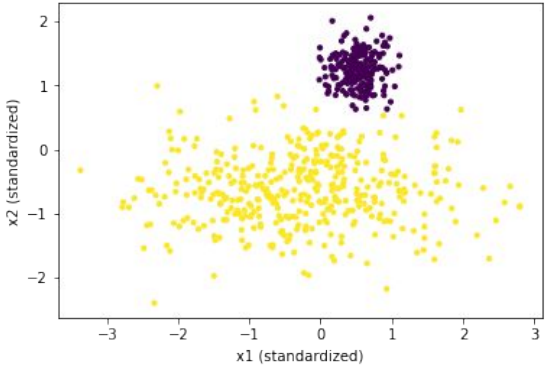
GMM vs. K-Means



K-Means



GMM
(probabilities)



GMM
(hard assignment)

Advantages & Disadvantages

Advantages

- Soft clustering allows for statements about the uncertainty of cluster assignments
- Covariance matrices allows for varying shapes across clusters, which leads to better clustering accuracy than k-means (which assumes spherical, equal volume clusters)
- Covariance matrices provide insights into the variability of the observations produced by the Gaussian process assumed to drive each cluster

Advantages & Disadvantages

Disadvantages

- Have to pre-specify the number of clusters
- Assumes clusters are generated by multivariate Gaussian density functions; may perform poorly for clusters that have irregular shapes
- Only appropriate for quantitative data, as Gaussian distributions are inherently quantitative; does not support categorical or mixed data
- Doesn’t scale well with high-dimensional data (because of covariance calculations)

Summary

- Gaussian Mixture Modeling is a form of soft clustering, in which observations are assigned a probability of belonging to clusters, rather than a hard assignment
- GMM uses the Expectation-Maximization, or E-M Algorithm to iteratively refine estimates of the mean, covariance matrix, and mixing proportion associated with the Gaussian density function underlying each cluster
- GMM is closely related to k-means, but has several advantages over it, such as the ability to identify clusters with varying shapes and sizes; however, it scales very poorly to high dimensional data

Assignment

1. Jupyter notebook
2. Data set: NBA player-seasons from 15-16 to 18-19
3. Filter players that do not play very much (i.e. few games started, few minutes player per game)
4. Isolate a subset of columns that would be useful for identifying player archetypes
5. Run both GMM and K-Means and compares the archetypes identified by both
6. Determine which hard clustering assignment we would be most/least confident about

Thank You

Soft Clustering: Gaussian Mixture Models and the E-M Algorithm

Warm Up

Consider the following question:

- Thus far, all of the clustering models we have discussed are “hard clustering” models – that is, each point either belongs fully to a cluster or not at all. What are the drawbacks of hard assignment?

High Level Agenda

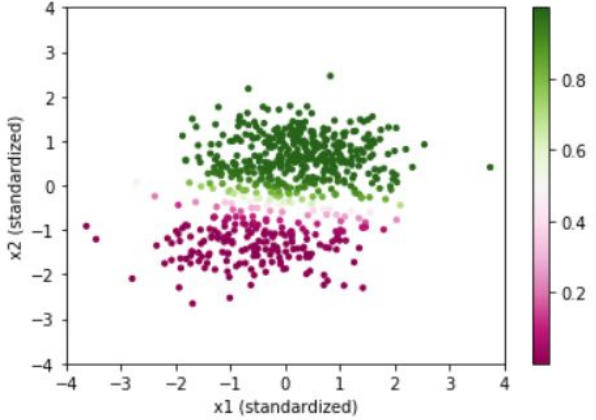
- Overview
- E-M Algorithm
- Gaussian Mixtures Modeling
- Advantages and Disadvantages

Overview

Overview

Gaussian Mixtures Models provide a “soft clustering” alternative to previously discussed techniques

- Assumes that the observations are generated by a mixture of processes governed by multivariate Gaussian density functions
- Model calculates a mean, covariance matrix, and mixing coefficient for each cluster (number of clusters must be prespecified) using the E-M algorithm
- Output: assigns a probability of membership to each cluster (rather than the hard assignment of k-means, k-medoids, etc).



2-component Gaussian mixtures on a synthetic data set. Color scale represents posterior probability of membership to the cluster with mean corresponding to top cluster. Data was generated by multivariate Gaussian functions with the following parameters:

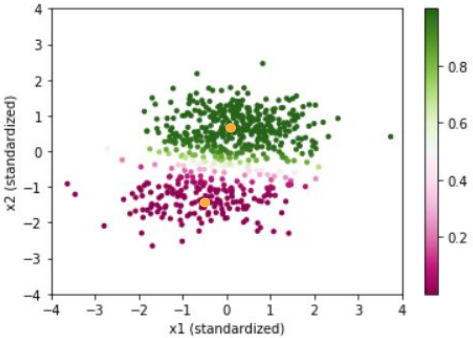
Cluster 1: $\mu = (23,23), \Sigma = \begin{bmatrix} 5 & 0 \\ 0 & 1 \end{bmatrix}, \tau = 2/3$

Clusters: $\mu = (20,20), \Sigma = \begin{bmatrix} 5 & 0 \\ 0 & 1 \end{bmatrix}, \tau = 1/3$

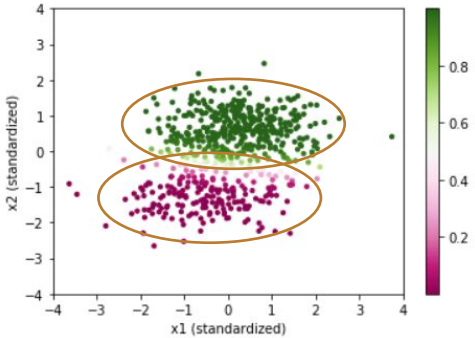
GMM Terminology

The Gaussian density functions estimated by GMM models are characterized by three parameters:

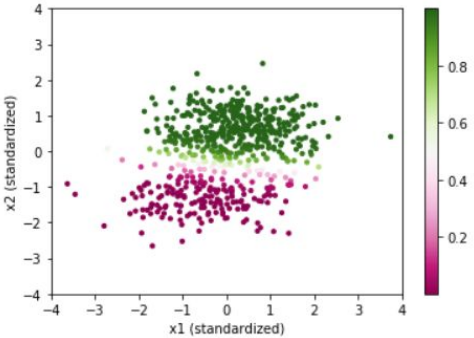
μ:
Mean



∑: Covariance
matrix



τ: Mixing
proportion

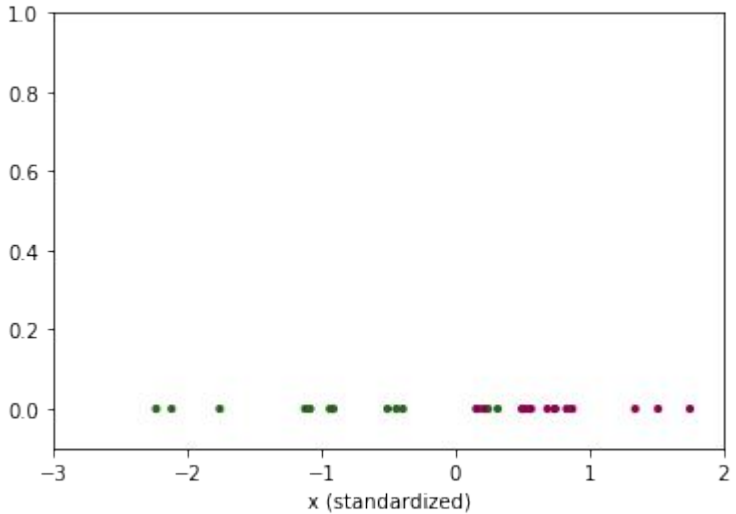


E-M (Expectation-Maximization Algorithm)

E-M Algorithm

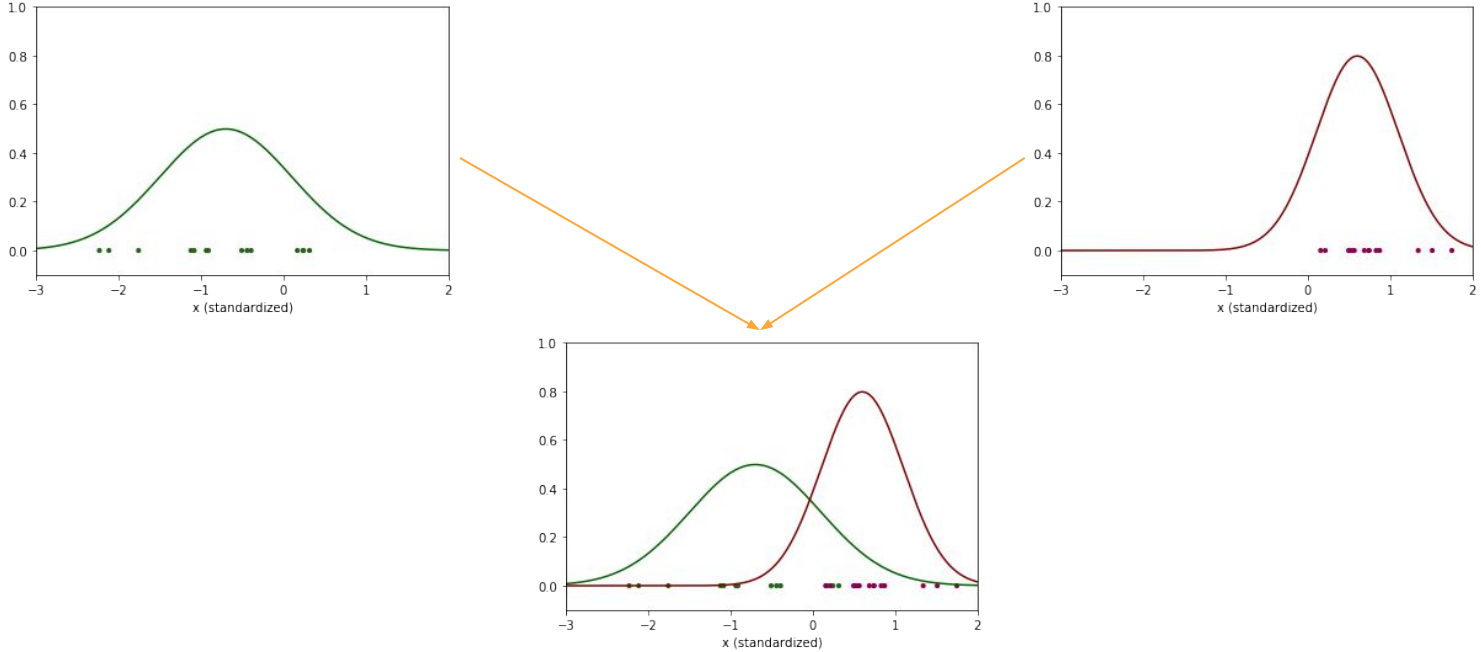
Imagine generating several 1-D observations, x_1-x_i, randomly from two processes with univariate Gaussian densities with equal mixing proportions. For each observation, we know which process created it, and we can also encode this information as a new variable, z_{ij}, which is equal to 1 if observation i was created by process j.

Observation # (i)	Value (x_i)	Generated by process 1? (z_{i1})	Generated by process 2? (z_{i2})
1	-1.2	1	0
2	0.7	0	1
3	-1.4	1	0
…	…	…	…



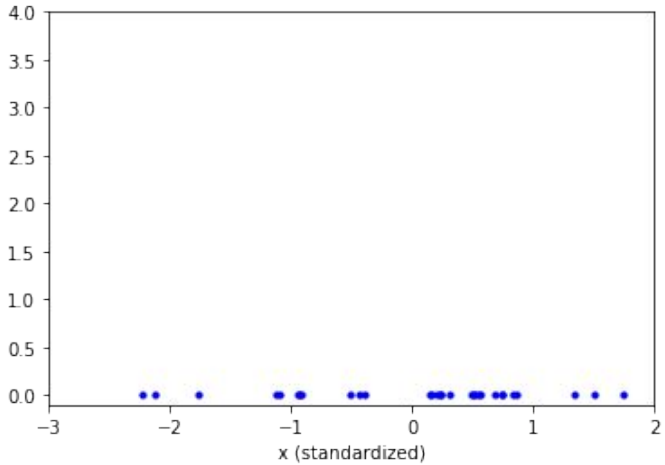
E-M Algorithm

If we knew which point was generated by which process (i.e. if we knew z_{ij}), we could use MLE to estimate the parameters of both underlying Gaussian density functions, which would characterize our clusters



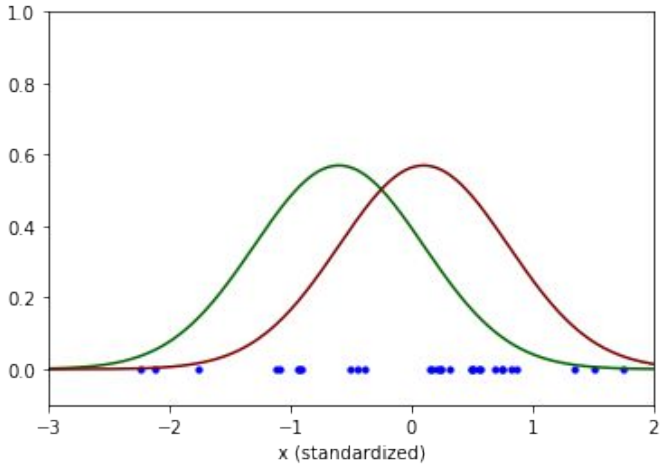
E-M Algorithm

However, in unsupervised learning, we do not know which process generated which point. Z_{ij} is a **hidden** or **latent** variable, and the E-M algorithm is one of the most popular algorithms for parameter estimation in contexts involving latent variables.



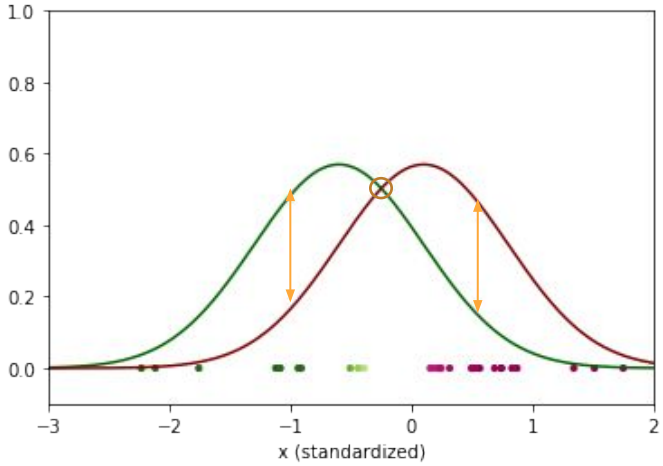
E-M Algorithm

Begin by making an initial guess for the parameters of our two Gaussian proportions, $\mu_1, \mu_2, \sigma_1,$ and σ_2, as well as the mixing proportions τ_1 and τ_2.



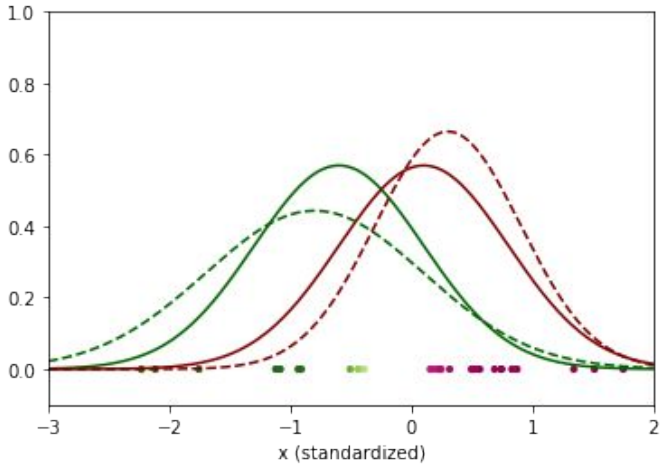
E-M Algorithm

E (Expectation) step: given our existing estimates of the parameters θ and mixing proportions τ, guess the values of the missing data (i.e. the missing Z_{ij}) by calculating the posterior probabilities for each point to be generated by process 1 and process 2 given the feature value(s). Theses are sometimes called the *responsibilities* and are mathematically equal to $\mathrm{E}[Z_{ij}|X, \theta, \tau]$.



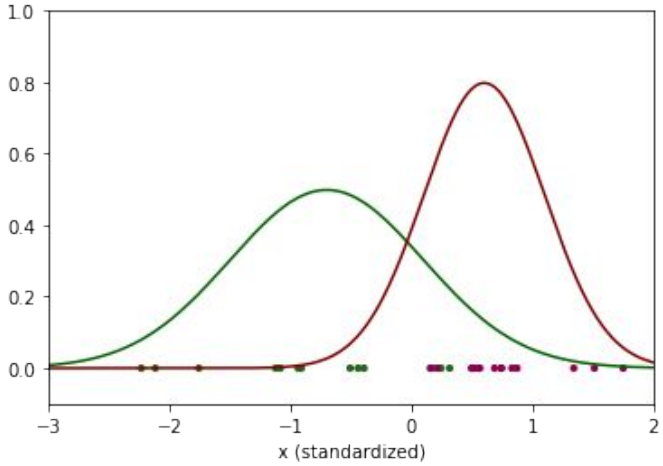
E-M Algorithm

M (Maximization) step: Update our estimates of $\mu_1, \mu_2, \sigma_1, \sigma_2, \tau_1$, and τ_2 using our newly completed data ($\mathrm{E}[Z_{ij}|X, \mu_1, \ldots]$) as weights; these parameters are estimated using MLE on the newly completed data



E-M Algorithm

Iteratively repeat the E and M steps until convergence



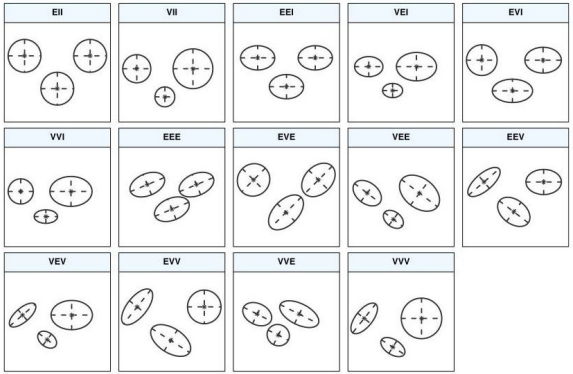
Gaussian Mixture Modeling

Gaussian Mixture Models

Gaussian Mixtures Models extend the E-M algorithm to the multivariate case

- Must specify the number of clusters upfront
- Sensitive to initialization; a common approach is to initial means using k-means
- In addition to the parameters we estimated in the univariate case, we also must estimate covariances between pairs of features
- We can make assumptions which restrict the shapes of the clusters to reduce computation time, or allow each cluster to have independent, unconstrained covariance matrices for greater flexibility

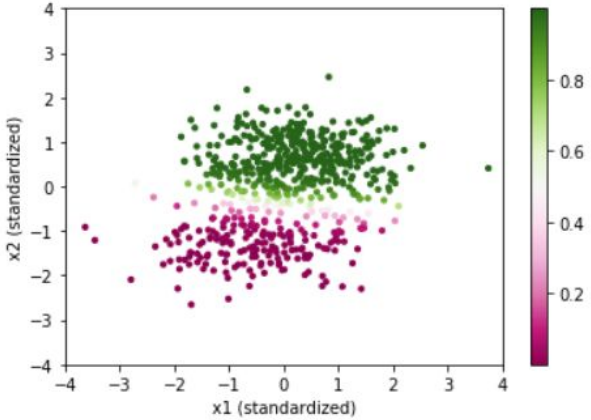
Model	Σ_k	Distribution	Volume	Shape	Orientation
EII	$\lambda \boldsymbol{I}$	Spherical	Equal	Equal	—
VII	$\lambda_k \boldsymbol{I}$	Spherical	Variable	Equal	—
EEI	$\lambda \boldsymbol{A}$	Diagonal	Equal	Equal	Coordinate axes
VEI	$\lambda_k \boldsymbol{A}$	Diagonal	Variable	Equal	Coordinate axes
EVI	$\lambda \boldsymbol{A}_k$	Diagonal	Equal	Variable	Coordinate axes
VVI	$\lambda_k \boldsymbol{A}_k$	Diagonal	Variable	Variable	Coordinate axes
EEE	$\lambda \boldsymbol{D}\boldsymbol{A}\boldsymbol{D}^\top$	Ellipsoidal	Equal	Equal	Equal
EVE	$\lambda \boldsymbol{D}\boldsymbol{A}_k\boldsymbol{D}^\top$	Ellipsoidal	Equal	Variable	Equal
VEE	$\lambda_k \boldsymbol{D}\boldsymbol{A}\boldsymbol{D}^\top$	Ellipsoidal	Variable	Equal	Equal
VVE	$\lambda_k \boldsymbol{D}\boldsymbol{A}_k\boldsymbol{D}^\top$	Ellipsoidal	Variable	Variable	Equal
EEV	$\lambda \boldsymbol{D}_k\boldsymbol{A}\boldsymbol{D}_k^\top$	Ellipsoidal	Equal	Equal	Variable
VEV	$\lambda_k \boldsymbol{D}_k\boldsymbol{A}\boldsymbol{D}_k^\top$	Ellipsoidal	Variable	Equal	Variable
EVV	$\lambda \boldsymbol{D}_k\boldsymbol{A}_k\boldsymbol{D}_k^\top$	Ellipsoidal	Equal	Variable	Variable
VVV	$\lambda_k \boldsymbol{D}_k\boldsymbol{A}_k\boldsymbol{D}_k^\top$	Ellipsoidal	Variable	Variable	Variable



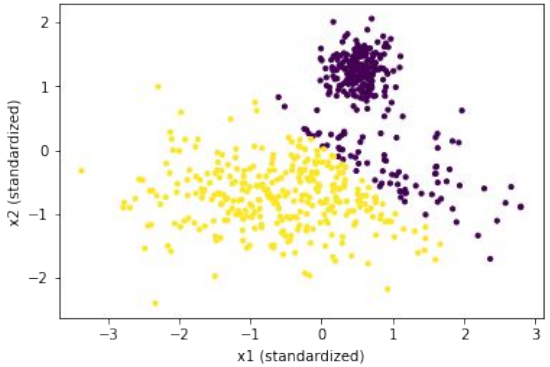
Gaussian Mixture Models

The output of a Gaussian model includes:

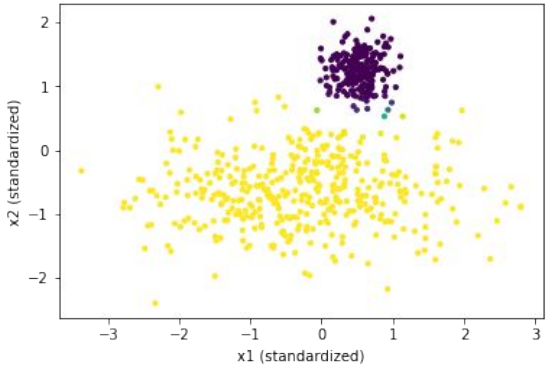
- Cluster means, which provide a centroid-like single point representation of each cluster
- Cluster covariance matrices, which provide greater detail about the spread of the observations generated by the underlying process
- Cluster mixing proportions, which indicate what proportion of the data is assumed to be generated by each underlying process
- The posterior probability that each point belongs to each cluster; alternatively, the responsibility of each cluster for each observation



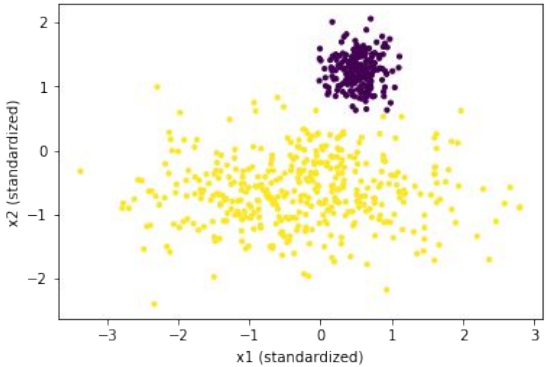
GMM vs. K-Means



K-Means



GMM
(probabilities)



GMM
(hard assignment)

Advantages and Disadvantages

Advantages and Disadvantages

Advantages

- Soft clustering allows for statements about the uncertainty of cluster assignments
- Covariance matrices allows for varying shapes across clusters, which leads to better clustering accuracy than k-means (which assumes spherical, equal volume clusters)
- Covariance matrices provide insights into the variability of the observations produced by the Gaussian process assumed to drive each cluster

Disadvantages

- Have to pre-specify the number of clusters
- Assumes clusters are generated by multivariate Gaussian density functions; may perform poorly for clusters that have irregular shapes
- Only appropriate for quantitative data, as Gaussian distributions are inherently quantitative; does not support categorical or mixed data
- Doesn’t scale well with high-dimensional data (because of covariance calculations)

Recap

- Gaussian Mixture Modeling is a form of soft clustering, in which observations are assigned a probability of belonging to clusters, rather than a hard assignment
- GMM uses the Expectation-Maximization, or E-M Algorithm to iteratively refine estimates of the mean, covariance matrix, and mixing proportion associated with the Gaussian density function underlying each cluster
- GMM is closely related to k-means, but has several advantages over it, such as the ability to identify clusters with varying shapes and sizes; however, it scales very poorly to high dimensional data

Exercise:
Gaussian Mixture Models

- Jupyter notebook
- Data set: NBA player-seasons from 15-16 to 18-19
- Filter players that do not play very much (i.e. few games started, few minutes player per game)
- Isolate a subset of columns that would be useful for identifying player archetypes
- Run both GMM and K-Means and compares the archetypes identified by both
- Determine which hard clustering assignment we would be most/least confident about