Handin 1

Søren Meldgaard 201303712, Malthe Bisbo 201303718

December 4, 2017

STATUS OF WORK

SECTION 1

The two algorithms was compared both by visual inspection and using the two scores. The visual inspection was carried out by inspecting 2d scatter plot with the data coloured according to the cluster it was assigned to. From visual inspection Lloyd's algorithm seems most promising, as it separated the data into well defined groups, while both the basic EM and the Lloyd initialized EM had a tendency to return overlapping distributions, which resulted in split clusters, when performing the hard clustering (assigning each point to the cluster with the largest probability).

SECTION 2.1 - SILHOUETTE COEFFICIENTS

The silhouette coefficients does not wary much with either k or algorithm. For Lloyd's it scores highest for k=4 and EM scores highest for k=2.

SECTION 2.2 - F1 SCORE

Lloyd's algorithm scores the most for k=3 and EM for k=2. As the data contained three different labels it seems preferable that the algorithms also has the highest F1 score for k=3, which

Table 0.1: My caption

	k=2	k=3	k=4
Lloyd's	0.9947	0.9966	0.9970
EM	0.9955	0.9926	0.9947

Table 0.2: My caption

	k=2	k=3	k=4
Lloyd's	0.7762	0.8120	0.6509
EM	0.8260	0.6933	0.6081

is only true for Lloyd's.

SECTION 2.3 - DIFFERENCES BETWENE MEASURES

F1 is an external/supervised measure and is thus only applicable when labels are available, whereas the silhouette coefficients can always be used.

SECTION 3 - COMPRESSION

The image was compressed from 127983 bytes to 72115 bytes which gived a compression ration of 1.775.
This was done py clustering all pixels in the image and then representing each pixel with the centroid it belongs to.

SECTION 4 - GENERATING IMAGES FROM EM

In the EM algorithm the clusters represents probability distributions. When clustering the MNIST dataset, the clusters generated seem to represent the data well, as digit like figures are generated when sampling from the probability distributions.