DDA 3020 Assignment 4

April 30, 2023

Homework due: 11:59 pm, May 14, 2023. Note that we have reduced the number of questions in both the written and programming assignments such that you could have more time to prepare for the final exam. This assignment accounts for 15/100 of the final score.

1 Written Problems (7 points)

1. (3 points) EM for mixtures of Bernoullis.
   (1) Show that the M step for maximum likelihood estimation of a mixture of Bernoullis is given by

   $$\mu_{kj} = \frac{\sum_i r_{ik} x_{ij}}{\sum_i r_{ik}}$$

   Hint1: The distribution of a mixture of Bernoulli's is given by:

   $$p(x_i|\mu_k) = \prod_{j=1}^{D} \mu_{kj}^{x_{ij}} (1-\mu_{kj})^{1-x_{ij}}$$

   Hint2: μ_{kj} is given by :

   $$\mu_{kj} = \arg\max_{\mu_{kj}} L(q;\mu) \tag{1}$$

   Where $L(q;\mu)$ is the Auxiliary function (check ppg.17 in L16 slide).

   (2) Show that the M step for MAP estimation of a mixture of Bernoullis with a $\beta(\alpha, \beta)$ prior is given by

   $$\mu_{kj} = \frac{(\sum_i r_{ik} x_{ij}) + \alpha - 1}{(\sum_i r_{ik}) + \alpha + \beta - 2}$$

   Hint3: $\beta(\alpha, \beta)$ prior is given by:

   $$\beta(\alpha, \beta, \mu_{kj}) = A\mu_{kj}^{\alpha-1}(1-\mu_{kj})^{\beta-1} \quad \left(\because A = \frac{(\alpha+\beta-1)!}{(\alpha-1)!(\beta-1)!}\right)$$

Hint4: μ_{kj} is given by :

$$\mu_{kj} = \arg\max_{\mu_{kj}} L(q;\mu) + \sum_j \sum_k \log \beta(\alpha, \beta, \mu_{kj}) \tag{2}$$

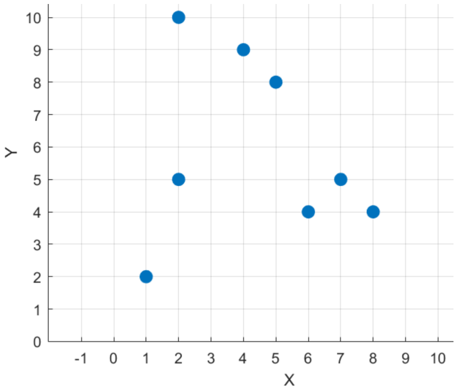

Figure 1: Training data set for K-means clustering

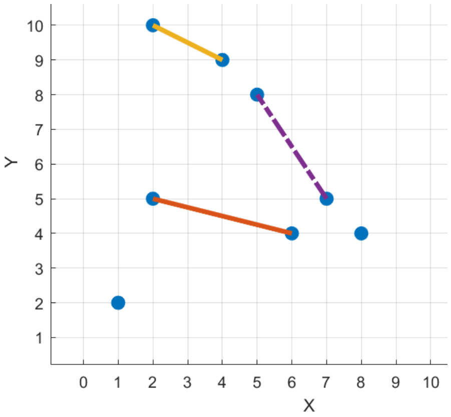

Figure 2: Training data set for constrained K-means clustering

2. (2 points) Given the data set as shown in Figure 1 and assume that points $A1$, $A4$ and $A7$ are chosen to be the initialized cluster centers. The coordinates of the data points are:

$$A1 = (2, 10), \quad A2 = (2, 5), \quad A3 = (8, 4), \quad A4 = (5, 8),$$
$$A5 = (7, 5), \quad A6 = (6, 4), \quad A7 = (1, 2), \quad A8 = (4, 9).$$

(1) (1 points) Use the K-means algorithm and Euclidean distance to cluster the 8 data points into K = 3 clusters.
(2) (1 points) Consider the case that there exist 2 must links and 1 cannot link as shown in Figure 2, derive the 3 clusters now.

3. (2 point) Consider the following 10 data points: $X = \{(7,4,3),(4,1,8),(6,3,5),(8,6,1),(8,5,7),(7,2,9),(5,3,3), (9,5,8),(7,4,5),(8,2,2)\}$. Calculate the projection of each data onto a two-dimensional subspace(i.e.K=2) using PCA. You could use python or matlab or online calculator to obtain eigenvectors and eigenvalues. You should show each step of deriving the result.(You don't need to show the calculation)

2 Programming using Python (8 points)

Task: Clustering on UCI seed dataset, which can be downloaded from `https://archive.ics.uci.edu/ml/datasets/seeds`. The number of clusters is set as 3.
You need to:

1. Implement **PCA** and **K-means from scratch** (*i.e.*, no third-party or off-the-shelf package or library are allowed). You should first calculate the projection of each data point onto a two-dimensional subspace using PCA. Explain briefly your source codes in the report. (5 points)

2. Implement 2 evaluation metrics including **Silhouette Coefficient** and **Rand Index from scratch** (*i.e.*, not calling off-the-shelf package) to evaluate the performance of above clustering algorithms. (3 points)

Note that you should submit A4_StudentID.pdf (report, together with the written answers), and A4_StudentID.ipynb (code). Please zip them into "A4_StudentID.zip". The reference report is in Assignment 1. You can check it on BlackBoard. (You can submit several files in one submission. Don't submit them in different submissions.) Your report for the programming questions should include necessary formulas, charts, and explanations. The number of pages should be 4-5.