

Assignment 5

Computational Intelligence, SS2018

Team Members		
Last name	First name	Matriculation Number
Lee	Eunseo	11739623
Shadley	Alex	11739595
Lee	Dayeong	11730321

- For your tests, select the correct number of components ($K = 3$), but also check the result when you use more or less components. How do you choose your initialization θ_0 ? Does this choice have an influence on the result

When the number of components is 2, the result is the following figure.

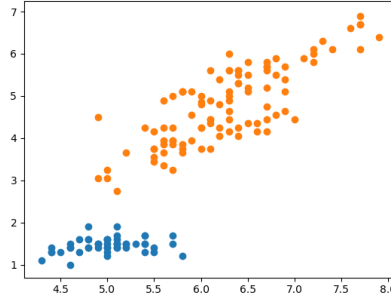


Figure 3: $K = 2$

When the number of components is 4, the result is the following figure.

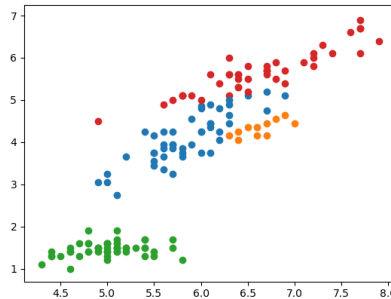


Figure 4: $K = 4$

For initialization θ_0 , we referred to the class pdf. The following figure is the reference class pdf.

4.2.1 Initialisierung

Eine Möglichkeit θ^0 zu initialisieren ist:

1. α_m^0 auf uniforme Verteilungsfunktion $\alpha_m^0 = \frac{1}{M}$
2. Σ_m^0 wird auf die Kovarianzmatrix Σ der Daten \mathbf{X} gesetzt d.h. $\Sigma = \frac{1}{N} \sum_{n=1}^N (\mathbf{x}_n - \boldsymbol{\mu})(\mathbf{x}_n - \boldsymbol{\mu})^T$ wobei $\boldsymbol{\mu} = \frac{1}{N} \sum_{n=1}^N \mathbf{x}_n$.
3. Für $\boldsymbol{\mu}_m^0$ wählt man m Samples zufällig aus oder man verwendet den k-means Algorithmus.

Figure 5: The initialization process in the class pdf

The initialization process influences the EM algorithm result. When the randomly selected points, for calculating mean value, is well chosen, the result is accurate. That is, the result is accurate when the first random selected points are from first answer label group, the second random selected points are from second answer label group and the third random selected points are from third answer label group.

- plot the log-likelihood function over the iterations! What is the behavior of this function over the iterations?

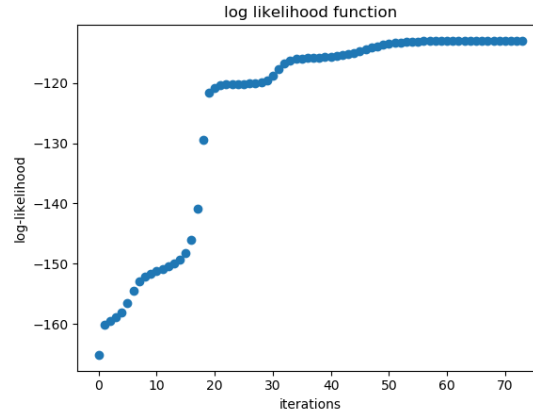


Figure 6: The log-likelihood function over iterations

As shown in Figure 4, the log-likelihood increases over iterations. That is, likelihood increased over iterations. And about 50th iteration, the function looks converging to the value, -112.96270776709655. Therefore, the process stops even though it didn't reach the max iteration number.

- Make a scatter plot of the data that shows the result of the soft-classification that is done in the E-step

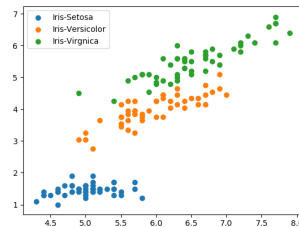


Figure 7: The EM algorithm soft-classification

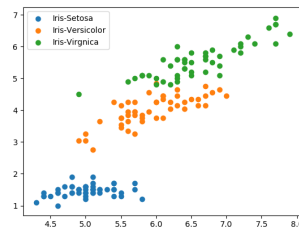


Figure 8: The answer classification

The EM algorithm classifies well the points when it is compared with the answer classification. EM algorithm fails to classify the points near the boundary of iris-Versicolor and iris-Virginica.

- 1.1.2 Perform all of the above-mentioned tasks for the K-means algorithm
- 1.1.3 You may additionally choose any other pair of features; how would this change the classification accuracy
- 1.2 4 dimensional feature
 - 1.2.1 How do the convergence properties and the accuracy of your classification change in comparison to scenario 2.1?
 - 1.2.2 Within your EM-function connect the structure of the covariance matrices to diagonal matrices! What is the influence on the result.
- 1.3 Processing the data with PCA
 - 1.3.1 How much of the variance in the data is explained this way?
 - 1.3.2 How does the performance of your algorithms compare to scenario 2.1 and scenario 2.2?
 - 1.3.3 Apply PCA with whitening, so that the transformed data has zero mean and a unit covariance matrix. How does this influence the choice of your initialization?

2 Samples from a Gaussian Mixture Model

- 2.1 Write a function $Y = \text{sample-GMM}(\alpha, \mu, \text{cov}, N)$
- 2.2 Using a GMM of your choice ($K > 3$), demonstrate the correctness of your function