Application of Adaptive Annealing Method to Generalized Incremental Algorithm

Una Radojičić

University J. J. Strossmayer of Osijek
Department of Mathematics
Trg Ljudevita Gaja 6
31000 Osijek, Croatia
http://www.mathos.unios.hr

uradojic@mathos.hr

[Supported by CSF through research grant IP-2016-06-6545]

11.5.2018.

Model assumptions

Goal is to estimate parameters $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_k)$, $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_k)$, $\mu_i \in \mathbb{R}^d$, $\sigma_i > 0$, of special case of Gaussian Mixture Model

$$f(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\sigma}) = \frac{1}{k} \sum_{i=1}^{k} f_i(\boldsymbol{x}; \mu_i, \sigma_i I_d) = \frac{1}{k} \sum_{i=1}^{k} \prod_{j=1}^{n} f_{ij}(x_j; \mu_i, \sigma_i),$$

where k is number of clusters, n is number of data in $\mathbb{R}^d$ known in advance.

- Data within the same cluster is independent and equally distributed from $f_{ij} \sim \mathcal{N}(\mu_i, \sigma_i)$
- Further on, focus will be placed on estimation of the expectation $\boldsymbol{\mu}$, due to fact that estimation of standard deviations $\boldsymbol{\sigma}$ follows trivially from it.

Model assumptions

Goal is to estimate parameters $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_k)$, $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_k)$, $\mu_i \in \mathbb{R}^d$, $\sigma_i > 0$, of special case of Gaussian Mixture Model

$$f(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\sigma}) = \frac{1}{k} \sum_{i=1}^{k} f_i(\boldsymbol{x}; \mu_i, \sigma_i I_d) = \frac{1}{k} \sum_{i=1}^{k} \prod_{j=1}^{n} f_{ij}(x_j; \mu_i, \sigma_i),$$

where k is number of clusters, n is number of data in $\mathbb{R}^d$ known in advance.

- Data within the same cluster is independent and equally distributed from $f_{ij} \sim \mathcal{N}(\mu_i, \sigma_i)$
- Further on, focus will be placed on estimation of the expectation $\boldsymbol{\mu}$, due to fact that estimation of standard deviations $\boldsymbol{\sigma}$ follows trivially from it.

Model assumptions

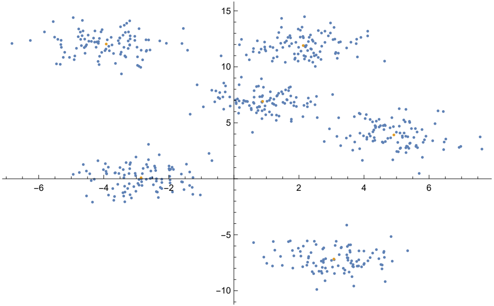
Goal is to estimate parameters $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_k)$, $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_k)$, $\mu_i \in \mathbb{R}^d$, $\sigma_i > 0$, of special case of Gaussian Mixture Model

$$f(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\sigma}) = \frac{1}{k} \sum_{i=1}^{k} f_i(\boldsymbol{x}; \mu_i, \sigma_i I_d) = \frac{1}{k} \sum_{i=1}^{k} \prod_{j=1}^{n} f_{ij}(x_j; \mu_i, \sigma_i),$$

where k is number of clusters, n is number of data in $\mathbb{R}^d$ known in advance.

- Data within the same cluster is independent and equally distributed from $f_{ij} \sim \mathcal{N}(\mu_i, \sigma_i)$
- Further on, focus will be placed on estimation of the expectation $\boldsymbol{\mu}$, due to fact that estimation of standard deviations $\boldsymbol{\sigma}$ follows trivially from it.

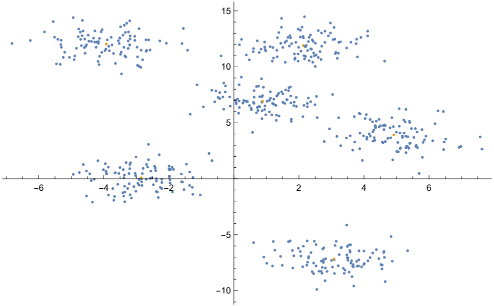
Main method idea



- Data within each of k clusters are independent and equally distributed from $f_{ij} \sim \mathcal{N}(\mu_i, \sigma_i I_d)$

- If we could detect data with upper property, it would be reasonable to tag them as one of k clusters.
- Appropriate average would be approximated with sample mean.
- This idea is conduced in the following way:

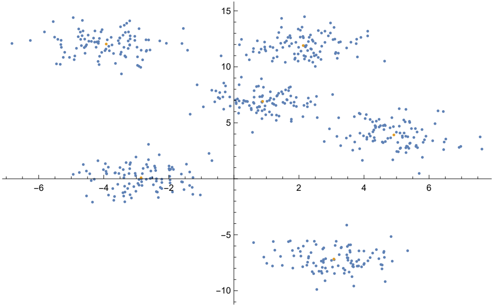
Main method idea



- Data within each of k clusters are independent and equally distributed from $f_{ij} \sim \mathcal{N}(\mu_i, \sigma_i I_d)$

- If we could detect data with upper property, it would be reasonable to tag them as one of k clusters.
- Appropriate average would be approximated with sample mean.
- This idea is conduced in the following way:

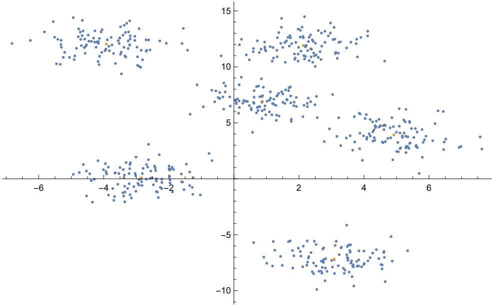
Main method idea



- Data within each of k clusters are independent and equally distributed from $f_{ij} \sim \mathcal{N}(\mu_i, \sigma_i I_d)$

- If we could detect data with upper property, it would be reasonable to tag them as one of k clusters.
- Appropriate average would be approximated with sample mean.
- This idea is conduced in the following way:

Main method idea



- Data within each of k clusters are independent and equally distributed from $f_{ij} \sim \mathcal{N}(\mu_i, \sigma_i I_d)$

- If we could detect data with upper property, it would be reasonable to tag them as one of k clusters.
- Appropriate average would be approximated with sample mean.
- This idea is conduced in the following way:

- Each of the n data is randomly (uniformly) joined to one of k clusters.
- Depending on such data labeling, we detect the data associated with each of k cluster.
- Using Shapiro-Wilks test, we test the normality of data in each cluster - due to the assumption of normality and independence in each cluster, the normality of d-dimensional data is equivalent to the normality of its margins.
- Cluster with the largest Shapiro-Wilks p-value is used to estimate one of k centers.
- Upper cluster is updated by placing into it closest n/k data to the estimated center.
- Last two steps are iterated until "convergence".
- The estimated cluster is removed from the data set and the procedure is repeated.

- Each of the n data is randomly (uniformly) joined to one of k clusters.
- Depending on such data labeling, we detect the data associated with each of k cluster.
- Using Shapiro-Wilks test, we test the normality of data in each cluster - due to the assumption of normality and independence in each cluster, the normality of d-dimensional data is equivalent to the normality of its margins.
- Cluster with the largest Shapiro-Wilks p-value is used to estimate one of k centers.
- Upper cluster is updated by placing into it closest n/k data to the estimated center.
- Last two steps are iterated until "convergence".
- The estimated cluster is removed from the data set and the procedure is repeated.

- Each of the n data is randomly (uniformly) joined to one of k clusters.
- Depending on such data labeling, we detect the data associated with each of k cluster.
- Using Shapiro-Wilks test, we test the normality of data in each cluster - due to the assumption of normality and independence in each cluster, the normality of d-dimensional data is equivalent to the normality of its margins.
- Cluster with the largest Shapiro-Wilks p-value is used to estimate one of k centers.
- Upper cluster is updated by placing into it closest n/k data to the estimated center.
- Last two steps are iterated until "convergence".
- The estimated cluster is removed from the data set and the procedure is repeated.

- Each of the n data is randomly (uniformly) joined to one of k clusters.
- Depending on such data labeling, we detect the data associated with each of k cluster.
- Using Shapiro-Wilks test, we test the normality of data in each cluster - due to the assumption of normality and independence in each cluster, the normality of d-dimensional data is equivalent to the normality of its margins.
- Cluster with the largest Shapiro-Wilks p-value is used to estimate one of k centers.
- Upper cluster is updated by placing into it closest n/k data to the estimated center.
- Last two steps are iterated until "convergence".
- The estimated cluster is removed from the data set and the procedure is repeated.

- Each of the n data is randomly (uniformly) joined to one of k clusters.
- Depending on such data labeling, we detect the data associated with each of k cluster.
- Using Shapiro-Wilks test, we test the normality of data in each cluster - due to the assumption of normality and independence in each cluster, the normality of d-dimensional data is equivalent to the normality of its margins.
- Cluster with the largest Shapiro-Wilks p-value is used to estimate one of k centers.
- Upper cluster is updated by placing into it closest n/k data to the estimated center.
- Last two steps are iterated until "convergence".
- The estimated cluster is removed from the data set and the procedure is repeated.

- Each of the n data is randomly (uniformly) joined to one of k clusters.
- Depending on such data labeling, we detect the data associated with each of k cluster.
- Using Shapiro-Wilks test, we test the normality of data in each cluster - due to the assumption of normality and independence in each cluster, the normality of d-dimensional data is equivalent to the normality of its margins.
- Cluster with the largest Shapiro-Wilks p-value is used to estimate one of k centers.
- Upper cluster is updated by placing into it closest n/k data to the estimated center.
- Last two steps are iterated until "convergence".
- The estimated cluster is removed from the data set and the procedure is repeated.

- Each of the n data is randomly (uniformly) joined to one of k clusters.
- Depending on such data labeling, we detect the data associated with each of k cluster.
- Using Shapiro-Wilks test, we test the normality of data in each cluster - due to the assumption of normality and independence in each cluster, the normality of d-dimensional data is equivalent to the normality of its margins.
- Cluster with the largest Shapiro-Wilks p-value is used to estimate one of k centers.
- Upper cluster is updated by placing into it closest n/k data to the estimated center.
- Last two steps are iterated until "convergence".
- The estimated cluster is removed from the data set and the procedure is repeated.

Problem!

- The likelihood that the data will be associated with the right cluster is $1/k$.
- The likelihood that the random labeling of data will "hit" the entire cluster is approximately $\left(\frac{1}{k}\right)^{n/k}$.
- $\left(\frac{1}{k}\right)^{n/k} << 1$ in case of large number of clusters, i.e. data.
- The chances of success are increased by repeating the same procedure a number of times.

Problem!

- The likelihood that the data will be associated with the right cluster is $1/k$.
- The likelihood that the random labeling of data will "hit" the entire cluster is approximately $\left(\frac{1}{k}\right)^{n/k}$.
- $\left(\frac{1}{k}\right)^{n/k} << 1$ in case of large number of clusters, i.e. data.
- The chances of success are increased by repeating the same procedure a number of times.

Problem!

- The likelihood that the data will be associated with the right cluster is $1/k$.
- The likelihood that the random labeling of data will "hit" the entire cluster is approximately $\left(\frac{1}{k}\right)^{n/k}$.
- $\left(\frac{1}{k}\right)^{n/k} << 1$ in case of large number of clusters, i.e. data.
- The chances of success are increased by repeating the same procedure a number of times.

Problem!

- The likelihood that the data will be associated with the right cluster is $1/k$.
- The likelihood that the random labeling of data will "hit" the entire cluster is approximately $\left(\frac{1}{k}\right)^{n/k}$.
- $\left(\frac{1}{k}\right)^{n/k} << 1$ in case of large number of clusters, i.e. data.
- The chances of success are increased by repeating the same procedure a number of times.

Step by step method explanation on the synthetic data in $\mathbb{R}^2$

We will demonstrate how the algorithm works in an example of 150 data points in $\mathbb{R}^2$ coming from 4 clusters.

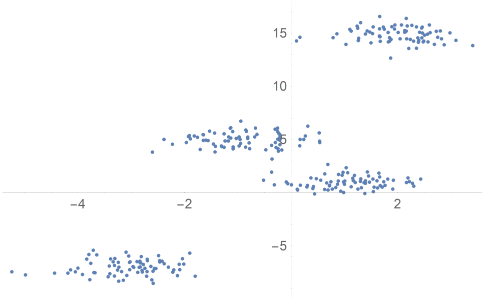
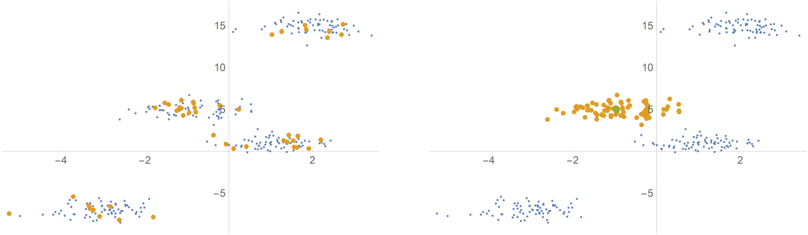


Figure: Genetated dataset in $\mathbb{R}^2$

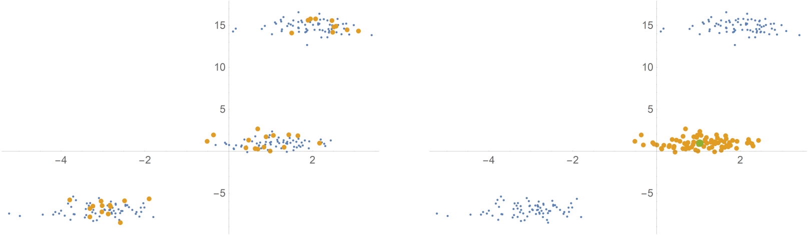


(a) Initially chosen 1. cluster

(b) 1. cluster after iterating cluster - center 10 times

Figure: Estimation of first center

The cluster specified in the previous iteration is removed from the data set

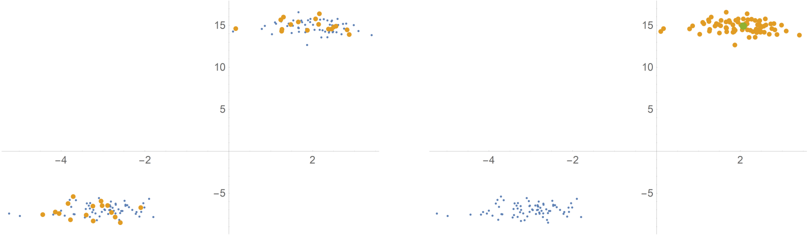


(a) Initially chosen 2. cluster

(b) 2. cluster after iterating cluster - center 10 times

Figure: Estimation of second center

The cluster specified in the previous iteration is removed from the data set

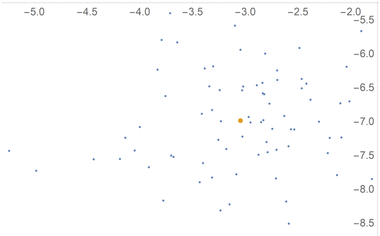


(a) Initially chosen 3. cluster

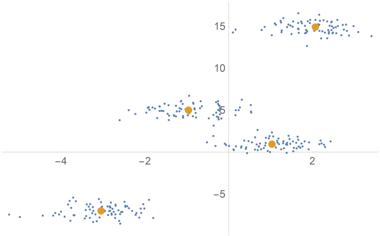
(b) 3. cluster after iterating cluster - center 10 times

Figure: Estimation of third center

The cluster specified in the previous iteration is removed from the data set



(a) The remaining data form the last cluster



(b) Initial data set with estimated centers

- The advantage of the method is manifested in its speed and precision on high-dimensional data sets, large data sets, and clusters with a large number of clusters.
- Normality assumption can be weakened to an large class of symmetric distribution.
- The estimation of centers obtained by the method can also be used as an initial approximation of the known k-Means algorithm to increase precision.
- The problem of estimating cluster centers can also be seen as a global minimization problem with the goal function

$$F(\mu_1, \dots, \mu_k; \boldsymbol{x}) = \sum_{i=1}^{n} \min_{1 \le j \le k} \mathrm{d}(x_i, \mu_j).$$

- The advantage of the method is manifested in its speed and precision on high-dimensional data sets, large data sets, and clusters with a large number of clusters.
- Normality assumption can be weakened to an large class of symmetric distribution.
- The estimation of centers obtained by the method can also be used as an initial approximation of the known k-Means algorithm to increase precision.
- The problem of estimating cluster centers can also be seen as a global minimization problem with the goal function

$$F(\mu_1, \dots, \mu_k; \boldsymbol{x}) = \sum_{i=1}^{n} \min_{1\le j\le k} \mathrm{d}(x_i, \mu_j).$$

- The advantage of the method is manifested in its speed and precision on high-dimensional data sets, large data sets, and clusters with a large number of clusters.
- Normality assumption can be weakened to an large class of symmetric distribution.
- The estimation of centers obtained by the method can also be used as an initial approximation of the known k-Means algorithm to increase precision.
- The problem of estimating cluster centers can also be seen as a global minimization problem with the goal function

$$F(\mu_1, \ldots, \mu_k; \boldsymbol{x}) = \sum_{i=1}^{n} \min_{1 \leq j \leq k} \mathrm{d}(x_i, \mu_j).$$

- The advantage of the method is manifested in its speed and precision on high-dimensional data sets, large data sets, and clusters with a large number of clusters.
- Normality assumption can be weakened to an large class of symmetric distribution.
- The estimation of centers obtained by the method can also be used as an initial approximation of the known k-Means algorithm to increase precision.
- The problem of estimating cluster centers can also be seen as a global minimization problem with the goal function

$$F(\mu_1, \ldots, \mu_k; \boldsymbol{x}) = \sum_{i=1}^{n} \min_{1 \le j \le k} \mathrm{d}(x_i, \mu_j).$$

Comparison of the "InitialEM" method with some known clustering algorithms

Table: Synthetic data in $\mathbb{R}^3$, $n = 20000$, $k = 5$

Method	Time(s)	Goal function value
InitialEm	6.0625	59647.9
InitialEm + K-means	21.0469	59639.9
Fdirect	747.609	59639.9
Zha	-	-
Zha+K-means	-	-

Table: Synthetic data in $\mathbb{R}^{15}$, $n = 10000$, $k = 2$

Method	Time(s)	Goal function value
InitialEm	4.39063	37565.9
InitialEm + K-means	9.67188	37564.5
Fdirect	61.4531	37564.5
Zha	5.32813	37564.5
Zha+K-means	6.71875	37564.5

Testing method on IRIS data set

- A typical test case for many statistical classification techniques in machine learning such as support vector machines.
- The data set contains a set of 150 records under 4 attributes - petal and sepal length and width .
- The quality of the classification is measured by the "Adjusted Random Index" (AdjRand).

Table: The value of AdjRandIndex depending on the classification method

Method	InitialEm	Fdirect	Zha+K-means
AdjRand	0.7142	0.686081	0.686081