Subject

Gaussian mixture model based clustering with TANAGRA: the EM algorithm.

In the **Gaussian mixture model-based clustering**, each cluster is represented by a Gaussian distribution. The entire dataset is modeled by a mixture (a linear combination) of these distributions.

The **EM** (Expectation Maximization) algorithm is used in practice to find the "optimal" parameters of the distributions that maximize the likelihood function.

The number of clusters is a parameter of the algorithm. But we can also detect the "optimal" number of clusters by evaluating several values, i.e. testing 1 cluster, 2 clusters, etc. and choosing the best one (which maximizes the likelihood or another criterion such as AIC or BIC).

Dataset

We use a synthetic dataset in a two dimensional space¹. We aim to discover two clusters (Figure 1).





¹ This dataset comes from the free distribution of «FAST EM Clustering» (AUTONLAB -- <u>http://www.autonlab.org/autonweb/10466.html</u>).

Clustering with the EM Algorithm

Create a diagram

We create a new diagram (FILE/NEW) and import the TWO_GAUSSIANS.XLS dataset.

TANAGRA 1.4.9				>
le Diagram Window Help				
Choose your dataset and				
	• .			
Diagram title :	-			
Data mining diagram	fla nama i			
D:\DataMining\Datab	ases for mining\dataset for s	dev and com	ipa 🖥	
		1 A.A.		
Dataset (".txt,".arft,".xls	s) :	2		
		G		
		Tanagra		
	ОК	Regarder dans :	: 🔁 clustering 🔤 🕜 🤣 🖽 -	
		Mes documents récents	間 care.xke 利 dog.xks 電 protein.xks ※ 1 two gaussians.xks	
		Direau		
			1 N	
Data visualization	Statistics		N	
Feature selection	Regression	Mes documents	N	
Spv learning	Meta-spv learning		1	
AGCT A	стр 🏑 ем-		1	
<		Poste de travail		
			Nom du lichier: two_gaussians.xls	Uuvrir
		Favoris réseau	Fichiers de type : 🔪 Excel File (97 & 2000) 🖉 🗸 🗸	Annuler

Plotting the examples

We use the SCATTERPLOT component in order to plotting the examples in a scatter plot. We distinguish well the two clusters.



INPUT attributes

We use the DEFINE STATUS component in order to select the input attributes (X0 and X1).

🌋 TANAGRA 1.4.9 - [Scatt	erplot 1]							
Tile Diagram Component	Window Help							a x
Default title		X1		~				~
🖵 🧰 Dataset (two_gaussia	ns.xls)			0111	XII vs. (X2) X1			
🔰 🧖 Scatterplot 1	Define attribu	ite statuses		()				
🦾 🎎 Define status 1	Derme attribu	no otaraboo						
	Parameters							
	و مع المانية الم							
	Attributes	:			Target	nput III	ustrative	
				X	(0			
				- 1				
			-	-				
Data visualization								
Feature construction								
PLS		- 1						
Spv learning assessment		B	<u>**</u>		Clea	r selection		
Correlation scatterplot								
÷				l	ок	Cancel	Hel	
<	1111							2

K-MEANS clustering algorithm

In the first step, we use the K-MEANS algorithm in order to create clusters. The aim is to have a reference, which will enable us to compare the later results.

We insert the K-MEANS (CLUSTERING tab) into the diagram. We modify the number of clusters parameter.

TANAGRA 1.4.9 - [Scatterplot 1]		
Tile Diagram Component Window Help		- 8 ×
D 📽 🔚 🔛		
Default title		~
🖃 🥅 Dataset (two gaussians.xls)	K-Means parameters	
Scatterplot 1	F	
🖃 🌄 Define status 1 🛛 👔	Parameters Results	
K-Means 1 Parameters Execute View 0	Number of clusters : 2 2 Max iterations : 10 2 Number of trials : 5 2 Distance normalization None O Variance Average computation O Forgy	
Data visualization Statistics	Mc Queen	Feature construction
Feature selection Regression	Seed random generator	Clustering
Spv learning Meta-spv learning	○ Random	Association
A-CT A-CTP Memory K-Means Memory Kommen-SOM	Standard	
	UK Cancel Help	

In order to visualize the clusters, we insert again a SCATTERPLOT component. We color the points with their cluster membership.



The K-MEANS method roughly found the two clusters. But we see that some examples seem visually misclassified. It is not surprising. The K-MEANS estimates only the barycentre of the clusters. The underlying hypothesis is that the shape of the clusters (covariance matrix) is spherical. This hypothesis is obviously wrong for our dataset.

Gaussian mixture based clustering

We insert the EM-CLUSTERING component in our diagram. The number of desired clusters is 2.

TANAGRA 1.4.9 - [Scatterplot 2	2] Holo			
	neip			
Default title				
Dataset (two gaussians.xk)			¥	Cluster KMeans I 💌
Scatterplot 1	-		(X1) X0 vs. (X2) X1 by (Y) Clust	er_KMeans_1
🖃 🌇 Define status 1		El	A clustering parameters	
🖃 💁 K-Means 1	1		in our of the second of the	
└─ <mark> </mark> <u>∦</u> Scatterplot 2 ── <mark></mark> EM-Clustering 1			Parameters Results	
Par Exe Viev	ameters cute v 0		Number of clusters :	2
			Max iteration :	50
			Ridge coef. :	0.001
			Standardization of	attributes for K-Means
Data visualization	Statistics	Non		
nic reat	ure selection		OK	Cancel Help
Spy learning assessment	Scoring	A	ssociation	
Ar CT	ering Anti-	IAC (-Means	📓 Kohonen-SOM 🏝 LVQ	Reighborhood Graph

The component internally initializes the clusters with a K-MEANS approach, and then it optimizes the likelihood with the EM algorithm. The computation is stopped when the likelihood is not increasing anymore or when we have reached the maximum number of iterations². TANAGRA shows the size of each cluster, their means and the quality of the clustering.

		EM-
		P
EM	parameter	s
Clusters		
Max Iteratio	n	5
Ridge		0.00100
Seed random	generator	Standar
Cluster	ing res	ults
Clusters		2
Cluster	Descriptio	n Size
cluster nº1	c_em_1	491
cluster n°2	c_em_2	509
Cluster	ing au	alitv
Criterion	Value	
Log-likelihoo	d 510.417	1
AIC	-998.834	1
BIC	-944.848	8
Mean of	folust	ore
Mean o	ciust	ers
Attribute C	luster_1 C	luster_2
		0.7862
XO	0.1121	0.7002

² For more details on the approach, see <u>http://fr.wikipedia.org/wiki/Algorithme_esp%C3%A9rance-maximisation and http://en.wikipedia.org/wiki/Expectation-maximization_algorithm</u>

We add again a SCATTERPLOT component in the diagram. We see that the results of the clustering are more in adequation with our visual insight.



Gaussian mixture model is at least as well as K-MEANS for clustering. But it is more complex, we have to estimate more parameters, and the risk of overfitting is higher, especially when we are in a high dimensional space and fewer examples.

Detecting automatically the number of clusters

Choosing the right number of clusters is a crucial question in a clustering problem. With the Gaussian mixture model, we have a criterion that we must optimize: the likelihood. So, we can evaluate various values of the number of clusters and select the best one.

In order to obtain an unbiased result, we must insert two enhancements in this basic outline: we use a resampling method (cross-validation) for obtaining an honest likelihood estimate; the likelihood must be counterbalanced with the model complexity (number of clusters) for avoiding the solutions with a high number of clusters (i.e. AIC Akaike or BIC Schwartz criteria).

The EM-SELECTION component must be inserted under the EM-CLUSTERING component. It tests various numbers of clusters and select the best one according to the chosen criterion.





In our dataset, we use a resubstitution AIC estimate. We evaluate a number of clusters from 1 to 10. The associated EM-CLUSTERING component is automatically updated.

We obtain the following results.

EM-Selection 1						
	Parameters					
Parameter	Value					
Criterion	AIC					
# clusters Start value	1					
# clusters End value	10					
Evaluation type	Resubstitution					
Folds for CV	5					
Random generator	Standard					
-						
	Results					

Criterion values w.r.t. # of clusters

# ol	f clusters	Criterion
	1	966.711776
	2	-998.834106
	3	-733.068350
	4	-702.453791
	5	-413.147625
	6	-395.647989
	7	-316.241112
	8	-211.967653
	9	-350.100880
	10	-249.646101

Computation time : 359 ms. Created at 28/08/2006 10:03:10 Selecting two clusters seems also numerically the best solution.

We can copying the results in a spreadsheet and plotting the criterion according to the number of clusters. It allows you to compare the various solutions.

Criterio	n values w.r.t. #	of clusters
# of clusters	Criterion	1500
1	966.711776	1000
2	-998.834106	
3	-733.06835	500
4	-702.453791	
5	-413.147625	0
6	-395.647989	1 2 3 4 5 6 7 9 9 10
7	-316.241112	-500
8	-211.967653	
9	-350.10088	-1000
10	-249.646101	
		-1500 J
Computation tim	ie : 359 ms.	

Conclusion

Gaussian mixture model with EM algorithm is a powerful approach for clustering. It is surprisingly rarely implemented in the data mining and statistical software.