I) Important Topics to review

1. Data Exploration
   1. Basic concepts: types of attributes, objects, etc.
   2. Proximity measures
   3. Basic Statistical Descriptions of Data: mean, mode, median, quartile, IQR, etc.
   4. Data Visualization: box plot, scatter plot
2. Data Preprocessing
   1. Basic concepts: basic tasks in data preprocessing
   2. Data Cleaning methods
   3. Data Integration: redundancy and correlation analysis
   4. Data Reduction: main ideas of Principle Component Analysis (PCA), attribute subset selection
3. Finding frequent itemsets
   1. Basic concepts: support, confidence, frequent items
   2. Apriori algorithm
   3. FP-tree algorithm
   4. Comparison between the above algorithms
4. Classification
   1. Naïve Bayes
   2. Decision Trees: GINI index, entropy, information gain, how to build decision trees
   3. Support Vector Machines:
   4. KNN
   5. Artificial Neural Networks and Backpropagation
   6. Evaluation measures: Accuracy, Precision, Recall, F1-measure, Sensitivity, etc.
5. Clustering
   1. Basic concepts: centroids, dendrogram,
   2. K-means
   3. K-medoids
   4. Hierarchical clustering: agglomerative clustering
   5. Gaussian Mixture Model
   6. DBSCAN
   7. Clustering with constraints
   8. Biclustering
   9. Evaluation: B-cubed precision, B-cubed recall, Silhouette coefficients, etc.
6. Outlier Detection
   1. Basic concepts: outliers, types of outliers
   2. Univariate outlier detection
   3. Multivariate outlier detection
   4. Distance-based outlier detection
7. Advanced Topics
   1. Map-Reduce programming model
   2. Streaming Model
   3. Bloom Filter
   4. Locality Sensitive Hashing

II. Exercises:

1) Given two objects represented by tuples (22, 1, 42, 10) and (20, 0, 36, 8)

1. Compute the Euclidean distance between two objects
2. Compute the Manhattan distance between two objects

2) A *partitioning* variation of Apriori subdivides the transactions of a database *D* into *n* nonoverlapping partitions. Prove that any itemset that is frequent in *D* must be frequent in at least one partition of *D*.

3) A database has five transactions. Let *min sup* = 60% and *min conf* = 80%.

T1: {L, I, O, N}

T2: {T, I, G, E, R}

T3: {M, A, K, E}

T4: {M, U, C, K, Y}

T5: {C, O, O, K, I, E}

Find all frequent itemsets using Apriori and FP-growth, respectively.

4) (Contributed by Hang Yu Deng)

K-means algorithm: prove that given point assignments for a cluster, the mean of the points is the desired centroid that minimizes the inter-cluster variance.

5) (Contributed by Hang Yu Deng)

In DBSCAN algorithm, if p doesn't have enough neighborhoods, then it is marked as noise. But what if p's neiborhood contains a core object q? Naturally, p should be added to the cluster of the core object q. Explain whether DBSCAN can obtain this property.

6) (Contributed by Hang Yu Deng)

In Biclustering, please explain why do we choose to minimize the mean-squared residual values?

7) Suppose that the data mining task is to cluster the following eight points (with (x, y) representing location) into three clusters.

 A1(2, 10), A2(2, 5), A3(8, 4), B1(5, 8), B2(7, 5), B3(6, 4), C1(1, 2), C2(4, 9).

The distance function is Euclidean distance. Suppose initially we assign A1, B1, and C1 as the center of  each cluster, respectively. Use the k-means algorithm to show only the three cluster centers after the first round of execution