Đề thi:

MACHINE LEARNING WITH PYTHON

Thời hạn nộp bài: 21h30 ngày 27/01/2022

*** HV tạo 1 thư mục **LDS6_k272_HoVaTen_Cuoi_ky** trong thư mục LDS6_k272_HoVaTen để lưu tất cả bài làm vào để chấm điểm ***

*** HV sẽ bị trừ điểm nếu bài làm giống nhau ***

Chú ý, với mỗi câu :

- HV cần kiểm tra xem dữ liệu đã sạch, chuẩn và dùng được hay chưa, nếu chưa thì cần tiền xử lý dữ liệu trước khi làm bài.
- Lần lượt thực hiện các bước làm bài theo quy trình đã được hướng dẫn làm demo/bài tập trong lớp.
- Mỗi câu là một file viết trên jupyter notebook, các yêu cầu nhận xét kết quả trong từng câu được viết trong cell dưới định dạng Markdown.
- Mỗi câu đều phải đưa ra nhận xét, giải pháp cho các lựa chọn.
- Câu nào có phần trực quan hóa kết quả thì vừa phải trực quan vừa phải giải thích.

1. Clustering

- Cho dữ liệu **bbc** (giải nén file bbc-fulltext). Đọc tất cả dữ liệu vào dataframe **news** theo định dạng sau:

id (số thự tự tự động)	**content** (nội dung của tin tức trong từng tập tin)	class (một trong 5 loại : business, entertainment, politics, sport, tech) – là tên của folder chứa tin tức
0	…	…
…	…	…

- Yêu cầu : Chuẩn hóa dữ liệu (nếu cần) và **chọn một thuật toán thích hợp** để thực hiện việc **phân cụm dữ liệu** dựa trên cột **content** của dataframe **news** trên.
    1. Áp dụng thuật toán thích hợp. Dựa trên cơ sở nào để cho rằng thuật toán này thích hợp?
    2. Tìm kết quả => có bao nhiêu cụm => mẫu (tin tức) nào thuộc cụm nào?
    3. Nhận xét trên từng cụm: mỗi cụm có những keywords nào ? Vẽ WordClouds cho từng cụm.
    4. Kiểm chứng lại với **class** đã có

2. Groceries dataset

- Cho dữ liệu **ItemList.xlsx**
- Yêu cầu: Đọc dữ liệu, chuẩn hóa dữ liệu (nếu cần) và **chọn một thuật toán thích hợp trong nhóm Association rule learning** để tính toán mức độ kết hợp giữa các item.
    1. Áp dụng thuật toán (tự lựa chọn các tham số phù hợp cho thuật toán, lưu ý với số lượng transaction càng nhiều thì các ngưỡng nên đặt càng nhỏ). In kết quả. Vẽ biểu đồ.
    2. Tìm kiếm thông tin từ kết quả: trong thông tin kết quả có 'sausage' không? Nếu có thì 'sausage' kết hợp với item nào?
    3. Cho biết 15 sản phẩm được mua nhiều nhất. Vẽ biểu đồ.
    4. Cho biết 15 sản phẩm được mua ít nhất. Vẽ biểu đồ.

3. Pen-Based Recognition of Handwritten Digits

- Cho dữ liệu **penbased-5an-nn.csv**
- Mô tả dữ liệu:

General information

Pen-Based Recognition of Handwritten Digits data set			
Type	Classification	Origin	Real world
Features	16	(Real / Integer / Nominal)	(0 / 16 / 0)
Instances	10992	Classes	10
Missing values?			No

Attribute description

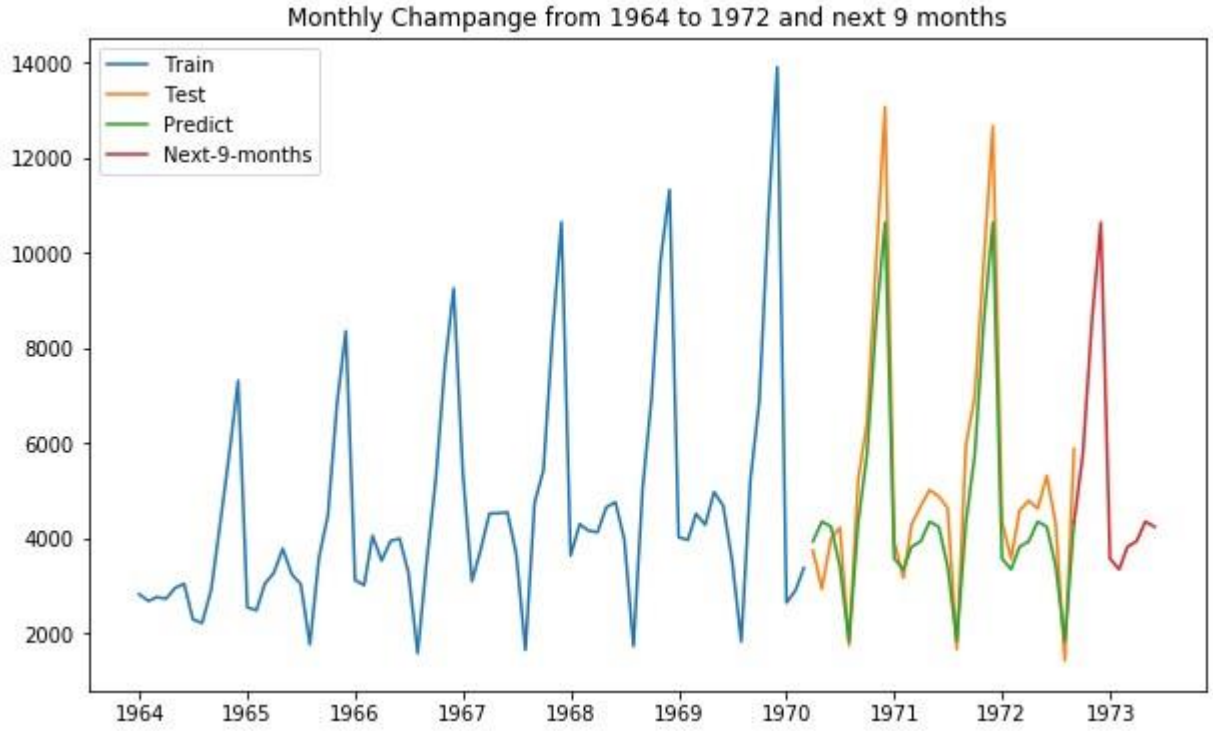
Attribute	Domain	Attribute	Domain
At1	[0, 100]	At9	[0, 100]
At2	[0, 100]	At10	[0, 100]
At3	[0, 100]	At11	[0, 100]
At4	[0, 100]	At12	[0, 100]
At5	[0, 100]	At13	[0, 100]
At6	[0, 100]	At14	[0, 100]
At7	[0, 100]	At15	[0, 100]
At8	[0, 100]	At16	[0, 100]
Class	{0,1,2,3,4,5,6,7,8,9}		

- Yêu cầu 1: Đọc dữ liệu, chuẩn hóa dữ liệu (nếu cần) và chọn một thuật toán thích hợp để thực hiện việc xác định một mẫu là loại (class) nào (trong các loại 0, 1, 2, 3, 4, 5, 6, 7, 8, 9) dựa trên các thông tin được cung cấp.
    1. Áp dụng thuật toán thích hợp để xây dựng model. Dựa trên cơ sở nào để cho rằng thuật toán này thích hợp?
    2. Đánh giá model dựa trên train/test.
    3. Trực quan hóa kết quả (nếu có). Đưa ra một số nhận xét dựa trên kết quả.
- Yêu cầu 2: Hãy **áp dụng thuật toán PCA và thuật toán đã chọn** ở Yêu cầu 1 để thực hiện việc xác định một mẫu là loại (class) nào (trong các loại 0, 1, 2, 3, 4, 5, 6, 7, 8, 9) dựa trên các thông tin được cung cấp. **Nhận xét kết quả giữa việc có áp dụng PCA và không áp dụng PCA**.

4. Monthly champagne sales millions

- Cho dữ liệu **champagne_new.xlsx** là dữ liệu bán champagne theo thời gian từ tháng 01-1964 đến tháng 09-1972
- Yêu cầu: Đọc dữ liệu, chuẩn hóa dữ liệu (nếu cần) và **chọn một thuật toán Time Series thích hợp** để thực hiện việc **dự đoán số tiền champagne bán được của 9 tháng tiếp theo** từ tháng 10-1972 đến tháng 06-1973 giá trị dựa trên các thông tin được cung cấp
    1. Thực hiện Decomposition, trực quan hóa, nhận xét.
    2. Tạo dữ liệu train/test với train chiếm 75% dữ liệu, test chiếm 25% dữ liệu.

3. Áp dụng thuật toán phù hợp.
4. Tìm kết quả.
5. Trực quan hóa kết quả (trong biểu đồ có cả train, test, predict và next_9_months) như gợi ý sau:



--- Chúc các bạn làm bài tốt ☺ ---