ML 2016/17
Exercise 5: Clustering

28/11/16

General information

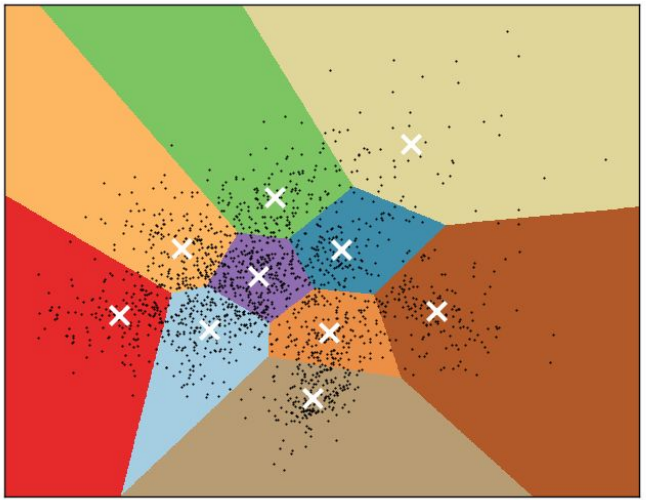
- The assignments are not graded on a scale: it's simply pass/no pass
  - If one homework is not sufficient you can simply redo it
- All assignments must be delivered one month before **you** take the exam
- Submission through email: send to fabiom.carlucci@dis.uniroma1.it
- Questions can be written to same email address.
- Office hours to meet in person: **Wednesday** at B004 (Via Ariosto, the door in front of library), 10AM-12PM.
- Python recommended: https://www.continuum.io/downloads
- *There is no need to replicate exactly the images I show!*

HW5: Clustering

We will see:

- Clustering with K-Means
- Clustering with GMM/EM
- Performance evaluation

Once you complete the experience send the report to fabiom.carlucci@dis.uniroma1.it with subject: **“[ML1617] Clustering report”**



K-Means intuition

Given a set of observations $(x_1, x_2, \ldots, x_n)$, where $x_i \in \mathbb{R}^d$

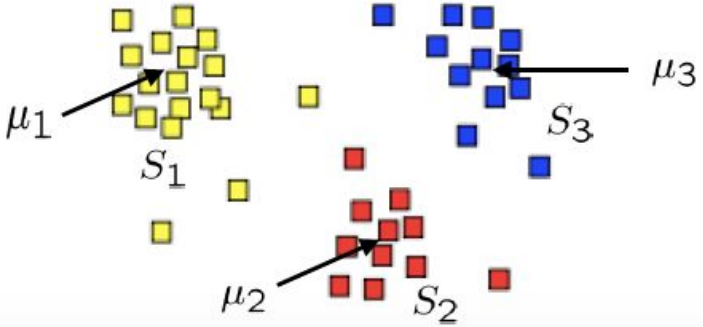
***K*-means clustering problem:**

Partition the *n* observations into *K* sets ($K \le n$) **S** = {S_1, S_2, ..., S_K} such that the sets minimize the within-cluster sum of squares:

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{K} \sum_{\mathbf{x}_j \in S_i} \left\| \mathbf{x}_j - \boldsymbol{\mu}_i \right\|^2$$

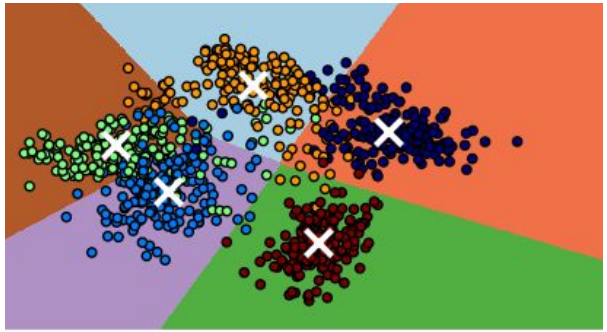
where μ_i is the mean of points in set S_i.

K=3



What to do ⅓ - K-Means

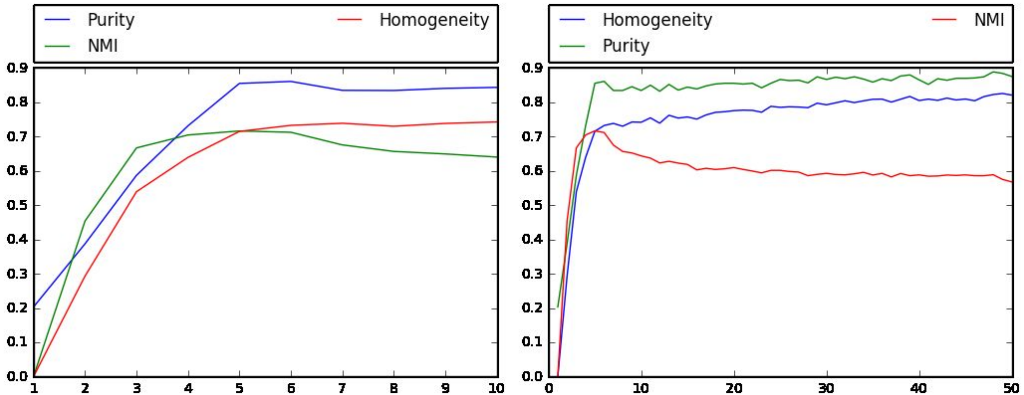
1. Load ***Digits*** dataset
2. Select all data (**X**) and labels (**y**) corresponding to classes {0, 1, 2, 3, 4}
3. Standardize and apply PCA in order to obtain 2D data
4. Cluster **X** into 5 clusters using K-Means
5. Plot **X**, the centroids and the boundaries between clusters



6. Repeat point 4 and 5, varying the number of clusters from 3 to 10

What to do ⅔ - GMM and Evaluation

7. Varying the number of clusters in {2, 3, ..., 10}
   a. apply GMM based clustering
   b. compute the cluster ***Purity*** score and plot it against the number of clusters.
   c. compute the ***Normalized Mutual Information*** score and plot it against the number of clusters
   d. compute the ***Homogeneity*** score and plot it against the number of clusters
8. Explain your observations - what is the difference between the scores we used?



Step-by-step: data loading

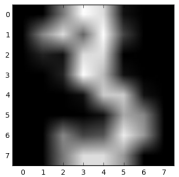
The **Digits** dataset contains 8x8px images of *digits*

Obtaining the data is easy:

```
from sklearn import datasets
digits = datasets.load_digits()
```

If you want, you can visualize some of the images:

```
plt.imshow(digits.images[3])
plt.show()
```



The flattened data is also provided:

```
X = digits.data
y = digits.target
```

You can select only the samples corresponding to certain classes by using mask indexing:

```
X = X[y<3] //get classes 0,1,2
y = y[y<3]
```

```
… apply standardization and
PCA…
```

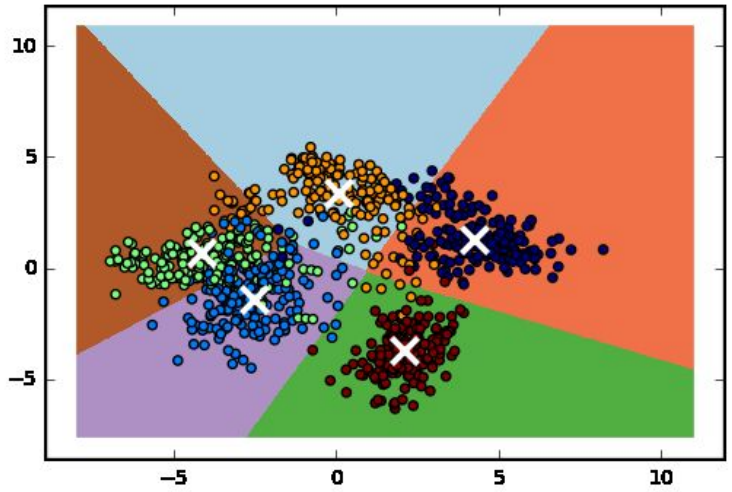
Step-by-step: K-Means and GMM

Thanks to sklearn, K-Means can be applied in a familiar way:

```
from sklearn.cluster import KMeans
kmeans = KMeans(5)
kmeans.fit(X)
```

Getting the coordinates of the centroids is also easy:

```
ccenters = kmeans.cluster_centers_
```



KMeans also has the ***predict*** function, which will help you to generate the boundaries (check past assignments for some code)

The same steps also apply for GMM

Step-by-step: performance evaluation

How to measure performances on **unlabeled** data?

Task is not trivial and many possible solutions exist:
http://scikit-learn.org/stable/modules/clustering.html#clustering-performance-evaluation

To Normalized Mutual Information score and Homogeneity score are easy to get, as sklearn does the work for us:

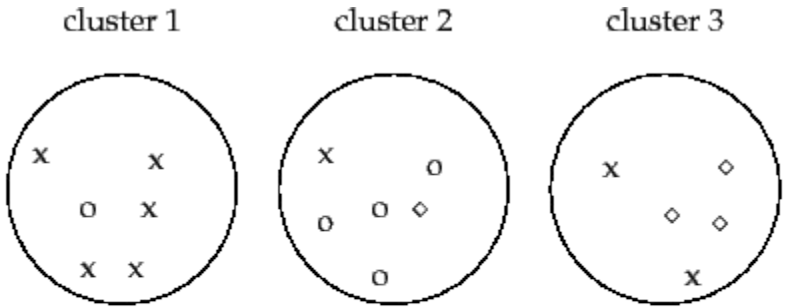
```
from sklearn.metrics import normalized_mutual_info_score, homogeneity_score
```

What about **Purity**?

Step-by-step: Purity

$$\text{purity}(\Omega, \mathbb{C}) = \frac{1}{N} \sum_k \max_j |\omega_k \cap c_j|$$

where C = {c 1 , . . . , c K } is the set of clusters and Ω = {ω 1 , . . . , ω J } is the set of classes.



► **Figure 16.1** Purity as an external evaluation criterion for cluster quality. Majority class and number of members of the majority class for the three clusters are: x, 5 (cluster 1); o, 4 (cluster 2); and ⋄, 3 (cluster 3). Purity is $(1/17) \times (5+4+3) \approx 0.71$.

Helpful Links

http://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

http://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_digits.html#sphx-glr-auto-examples-cluster-plot-kmeans-digits-py

http://scikit-learn.org/stable/modules/generated/sklearn.metrics.homogeneity_score.html

https://stats.stackexchange.com/questions/95731/how-to-calculate-purity/154379#154379

http://scikit-learn.org/stable/modules/generated/sklearn.metrics.normalized_mutual_info_score.html

Your turn now! Questions?