Assignment 3 [Part 1 of 2]

Questions 1-6 are pen-and-paper exercises (brief answers and justifications are expected). Questions 7-8 are coding exercises. In all responses, please show your workings (equations, justifications, or code when applicable).

1. What is the advantage of using the Apriori algorithm in comparison with computing the support of every subset of an itemset in order to find the frequent itemsets in a transaction dataset? [0.5 marks out of 5]
2. Let $\mathcal{L}_1$ denote the set of frequent 1-itemsets. For $k \geq 2$, why must every frequent k-itemset be a superset of an itemset in $\mathcal{L}_1$? [0.5 marks out of 5]
3. Let $\mathcal{L}_2 = \{\{1, 2\}, \{1, 4\}, \{2, 3\}, \{2, 4\}, \{3, 5\}\}$. Compute the set of candidates $\mathcal{C}_3$ that is obtained by joining every pair of joinable itemsets from $\mathcal{L}_2$. [0.5 marks out of 5]
4. Let S_1 denote the support of the association rule $\{\text{boarding pass}, \text{passport}\} \Rightarrow \{\text{flight}\}$. Let S_2 denote the support of the association rule $\{\text{boarding pass}\} \Rightarrow \{\text{flight}\}$. What is the relationship between S_1 and S_2? [0.5 marks out of 5]
5. What is the support of the rule $\{\} \Rightarrow \{\text{Eggs}\}$ in the transaction dataset used in Section 1 of this lab notebook? [0.5 marks out of 5]
6. In the transaction dataset used in the tutorial presented above, what is the maximum length of a frequent itemset for a support threshold of 0.2? [0.5 marks out of 5]
7. Implement a function that computes the Kulczynski measure of two itemsets $\mathcal{A}$ and $\mathcal{B}$. Use your function to compute the Kulczynski measure for itemsets $\mathcal{A} = \{\text{Onion}\}$ and $\mathcal{B} = \{\text{Kidney Beans}, \text{Eggs}\}$ in the transaction dataset used in this lab notebook. [1 mark out of 5]
8. Implement a function that computes the imbalance ratio of two itemsets $\mathcal{A}$ and $\mathcal{B}$. Use your function to compute the imbalance ratio for itemsets $\mathcal{A} = \{\text{Onion}\}$ and $\mathcal{B} = \{\text{Kidney Beans}, \text{Eggs}\}$ in the transaction dataset used in this lab notebook. [1 mark out of 5]

Assignment 3 [Part 2 of 2]

For your answers to the assignment, please include include your workings (e.g. equations, code) when this is relevant to the question. Questions 1-2 are pen-and paper exercises. Question 3 can be addressed either on paper or using code. Questions 4-5 are coding exercises.

1. For an application on credit card fraud detection, we are interested in detecting contextual outliers. Suggest 2 possible contextual attributes and 2 possible behavioural attributes that could be used for this application, and explain why each of your suggested attribute should be considered as either contextual or behavioural. [1 mark out of 5]

2. Assume that you are provided with the University of Wisconsin breast cancer dataset from the Week 3 lab, and that you are asked to detect outliers from this dataset. Additional information on the dataset attributes can be found online. Explain one possible outlier detection method that you could apply for detecting outliers for this particular dataset, explain what is defined as an outlier for your suggested approach given this particular dataset, and justify why would you choose this particular method for outlier detection. [1 mark out of 5]

3. The monthly rainfall in the London borough of Tower Hamlets in 2019 had the following amount of precipitation (measured in mm, values from January-December 2018): {22.93, 20.69, 25.75, 23.84, 25.34, 3.25, 23.55, 28.28, 23.72, 22.42, 26.83, 23.82}. Assuming that the data is based on a normal distribution, identify outlier values in the above dataset using the maximum likelihood method. [1 mark out of 5]

4. Using the stock prices dataset used in sections 1 and 2 of this lab notebook, estimate the outliers in the dataset using the one-class SVM classifier approach. As input to the classifier, use the percentage of changes in the daily closing price of each stock, as was done in section 1 of the notebook. Use the same SVM settings as in the lab notebook. Plot a 3D scatterplot of the dataset, where each object is color-coded according to whether it is an outlier or an inlier. Also compute a histogram and the frequencies of the estimated outlier and inlier labels. In terms of the plotted results, how does the one-class SVM approach for outlier detection differ from the parametric and proximity-based methods used in the lab notebook? What percentage of the dataset objects are classified as outliers? [1 mark out of 5]

5. This question will combine concepts from both data preprocessing and outlier detection. Using the house prices dataset from Section 3 of this lab notebook, perform dimensionality reduction on the dataset using PCA with 2 principal components (make sure that the dataset is z-score normalised beforehand, and remember that PCA should only be applied on the input attributes). Then, perform outlier detection on the pre-processed dataset using the k-nearest neighbours approach using k=2. Display a scatterplot of the two principal components, where each object is colour-coded according to the computed outlier score. [1 marks out of 5]