Homework set on GMM, EM and Kmeans

Guidelines

- The solution for this homework is to be posted as a .pdf file.
- You may choose the programming language you prefer for the implementations.
- All plots must have named axis, grids and title. If more than one plot is on the same figure, provide legend.

GMM implementation

A Gaussian Mixture Model (GMM) is a statistical model. It assumes that the observations were generated from a distribution of the form:

$$f_X(x) = \sum_{z=1}^{K} p_Z(z)\mathcal{N}(x|\mu_z, \Sigma_Z),$$

where $\mathcal{N}(x|\mu, \Sigma)$ is a Gaussian distribution with expectation μ and covariance matrix Σ. The random variable Z is called a latent variable.

1. Data Generation: generate synthetic data from a Gaussian mixture model with two Gaussians. Use the following parameters:

$$\begin{aligned}
\mu_1 &= [-1, -1]^T, \\
\mu_2 &= [1, 1]^T, \\
\Sigma_1 &= \begin{pmatrix} 0.8 & 0 \\ 0 & 0.8 \end{pmatrix}, \\
\Sigma_2 &= \begin{pmatrix} 0.75 & -0.2 \\ -0.2 & 0.6 \end{pmatrix}, \\
P_Z(z = 1) &= 0.7.
\end{aligned}$$

   If you are using Python, you can use *numpy.random.multivariate_normal* function. Alternatively, you can simply draw a uniform random variable in [0, 1] and transform it to Bernoully. Then, based on the outcome, draw a multi-dimensional normal random variable with the corresponding parameters. Scatter 1000 points of the generated data, using scatter plots.

2. K-Means implementation:
   (a) Generate 50 samples from the distribution above and plot them.
   (b) Implement a K-Means algorithm with two centers. You may start the algorithm with 2 random points.

(c) Plot the results after each iteration (the centers and which points belong to which center)

(d) Repeat the experiment with different initializations.

3. EM implementation:

    (a) generate 10000 samples from the distribution you created. This will be used as the realization of the distribution.

    (b) Implement the Expectation Maximization (EM) algorithm to fit a GMM of two Gaussians to the generated data. Try different initialization methods (Kmeans and random samples).

    (c) Plot the log-likelihood function value of each iteration. Set the horizontal axis to iteration number, and the vertical axis to the log-loss.

    (d) Plot the data and both of the Gaussain's contour* on the same figure.

    (e) Repeat (b)-(d) using three Gaussians instead of two.