Assignment 2, Frequent Itemset Mining and Representative Clustering: 50 points

Due September 27, 2024, 11:59 PM CST

Instructions

Your homework submission must cite any references used (including articles, books, code, websites, and personal communications). All solutions must be written in your own words, and you must program the algorithms yourself. Submit your solutions as a PDF to Canvas: `https://canvas.tamu.edu/`.

Your programs must be written in Python. The relevant code to the problem should be submitted as a separate file (*e.g.*, Python file (.py), Jupter/Google Colab Notebook (.ipynb)). Please **do not** include screenshots of your code in your submission. If a problem involves programming, then the code should be shown as part of the solution to that problem. If you solve any problems by hand just digitize that page and submit it (make sure the problem is clearly labeled and legible). **Clearly** label your figures and tables.

If you have any questions, please reach out to Dr. Peeples.

1 Frequent Itemset Mining- 20 points

Given the database in Table 1, solve the following problems. Please complete by **hand** or **typeset**:

Table 1: Transaction database for Problem 1.

tid	Itemset
t_1	ABCD
t_2	ACDF
t_3	ACDEG
t_4	ABDF
t_5	BCG
t_6	DFG
t_7	ABG
t_8	CDFG

1. Convert Table 1 to binary database.
2. Convert Table 1 to vertical database.
3. Using minimum support of 3, show how the Apriori algorithm enumerates all frequent patterns from this dataset (*i.e.*, find $\mathcal{F}^{(3)}$, hint: be sure to show all of your steps and clearly label each step, *e.g.*, C_1, L_1).
4. Using minimum support of 2, show how the FPGrowth algorithm enumerates all frequent patterns from this dataset (*i.e.*, find $\mathcal{F}^{(2)}$, hint: generate FP-tree and conditional pattern base).

2 k-Means 10 points

Your k-Means clustering algorithm has just assigned the points

$$\mathbf{x}_1 = (0, 3), \mathbf{x}_2 = (3, 3), \mathbf{x}_3 = (0, 0)$$

to cluster centroid at location $\mathbf{c}_1 = (3.5, -1)$. Answer the following questions. Please complete by **hand** or **typeset**. Show your work and clearly state equations used in your calculations.

1. Compute the sum of squared errors for the initial clustering assignments.
2. When it completes the next iteration of the algorithm, where will the centroid be located?

3 Gaussian Mixture Model- 10 points

Given a dataset, $\mathbf{D}$: $[1.0, 1.3, 2.2, 2.6, 2.8, 5.0, 7.3, 7.4, 7.5, 7.7]^T$, cluster this dataset using a Gaussian Mixture Model with 2 clusters. The initial random means, variances, and mixture weights/prior probabilities:

Table 2: Initial parameters of Gaussian Mixture Model.

$\mu_1 = 6.63$	$\sigma_1 = 1$	$P(C_1) = 0.5$
$\mu_2 = 7.57$	$\sigma_2 = 1$	$P(C_2) = 0.5$

Please answer the following questions. You may either solve by hand, typeset, or creating python script:

1. Compute the log-likelihood of the initial clustering assignments.
2. Expectation (E-step): Find the cluster posterior probability (*i.e.*, w_{ij}) for each data point.
3. Maximization (M-step): Update the parameters of the model:

(a) Means

(b) Variances

(c) Mixture weights/prior probabilities

4 Cluster Evaluation - 10 points

Please access the Iris Flower Dataset via Sklearn and select the following two attributes: sepal length and sepal width.

1. Plot the two features with the associated class labels for each data point.
2. Use k-Means to cluster the data into three clusters (set the random state to be 0 and use the default settings of the algorithm in Sklearn).

   (a) Plot the two features with the associated cluster labels for each data point.

   (b) Report the Silhouette index for the resulting clustering assignments.

   (c) How does the cluster assignment compare to the class labels? Does the Silhoutte index provide insight into the clustering assignments vs the class labels? Why or why not?
3. Experiment with different values of k (*i.e.*, 2 to 50) and plot the Silhouette index as a function of k. Is there a value of k that performs better than $k = 3$?
4. (Extra credit: up to 2 points) Use k-Mediods to cluster the dataset with $k = 3$ and plot the cluster assignments. How does this compare to k-means?