Machine Learning

Bộ môn Khoa học dữ liệu
Khoa Công nghệ thông tin
Trường Đại học Công nghiệp thành phố Hồ Chí Minh-IUH

Bài 1:
Cho bộ dữ liệu với 5 giao dịch của các mặt hàng đã mua như sau:

```
[['Milk', 'Onion', 'Nutmeg', 'Kidney Beans', 'Eggs', 'Yogurt'],
 ['Dill', 'Onion', 'Nutmeg', 'Kidney Beans', 'Eggs', 'Yogurt'],
 ['Milk', 'Apple', 'Kidney Beans', 'Eggs'],
 ['Milk', 'Unicorn', 'Corn', 'Kidney Beans', 'Yogurt'],
 ['Corn', 'Onion', 'Onion', 'Kidney Beans', 'Ice cream', 'Eggs']]
```

hãy thực hiện các yêu cầu:

1.1. Cài đăt thư viện mlxtend trên Colab (!pip install mlxtend)
1.2. Đọc dữ liệu ở trên
1.3. Mã hóa dữ liệu ở trên
1.4. Sử dụng thư viện mlxtend cho giải thuật Apriori với min_support = 0.6, in ra kết quả
1.5. Chạy lại câu 1.4 sử dụng thêm tính năng use_colnames=True, in ra kết quả
1.6. Lấy kết quả từ câu 1.5 là đầu vào, in ra các kết quả với điều kiện mặt hàng có độ dài >=2
1.7. Lấy kết quả từ câu 1.5 là đầu vào, in ra các kết quả với điều kiện mặt hàng có độ dài >=2 và support >= 0.8

Bài 2:
Cho bộ dữ liệu như ở dưới đây, hãy thực hiện các yêu cầu sau:

2.1. Đọc dữ liệu trong bảng đó
2.2. Mã hóa dữ liệu
2.3. Sử dụng thư viện mlxtend cho giải thuật FPgrowth với min_support = 0.6, in ra kết quả
2.4. Chạy lại câu 2.3 sử dụng thêm tính năng use_colnames=True, in ra kết quả
2.5. Lấy kết quả từ câu 2.4 là đầu vào, in ra các kết quả với điều kiện mặt hàng có chữ 2 ký tự cuối có giá trị >=13

Transaction ID	List of items
T100	I11, I12, I15
T200	I12, I14
T300	I12, I13
T400	I11, I12, I14
T500	I11, I13
T600	I12, I13
T700	I11, I13
T800	I11, I12, I13, I15
T900	I11, I12, I13

Bài 3:

Trong bài tập này, ta sẽ khai thác các tập phổ biến (frequent itemset) trên tập dữ liệu Plants (sự phân bố của một số loài thực vật ở khu vực Mỹ và Canada) sử dụng ứng dụng Weka.

3.1. Đọc dữ liệu Plants trên Colab, hiển thị một số thông tin cơ bản như:

- ✓ Tổng số các loài cây
- ✓ Số lượng vùng phân bố
- ✓ Vùng phân bố có ít loài cây nhất
- ✓ Vùng phân bố có nhiều loài cây nhất
- ✓ Số loài cây trung bình phân bố trên từng vùng

3.2. Cài đặt ứng dụng Weka
3.3. Viết đoạn code trên Colab để mã hóa dữ liệu ở trên, lưu kết quả vào 1 file (hint: cần chuyển sang dạng nhị phân để có thể thao tác trên Weka).
3.4. Sử dụng thuật toán Apriori trong Weka để khai thác tất cả tập hạng mục có độ phổ biến (thông số lowerBoundMinSupport) từ 0.1 trở lên, lưu các kết quả đạt được trên 1 file.
3.5. Sử dụng thuật toán FP-Growth trong Weka để khai thác tất cả các luật kết hợp. Với độ tin cậy (Confidence từ menu sổ xuống metricType) từ 0.95 (thông số minMetric) trở lên, lưu các kết quả đạt được trên 1 file.
3.6. Nhận xét về các kết quả 3.4 và 3.5, vẽ biểu đồ trực quan để so sánh, đánh giá kết quả của 2 thuật toán này trên Colab.