Sheet 6 - Clustering

Jan Scheffczyk - 3242317
Leif Van Holland - 2563657
Oliver Leuschner - 3205025

November 28, 2019

Practical part

Please find the solution in the accompanying .py file.

Theoretical Part

Assignment 2) Convergence of k-means

a) We perform only two operations both of which strictly decrease the cumulative distance and therefore the error function μ. First we move the cluster centers to the mean of their respective point clouds which strictly decreases the the error function, The only exception would be if all the centers are already on their means, which is the termination criteria. Secondly points only change clusters if their distance to a center is reduced as a result.

b) Given a finite number of k clusters and N points we have exactly k^N possible assignments which is a finite number.

c) As the error function is strictly decreasing with each successive step we never get the same assignment twice. As we've just seen there is only a finite amount of possible assignments. Therefore the algorithm terminates after a finite amount of steps.

Assignment 3) Expectation Maximization

a)

The result strongly depends on the initialization, as the assignment of the data points are iteratively refined in subsequent steps. A weak initialization can therefore lead to slow convergence speeds or convergence towards a subpar optimum. The refinement of a GMM with EM occurs locally, because the assignment

of a Gaussian distribution is dependent on the distance to the mean of the distribution and the following parameter maximization (M-step) is based on these local assignments.

b)

An incorrect number of clusters can cause hard-to-interpret results, as e.g. multiple clearly separated data clusters are be described by only one Gaussian or a single cluster is wrongly described by multiple (small) Gaussians.

c)

Two possibilities to improve convergence could be (a) to run the algorithm multiple times and select the best result and/or (b) to initialize it with a simple clustering algorithm like k-means.

d)

The sampling influence result/run-time, as the parameter maximization solely depends on the given data. A very sparse or ambiguous sampling can lead to slow convergence or convergence against a bad optimum.