Clustering and Embedding

In this problem set we will implement and apply the standard *K-means* and the *"online" K-means* clustering procedures. The file `cluster.dat` contains a data set of $p = 500$ (2-dimensional) observations generated from four different Gaussians with four different means.

10.1 K-means Clustering (5 points)

Write a program that implements the *standard* version of K-means clustering and partitions the given data set into K clusters. Repeat the clustering procedure for different initializations of the prototypes and $K = 2, 3, 4, 5$. Include the following steps:

Initialization –

- Set the initial prototypes $\mathbf{w}_q$ randomly around the data set mean
- Set the maximum number of iterations t_{max}, e.g. 5

Optimization – implement the k-means update (see script section 4.1.2). Each iteration should contain the followin two steps

- assign all datapoints to their closest prototype
- re-compute the new positions of the prototypes for this assignment

Plotting –

- Visualize data points and prototypes for each iteration in a sequence of scatter plots.
- Plot the error function E against the iteration number t

$$E_{\{m_q^{(\alpha)}\},\{\mathbf{w}_q\}} = \frac{1}{2p} \sum_{q,\alpha} m_q^{(\alpha)} \left\| \mathbf{x}^{(\alpha)} - \mathbf{w}_q \right\|$$

Visualization: Create a plot to show how the resulting solution assigns different regions of input space (e.g. new data points) to the different clusters.

10.2 Online K-means Clustering (5 points)

Write a program that implements the *online* version of K-means clustering and partitions the given data set into $K = 4$ clusters. Include the following steps:

Initialization –

- Set the initial prototypes $\mathbf{w}_q$ randomly around the data set mean
- Select an initial learning step η_0
- Set the maximum number of iterations t_{max}, e.g. equal to the data set size p.

Optimization –

- Choose a suitable $\tau < 1$ and implement online K-means clustering using the following "annealing" schedule for η:

$$\eta_t = \eta_0 \quad \text{for } t = 0, ..., \frac{t_{max}}{4} \quad \text{and} \quad \eta_t = \tau \eta_{t-1} \quad \text{for } t = \frac{t_{max}}{4} + 1, ..., t_{max}$$

Plotting –

- Visualize data points and the prototypes for each iteration in a sequence of scatter plots. Only include the first, the final, and four intermediate plots in your report.
- Plot the error function E against the iteration number t (see above)

Total points: 10