Exercise Sheet 9 due: 05.07.2017 at 23:55

K-means Clustering

In this problem set we will implement and apply the standard (batch) K-means algorithm, the online version, and the "soft" 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.

9.1 K-means Clustering (3 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, 6, 7, 8. Include the following steps:

Initialization -

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

Optimization -

Implement the k-means update (see lecture notes). Each iteration should contain the following two steps

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

Visualization -

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

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

(c) Create a plot (Voronoi-Tesselation) to show how the resulting solution assigns different regions of input space (e.g. new data points $\mathbf{x} \in \mathbb{R}^2$) to the different clusters.

9.2 Online K-means Clustering (3 points)

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

Initialization -

- ullet Set the initial prototypes ${f w}_q$ randomly around the data set mean
- Select an initial learning step ε_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 ε :

$$\varepsilon_t = \varepsilon_0 \quad \text{ for } t = 0, ..., \frac{t_{max}}{4} \quad \text{ and } \quad \varepsilon_t = \tau \varepsilon_{t-1} \quad \text{ for } t = \frac{t_{max}}{4} + 1, ..., t_{max}$$

Visualization -

- (a) Visualize data points and the prototypes for each iteration in a sequence of scatter plots, but only show the first, the final, and four intermediate iterations. In the final plot additionally show for each cluster the sequence of centroid positions \mathbf{w}_q by connecting them with straight lines.
- (b) Plot the error function E (as above) against the iteration number t

9.3 Soft K-means Clustering (4 points)

"Soft" clustering is a mean-field approximation of pairwise clustering with squared Euclidean distances. Implement the *soft* K-means algorithm with squared Euclidean distances (cf. lecture notes) and apply it to the same data as before. Proceed as follows:

- (a) Set K=8 initial prototypes \mathbf{w}_q randomly around the data set mean and choose a convergence tolerance γ .
- (b) For fixed β (no annealing), let the optimization procedure run until convergence, that is $\|\mathbf{w}_q^{new} \mathbf{w}_q^{old}\| < \gamma \ \forall q$. Repeat this for different $\beta \in [0.2, 20]$ e.g. in steps of $\Delta\beta = 0.2$. Use the same initial prototypes for all runs.
- (c) Visualize the data set, initial and final prototypes for each (fixed) β in one scatter plot. Therein show how "soft" each data point $\mathbf{x}^{(\alpha)}$ is assigned to a cluster, e.g. by scaling the brightness of the respective plot symbol with the largest assignment probability, i.e., $\max_q \langle m_q^{(\alpha)} \rangle$.
- (d) Plot in two separate subplots the first and second coordinate of the final prototypes \mathbf{w}_q against the β (i.e. K lines per subplot) and interpret the result.
- (e) In additional simulations, run the optimization for K=2,4,6,8 using an annealing schedule: increase β after each iteration. E.g. $\beta_0=0.2,\ \tau=1.1,\ \beta_{t+1}=\tau\beta_t$.
- (f) Show the data set, initial and final prototypes of the "annealed" clustering solutions for K=2,4,6,8 in a scatter plot. How "soft" are data points assigned now?

Total points: 10