Homework 8, due March 6th, 11:59pm

March 9, 2024

In all homeworks, you are required to include in your report the code that you implemented. If you use some code from the web or package, also mention in your report the origin of the code. The code should be included as searcheable text, not as a picture.

1. We will experiment with k-means and EM for clustering to see their advantages and disadvantages. To evaluate the clustering result we will use two measures. One is the accuracy, computed as follows:

- First compute the contingency matrix, which is a 2D histogram of the $(y_i, \hat{y}_i)$ combinations. See `sklearn.metrics.cluster.contingency_matrix` for details.
- We will assign the cluster labels to the true labels by finding a permutation of the labels that maximizes the sum of the diagonal elements on the resulting contingency matrix. For that we will solve the linear sum assignment problem, see `scipy.optimize.linear_sum_assignment` for details on the contingency matrix. Be aware that the default linear sum assignment is a minimization, but we want to maximize the assignment of clusters to labels.
- Finally we obtain the accuracy score as the value of the linear sum assignment (the sum of the diagonal elements of the permuted contingency matrix) divided by the number of observations.

The other one is the Adjusted Rand Index (from `sklearn.metrics.adjusted_rand_score`). We will also use the KL divergence between distributions, which for two Gaussians $P = (\mu_1, \Sigma_1)$ and $Q = (\mu_2, \Sigma_2)$ in dimension d is

$$D_{KL}(P \parallel Q) = \frac{1}{2}\{\log \frac{|\Sigma_2|}{|\Sigma_1|} - d + \mathrm{Tr}(\Sigma_2^{-1}\Sigma_1)\} + (\mu_i - \mu)^T \Sigma_2^{-1}(\mu_1 - \mu_2)/2$$

where $\mathrm{Tr}(A)$ is the trace of A (sum of elements on the diagonal).

a) Generate a set X_Q of 500 observations $x_i \in \mathbb{R}^2, xi \sim Q = \mathcal{N}(0, \sigma^2 I_2)$ with $\sigma = 3$ and labels $y_i = 0$. For each a in $\{0, 1, 2, 3, 4\}$ perform the following steps:

1. Generate set X_a of 500 observations in $x_i \in \mathbb{R}^2, xi \sim P_a = \mathcal{N}(\mu, I_2)$ with $\mu = (a, 0)$ and label $y_i = 1$. Merge X_a and X_Q to obtain a dataset of 1000 observations in $\mathbb{R}^2$.
2. Perform k-means and EM clustering on this dataset using 10 different random initializations (runs). Plot the clustering results obtained by k-means and EM for one of the runs for $a = 0$ as two different graphs (one for k-means and one for EM).
3. Compute the accuracy and Adjusted Rand Index obtained for each run.

On the same graph, plot the accuracy obtained for each run vs a as separate dots, with different colors for k-means (say red) and EM (say black). Since you have 10 runs and 5 datasets you will have 50 red dots and 50 black dots on the plot.

On a separate graph, plot the Adjusted Rand Index obtained for each run vs a as separate dots, with different colors for k-means and EM. (4 points)

b) Repeat 10 times (runs) the following steps:

1. Generate a 2×2 random matrix M with $M_{ij} \sim \mathcal{N}(0, 1)$. Use SVD `numpy.linalg.svd` to decompose M as $UDV^T = M$. We obtain this way a random rotation matrix U.
2. Compute the covariance matrix $\Sigma = U \begin{pmatrix} 100 & 0 \\ 0 & 1 \end{pmatrix} U^T$. Generate a dataset with 500 points $x_i \in \mathbb{R}^2$ from $x_i \sim Q = \mathcal{N}(0, \Sigma)$ and 500 points $x_i \in \mathbb{R}^2$, $x_i \sim P = \mathcal{N}(\begin{pmatrix} 10 \\ 0 \end{pmatrix}, \Sigma)$. Also compute the KL divergence $D_{KL}(P \parallel Q)$.
3. Run k-means with isotropic covariance matrices (or with no covariance modeling), k-means with full covariance matrices and EM clustering on this dataset, obtaining the accuracy and Adjusted Rand Index of the three methods.
4. Plot the three clustering results obtained at step 3 as three different graphs (one for each method) for the first four runs.

Plot the accuracy vs. KL divergence $D_{KL}(P \parallel Q)$ for the three methods on the 10 runs obtained above as dots with three different colors on the same graph. You will have a total of 30 dots in the plot. Similarly, plot the Adjusted Rand Index vs. KL divergence $D_{KL}(P \parallel Q)$ for the three methods on the 10 runs obtained above as dots with three different colors. Report the same results in a table with the 10 runs as rows and the three methods' accuracies and Adjusted Rand indexes and KL divergences as columns. (5 points)