Antonio Norelli

Machine learning *a.a 2016/2017* *prof. Barbara Caputo*

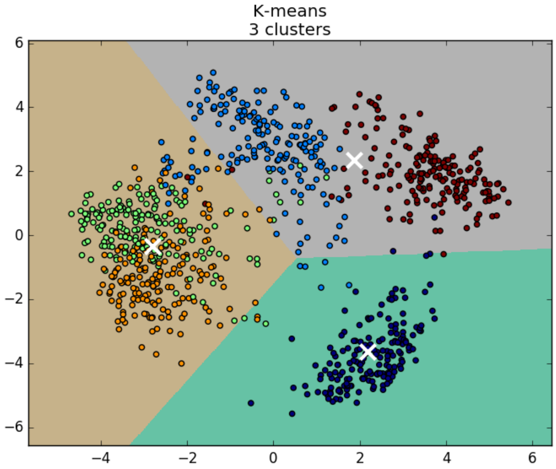
Homework 5: Clustering

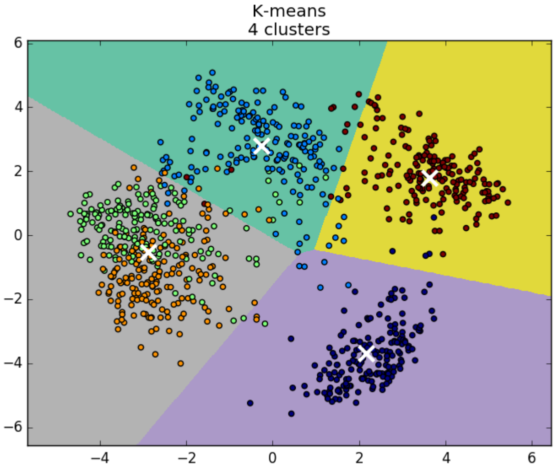
K-Means

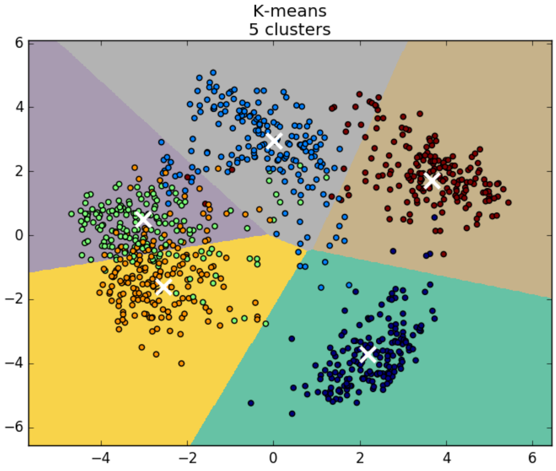
The dataset are the first two principal components of the first five classes of the digits dataset available in the sklearn's standard dataset library.

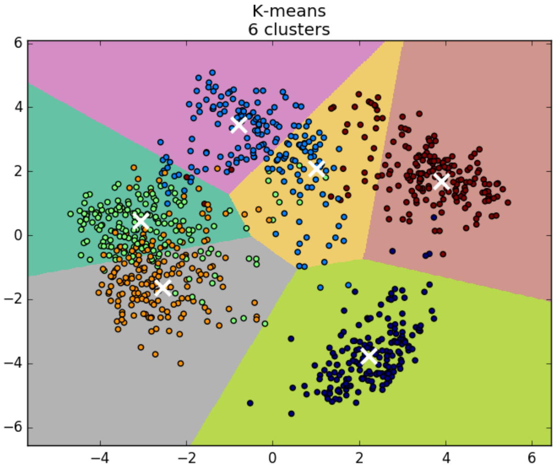
We have five classes, corresponding to the digits 0, 1, 2, 3, 4.

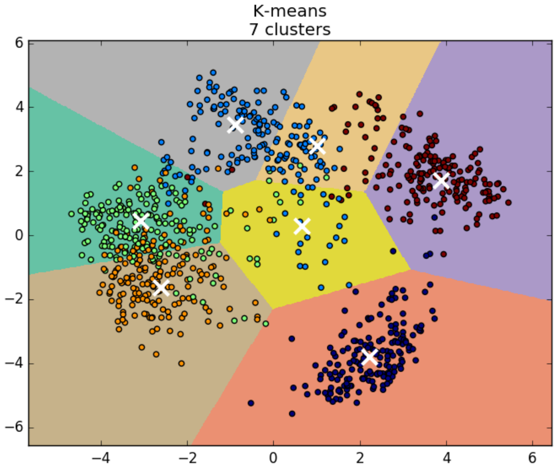
K-Means belongs to the family of unsupervised learning, we want to see the results varying the number of clusters from 3 to 10. In each plot is reported the K-means classification (decision boundaries), the true classification (colour of the points) and the centroid of each cluster (white crosses).

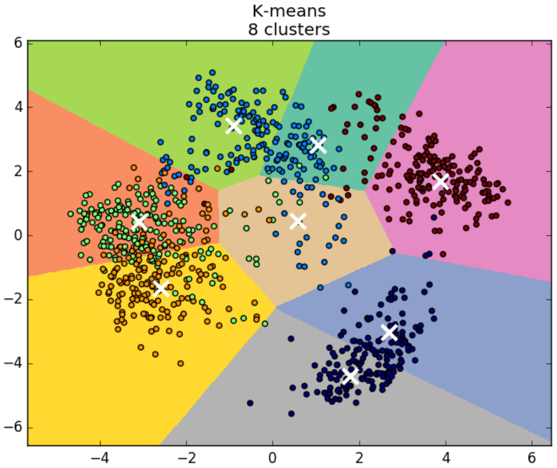


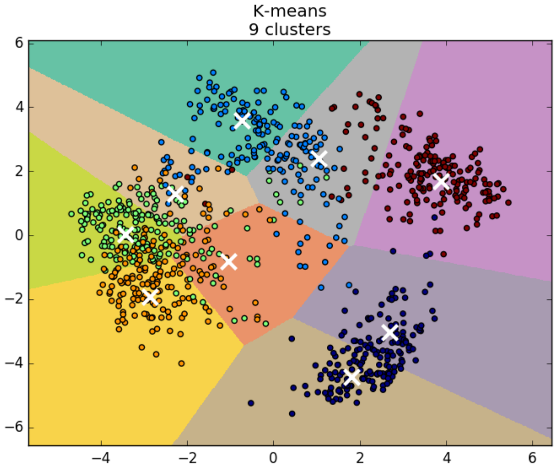


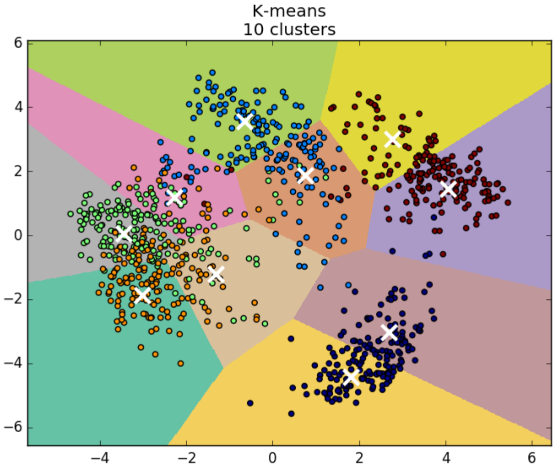






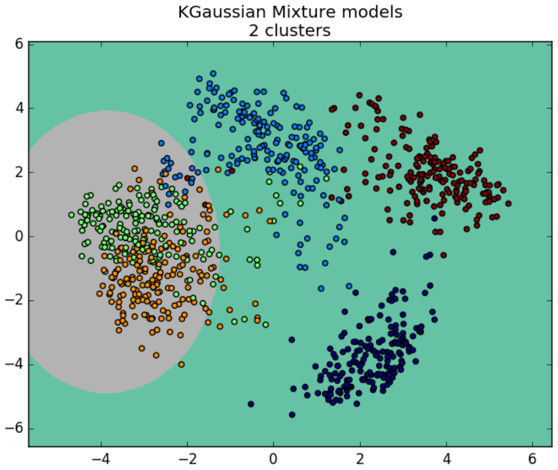


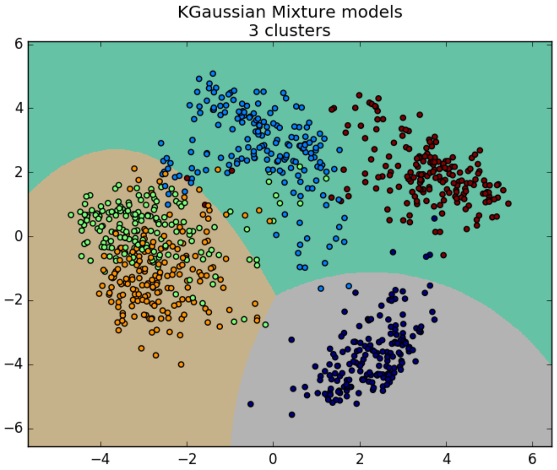


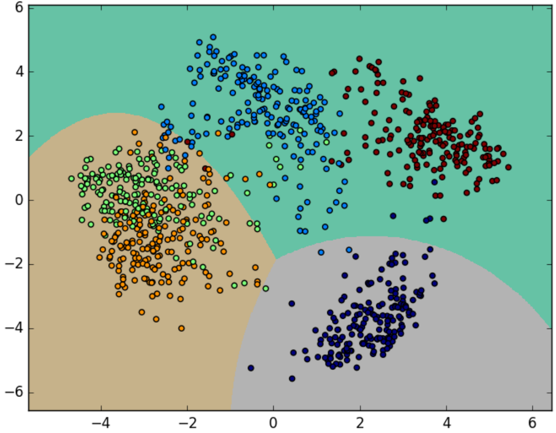


Gaussian mixture models (GMM) and Evaluation

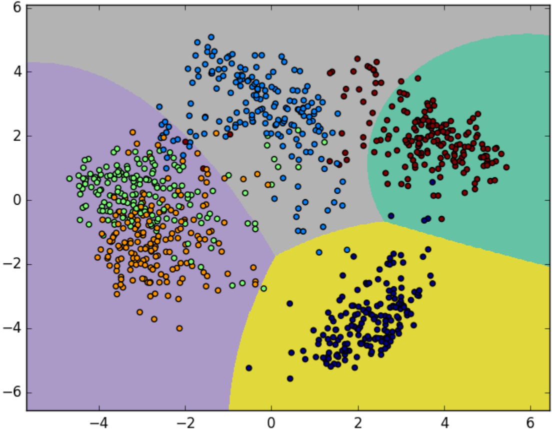
Another way to find clusters is performed using GMM, we are assuming that all the data points are generated from a mixture of a finite number of Gaussian distributions with unknown parameters. In the following plots are shown the clusters founded with this method varying the number of Gaussians. As before the colour of the points represents the true classification.



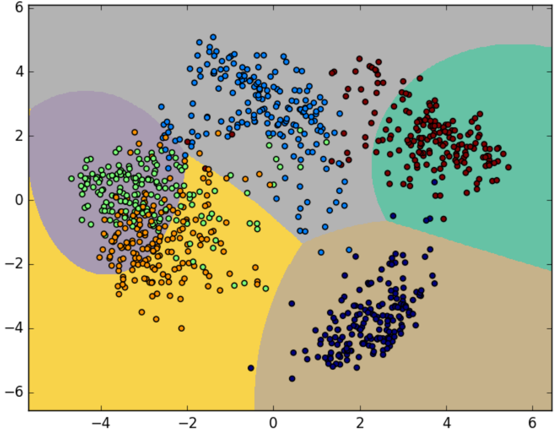




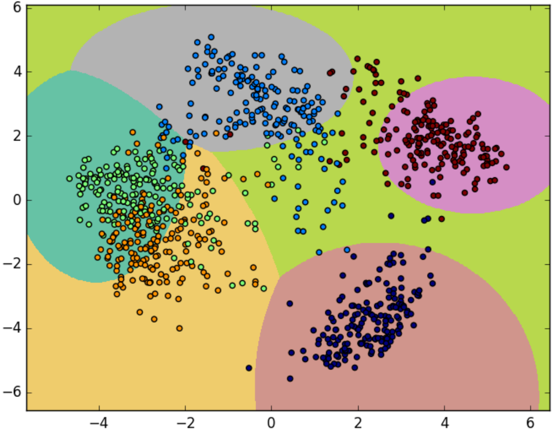
KGaussian Mixture models
3 clusters



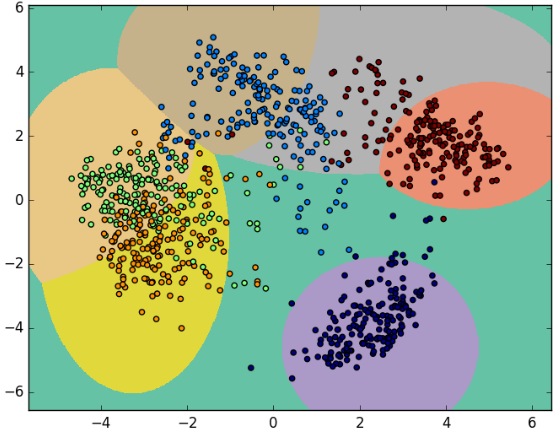
KGaussian Mixture models
4 clusters



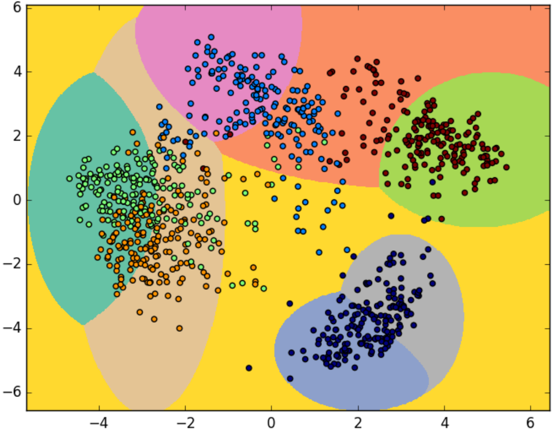
KGaussian Mixture models
5 clusters



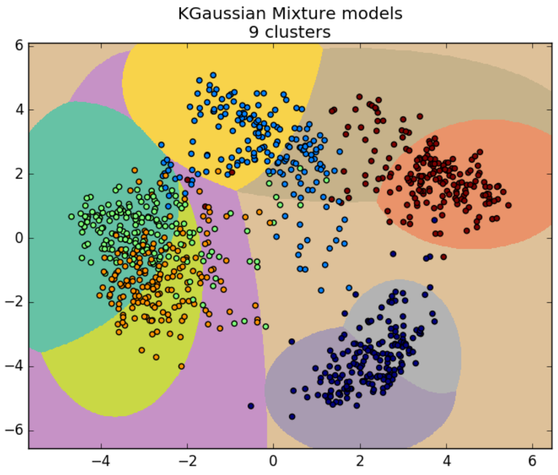
KGaussian Mixture models
6 clusters

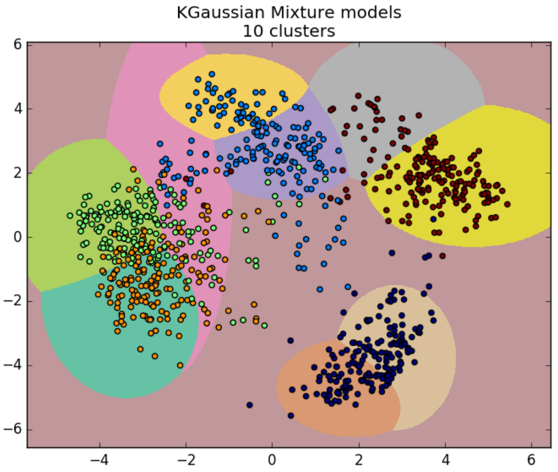


KGaussian Mixture models
7 clusters

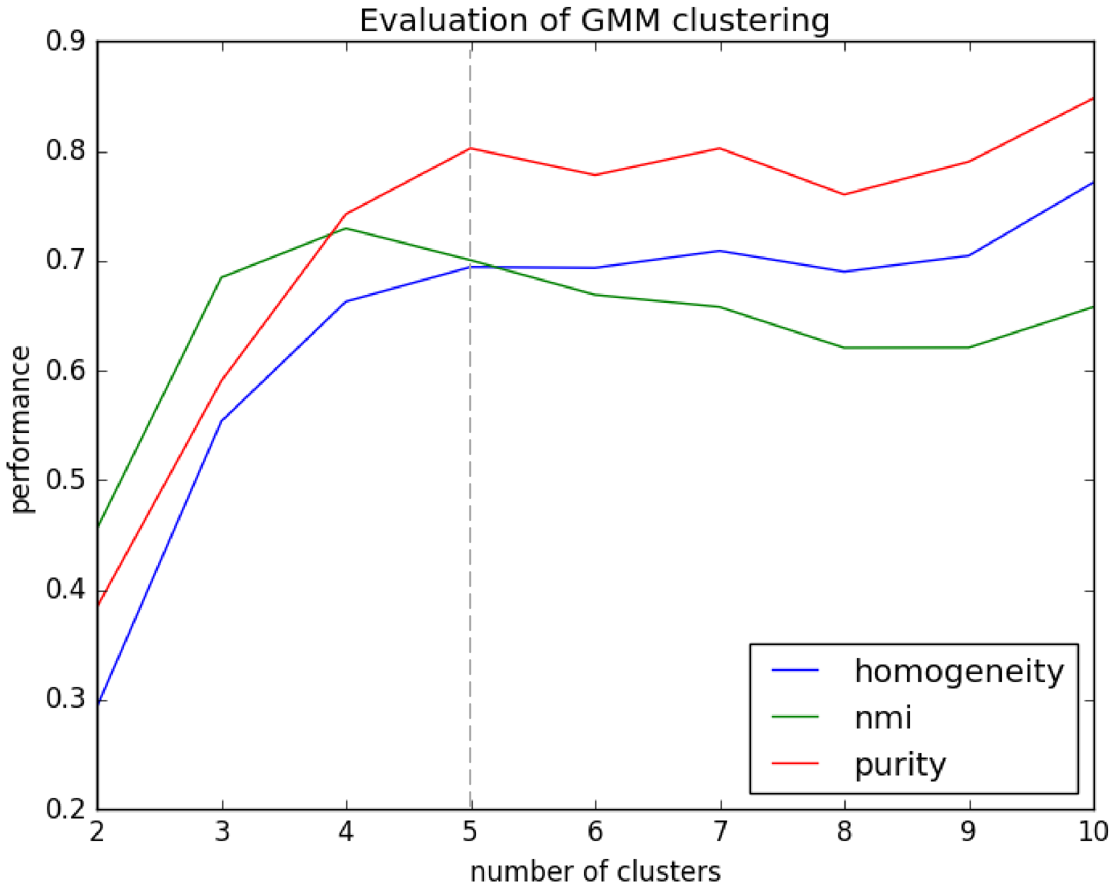


KGaussian Mixture models
8 clusters

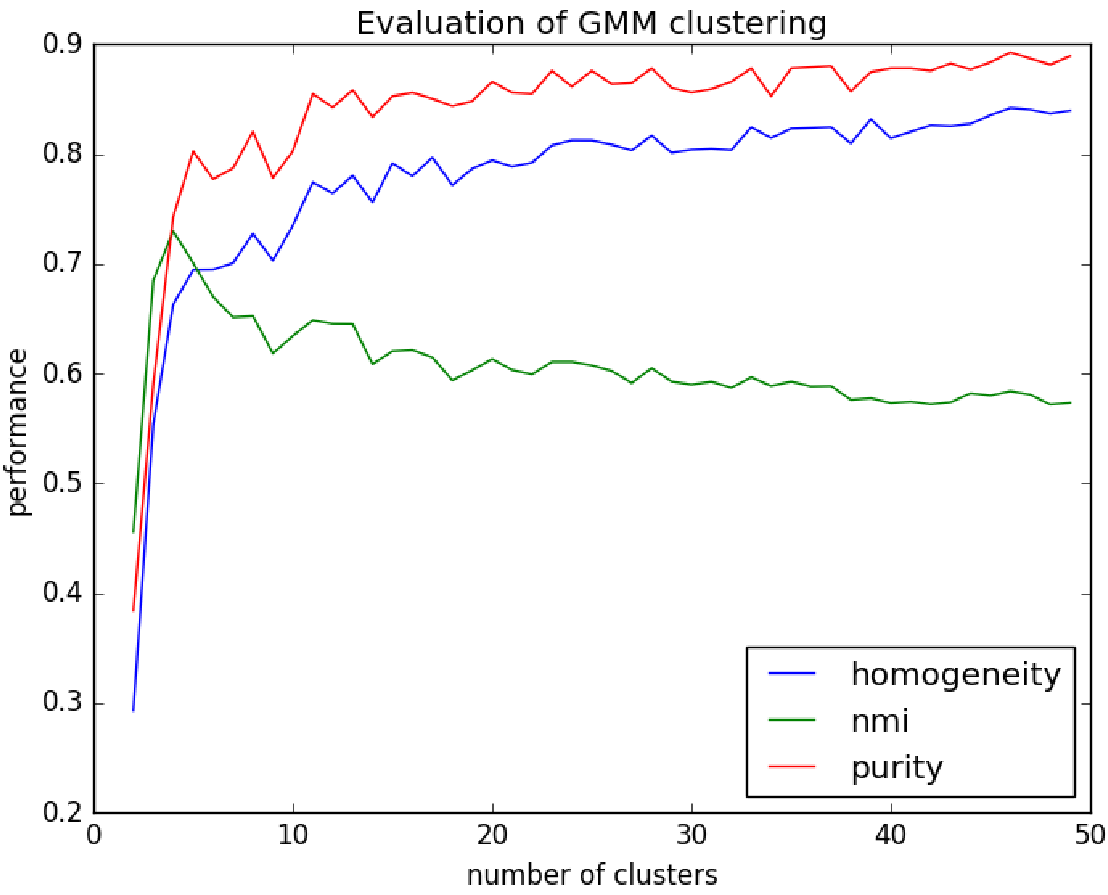




In the plot below are reported different scores for the GMM clustering with different number of Gaussians (clusters).

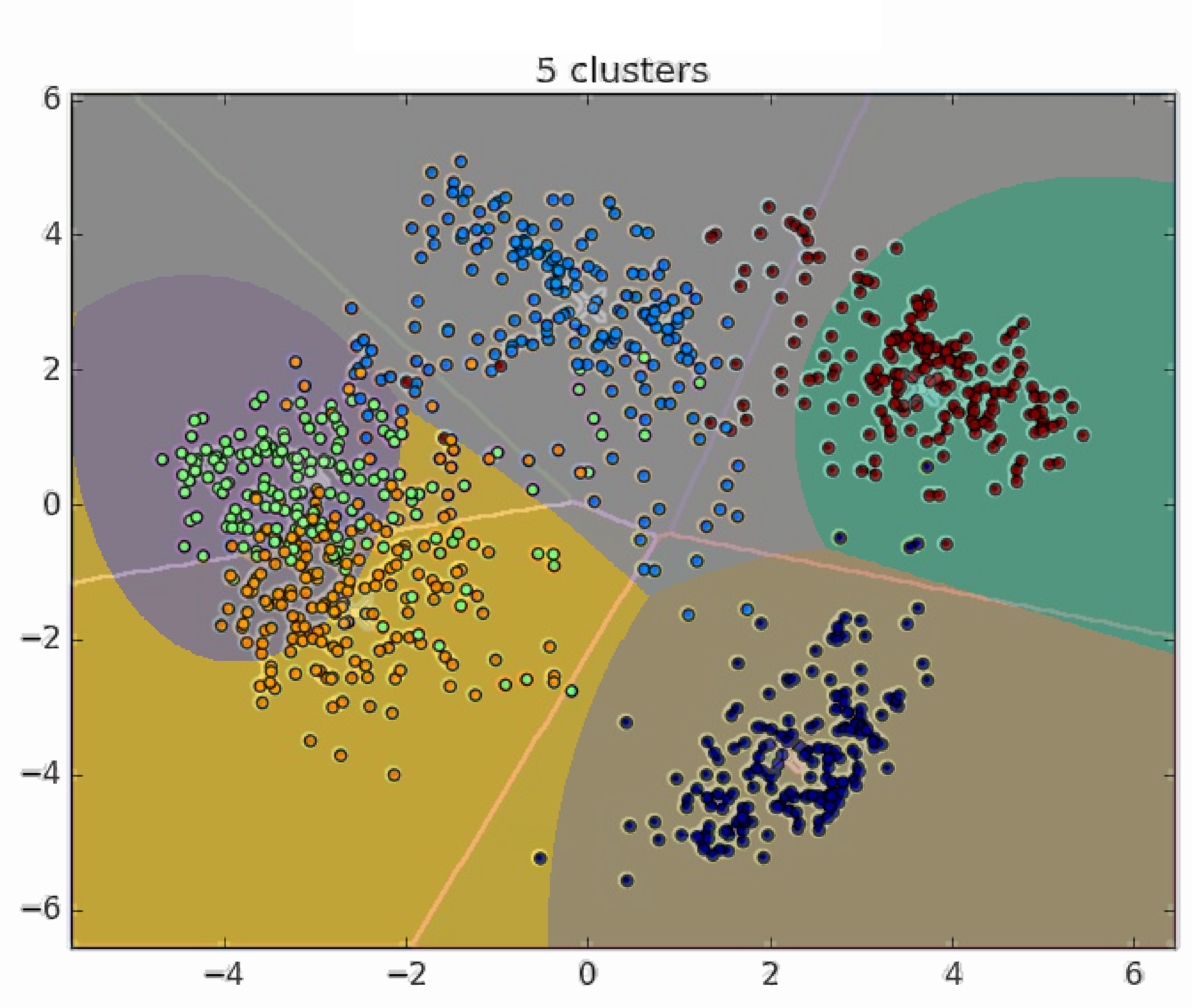


Explain your observations, what is the difference between the scores we used?
It is useful to look at the trend of the chosen performance testers with an increment of the maximum number of clusters. In the plot below is shown the same graph but with a larger limit in the number of clusters (50).



- **Homogeneity**: the homogeneity parameter measures a desirable property for the clustering, how much each cluster contains only members of a single class. Unfortunately, this parameter alone is not useful to find the best number of clusters as an extreme (for example a maximum), homogeneity is always growing (except from local fluctuations). The right way to deal with homogeneity is to look at the derivative: where the homogeneity ceases to grow fast we have a good candidate. Looking at our results a good candidate is 4 or 5, looking at the peak at 10 (multiple of 5) and the local decrease at 8 (multiple of 4) we are led to choose 5 clusters, the correct answer.
- **Normalized mutual information (NMI)**: the normalized mutual information parameter measures the agreement of the correct labels and the predicted labels ignoring the permutations of the classes or clusters, more specifically it quantifies the amount of information obtained about the true label vector, through the predicted label vector. It is a measure of mutual dependence. A too large or too small number of clusters are both penalized, so we can search for a maximum. Looking at our results the maximum is located at 4, near the correct answer.
- **Purity**: the purity parameter is very similar to the homogeneity parameter, both measure how much each cluster contains only members of a single class. The difference is in the way the result is translated into a number between 0 and 1. Probably because of this similarity this score is not implemented in the sklearn library. All the conclusions with homogeneity are still valid with purity so looking at our results we can conclude that we are led to choose 5 clusters, the correct answer.

Bonus: K-Means vs GMM



As we see in the plot K-means performance are better, a confirmation comes from one of the scores we used so far: NMI_{KMEANS}=0.76 NMI_{GMM}=0.73.