Machine Learning I
Practice Session II – Clustering

1 Goals

The goal of this practice session is two-fold: (a) to build a better understanding of how the hyper-parameters of K-means, soft K-means, DBSCAN and GMMs shape the clusters. (b) to learn how to determine the optimal hyper-parameters for these methods by using BIC, AIC and the F1-measure.

The document is divided into two parts:

- Part 1 – How-to: Instructions on how to perform clustering with K-means, soft K-means, DBSCAN, GMM and how to derive the evaluation metrics (log-likelihood, BIC, AIC and F1-measure).
- Part 2 – Tasks and Questions: Set of tasks to be performed during the practical and questions you must answer.

In this second practical, we will use two datasets provided in the previous practice session:

1. Wine cultivar
2. Activity recognition

You can download and find a description of the datasets here.

2 How-to:

2.1 Use the MLDemos interface for clustering

In order to use the clustering algorithms in MLDemos, select the second box from the left in the algorithms panel (Box 1 in Fig. 1)

For this practical, you will need to use the following set of options (Fig. 1) :

1. **Clustering Tab**: Let all clustering options appear.
2. **Cluster**: It launches clustering on the set of data you have drawn or uploaded, using the clustering algorithm you have selected (Box 8) and its hyper-parameters (Box 9).
3. **Clear**: Clears the created model
4. **One iteration:** (available only for K-means and soft K-means). It executes one iteration of K-means or soft k-means algorithms. This allows you to follow the progresses of the clustering.

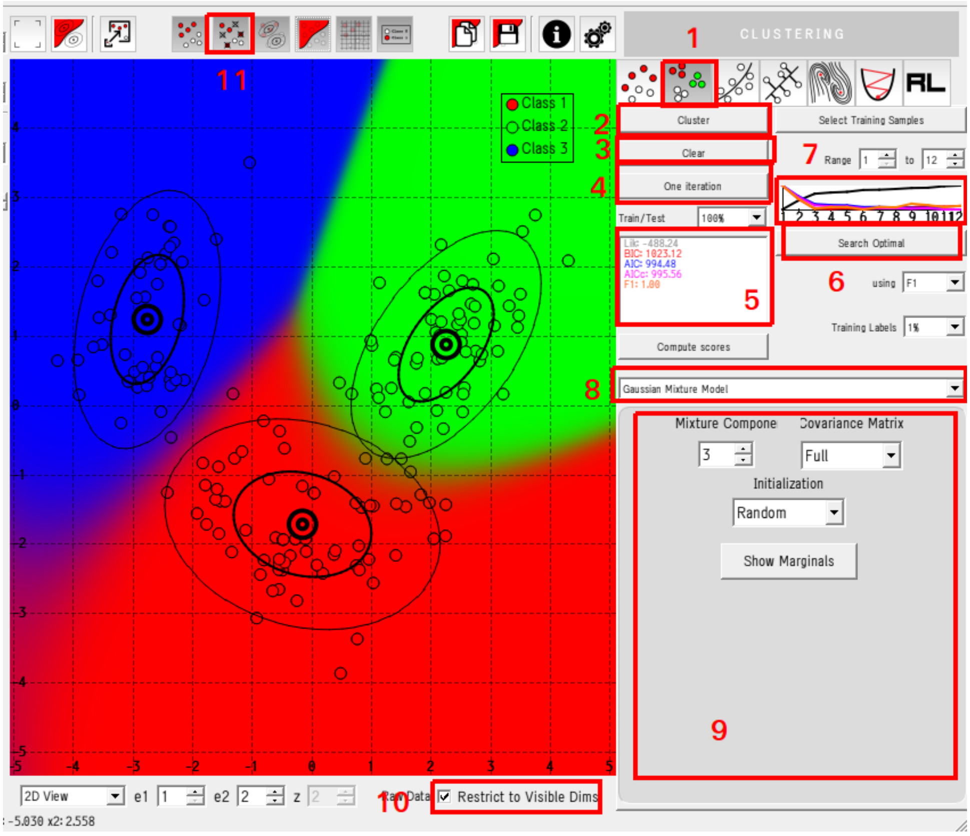

Figure 1: The MLDemos clustering interface

5. **Evaluation block** displays the values of clustering metrics (Log-likelihood / RSS, BIC, AIC, F1-measure) once clustering has converged.

6. **Search optimal:** Performs grid search for the optimal number of clusters according to each criterion (Log-lik, BIC, AIC, AICc, F1-measure)

7. **Curves of evaluation metrics:** Illustrates the curves created by the search optimal function

8. **Clustering algorithm:** Select the desired algorithm to cluster your data

9. **Hyper-parameters block:** You can set the hyper-parameters of the clustering algorithm. Also, for the case of k-means you can select the soft clustering variant from here.

10. **Restrict to visible dims:** If selected, the algorithm is applied only on the dimensions that currently appear in 2D view. Otherwise, the algorithm is applied on all the dimensions that the samples currently have.

11. **Display learned output:** When selected, it colors the points based on the color code of the cluster that they belong to. This can be turned off to get a visualization of samples that clustered together but belong to different classes .

2.2 Perform Clustering

This example demonstrates how to use the clustering interface in order to cluster the *wine* dataset with K-means and perform grid-search to find the optimal number of clusters K.

- Load the *Wine* dataset, perform PCA and reduce its dimensionality by keeping the eigenvectors that explain 90% of the variance
- Using the *Cluster tab*, select k-means with three clusters and click the *cluster* button. This will create a visualization of the clustering model on the 2D view and update the evaluation metrics (Box 5). Please note that the *lik* in Box 5 refers to the log-likelihood and thus it can be negative. You can turn the *Display learned output* off to get a better visualization of samples that clustered together but belong to different classes.
- In order to perform grid search for the optimal number of k, click the search optimal button. This will plot the BIC, AIC and F-measure curves above. The legend for the color-codes in Box 7 is given in Box 5.

3 Questions

Q1: Load the *Wine* dataset, perform PCA and keep the first two eigenvectors. Select the K-means algorithm with K=3 and the euclidean distance as similarity metric. Use the *One iteration button* (Box 4 in Fig 1)and try to predict where the means of the clusters will be located at the next step of the algorithm.

Q2: On the same data, apply the k-Means, soft k-Means, DBSCAN and GMM algorithms by varying their hyper-parameters. Answer the following questions for each algorithm:

- K-means hyper-parameters: k and similarity metric
  * What is the optimal value for k, in terms of class separation?
  * How having k larger than the optimal affects the result?
  * What is the impact of different distance metrics to the result?
  * Find the optimal set of hyper-parameters in terms of class-separation
- soft K-Means hyper-parameters: k and beta
  * How having k larger than the optimal affects the result? Is the effect of k the same as k-means?
  * Vary the beta parameter using high and low values. How and why the value of beta affects the result?
  * Find the optimal set of hyper-parameters in terms of class-separation
- DBSCAN hyper-parameters: *Max distance* (ϵ), *Min samples* and similarity metric
  * Set *Min samples=4*, the euclidean distance as similarity metric and vary the *Max distance* (ϵ). How *Max distance* (ϵ) affects the result?
  * Do the same as above but this time keep both the *Max distance* (ϵ) and similarity measurement fixed and vary the *Min samples* hyper-parameter. How the *Min samples* hyper-parameter affects the result?
  * Test different similarity measurements and mention how they affect the result.
  * Find the optimal set of hyper-parameters in terms of class-separation
- GMM hyper-parameters: Number of components and type of covariance matrix
  * Vary the number of GMM components and mention how they affect the result.
  * Select different types of covariance matrix and mention how they affect the result.
  * Find the optimal set of hyper-parameters in terms of class-separation

- Which of the above algorithms are sensitive to initialization and output different solutions at each run and why?

Q3: On the same data, use the k-means algorithm, vary the number of clusters (from 1 to 10) and create a line plot of the log-likelihood, BIC and AIC w.r.t to the number of clusters. Use the results illustrated in Box 5. of Fig 1 to get the values of log-likelihood, BIC and AIC and create a simple plot at Excel. Answer the following questions:

- How and why the evaluation metrics change with respect to the number of clusters?
- What causes the appearance of a "plateau" at the line plot of AIC and why?
- Why BIC is always larger than AIC?
- What is the optimal k based on the line plot of the evaluation metrics?

You can be helped by righting down the BIC and AIC formulas

Q4: On the same data, use the GMM algorithm with K=3 and k-means initialization. Decide which type of covariance matrix would you use based on the distribution of the data. Then compare the performance of different types of matrices using the F1-measure and BIC measurements as they appear in Box 5. of Fig 1. Which type of covariance matrix would you choose taking into account the F1-measure and which taking into account BIC? Is the choice the same for both criteria and why?