Đề thi:

MACHINE LEARNING WITH PYTHON

Hạn chót nộp bài: 08h30 ngày 10/12/2019

(Học viên ký tên nhưng không nộp bài sẽ nhận 1.0đ (Một điểm,

Học viên không ký tên sẽ không có điểm)

**** Học viên (HV) tạo 1 thư mục là LDS6_HoVaTen, lưu tất cả bài làm vào để nộp 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.
- Cần hiển thị thông tin chung của dữ liệu bằng cách dùng shape, head(), tail(), info()… để có cái nhìn ban đầu về dữ liệu.
- Trong dữ liệu có thể có rất nhiều thông tin (cột), cần xác định xem cột nào thật sự cần thiết dùng trong thuật toán thì đưa vào, cột nào không cần thiết thì không đưa vào.
- Với thuật toán chỉ áp dụng cho dữ liệu số, thì cần chuyển các cột dữ liệu chuỗi thành dữ liệu số tương ứng.
- Lần lượt thực hiện các bước làm bài như đã được hướng dẫn làm 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 cần phải thực hiện việc trực quan hóa kết quả.
- *Tổng số điểm của bài thi là 10.0 điểm*

1. Asian and Indian Cuisines (1.5 điểm)

- Cho dữ liệu ***asian_indian_recipes.csv*** chứa công thức (thành phần) nấu ăn của các món ăn một số nước châu Á như Hàn Quốc, Nhật Bản, Trung Quốc, Thái và Ấn Độ.
- Yêu cầu: Đọc dữ liệu, chuẩn hóa dữ liệu (nếu cần) và **chọn một thuật toán thích hợp** để thực hiện việc xác định một mẫu có **"cuisine" là nước nào** dựa trên các thông tin được cung cấp.

    1. Tạo X_train, X_test, y_train, y_test từ dữ liệu đã đọc và chuẩn hóa, với tỷ lệ dữ liệu test là 0.3
    2. Áp dụng thuật toán thích hợp để xây dựng model. Dựa trên cơ sở nào để cho rằng thuật toán này thích hợp?
    3. Đánh giá model dựa trên train/test.
    4. Trực quan hóa kết quả (nếu có). Đưa ra một số nhận xét dựa trên kết quả.
    5. Dựa trên kết quả, hãy cho biết nếu trong thành phần của món ăn **có** *"cumin"* và "*fish*" nhưng **không** có *"yogurt"* thì món ăn đó thường là món ăn của nước nào?
    6. Dựa trên kết quả, hãy cho biết nếu trong thành phần của món ăn **có** *cumin* nhưng **không** có "*fish*" *và* ***không*** *có "soy_sauce"* thì món ăn đó thường là món ăn của nước nào?
    7. In nội dung **confusion matrix** sau đó trả lời các câu hỏi sau :
        a. Tỷ lệ % các công thức món ăn của Nhật Bản được dự đoán chính xác ?

b. Tỷ lệ % các công thức món ăn của Hàn Quốc bị gán sai nhãn thành Nhật Bản?

c. Nước nào có tỷ lệ % các công thức món ăn bị gán nhãn sai nhiều nhất ?

2. Connectsionist Bench (Sonar, Mines vs. Rocks) (1.0 điểm)

- Cho dữ liệu ***sonar.all-data.txt***
- Yêu cầu: 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 **mine (M)** hay **rock (R)** dựa trên các thông tin được cung cấp.

1. Tạo X_train, X_test, y_train, y_test từ dữ liệu đọc và được chuẩn hóa, với tỷ lệ dữ liệu test là 0.3
2. Áp dụng thuật toán thích hợp để xây dựng model. Dựa trên cơ sở nào để cho rằng thuật toán này thích hợp?
3. Đánh giá model dựa trên train/test.
4. Trực quan hóa kết quả (nếu có). Đưa ra một số nhận xét dựa trên kết quả.
5. Hãy cho biết với X_new (từng giá trị ứng với từng cột trong dataset) như sau thì mẫu nào là mine (M), mẫu nào là rock (R) ?

X_new =
[[0.0123,0.0309,0.0169,0.0313,0.0358,0.0102,0.0182,0.0579,0.1122,0.0835,0.0548,0.0847,0.2026,0.2557,0.1870,0.2032,0.1463,0.2849,0.5824,0.7728,0.7852,0.8515,0.5312,0.3653,0.5973,0.8275,1.0000,0.8673,0.6301,0.4591,0.3940,0.2576,0.2817,0.2641,0.2757,0.2698,0.3994,0.4576,0.3940,0.2522,0.1782,0.1354,0.0516,0.0337,0.0894,0.0861,0.0872,0.0445,0.0134,0.0217,0.0188,0.0133,0.0265,0.0224,0.0074,0.0118,0.0026,0.0092,0.0009,0.0044],
[0.0203,0.0121,0.0380,0.0128,0.0537,0.0874,0.1021,0.0852,0.1136,0.1747,0.2198,0.2721,0.2105,0.1727,0.2040,0.1786,0.1318,0.2260,0.2358,0.3107,0.3906,0.3631,0.4809,0.6531,0.7812,0.8395,0.9180,0.9769,0.8937,0.7022,0.6500,0.5069,0.3903,0.3009,0.1565,0.0985,0.2200,0.2243,0.2736,0.2152,0.2438,0.3154,0.2112,0.0991,0.0594,0.1940,0.1937,0.1082,0.0336,0.0177,0.0209,0.0134,0.0094,0.0047,0.0045,0.0042,0.0028,0.0036,0.0013,0.0016]]

(Chú ý: Nếu cần có thể đọc thêm thông tin tại: http://archive.ics.uci.edu/ml/datasets/Connectionist+Bench+%28Sonar%2C+Mines+vs.+Rocks%29)

3. Computer Hardware (1.5 điểm)

- Cho dữ liệu ***machine.data.txt***
- Yêu cầu: Đọc dữ liệu, chuẩn hóa dữ liệu (nếu cần) và **chọn một thuật toán thích hợp** để thực hiện việc dự đoán giá trị **ERP (estimated relative performance),** là **cột cuối cùng (cột 9)** trong dataframe dựa trên các thông tin được cung cấp (loại bỏ các cột dữ liệu nếu không cần thiết)

1. Tạo X_train, X_test, y_train, y_test từ dữ liệu đọc và được chuẩn hóa với tỷ lệ dữ liệu test là 0.3.
2. Áp dụng thuật toán thích hợp để xây dựng model. Dựa trên cơ sở nào để cho rằng thuật toán này thích hợp?
3. Đánh giá model dựa trên train/test.
4. Trực quan hóa kết quả (nếu có). Đưa ra một số nhận xét dựa trên kết quả.
5. Với X_new như sau : X_new = [['amdahl','470v/7b',29,8000,32000,32,8,32,172], ['sperry','1100/83',50,2000,32000,112,52,104,307]] thì ERP lần lượt là bao nhiêu?

(Chú ý: Nếu cần có thể đọc thêm thông tin tại: http://archive.ics.uci.edu/ml/datasets/Computer+Hardware)

4. Clustering (1.5 điểm)

- Cho dữ liệu ***data3.txt***
- Yêu cầu: Đọc dữ liệu, chuẩn hóa dữ liệu (nếu cần) và **chọn một thuật toán thích hợp** để thực hiện việc **phân cụm dữ liệu** dựa trên hai cột trong dữ liệu được cung cấp.
    1. Vẽ biểu đồ thể hiện mối quan hệ giữa hai cột dữ liệu nói trên. Cho nhận xét dựa trên biểu đồ.
    2. Áp dụng thuật toán thích hợp. Dựa trên cơ sở nào để cho rằng thuật toán này thích hợp?
    3. Tìm kết quả => có bao nhiêu cụm => mẫu nào thuộc cụm nào?
    4. Nhận xét trên cụm.
    5. Với X_new như sau : X_new = [[5.5, 9.5], [16.5, 15.5], [30.0, 10.0], [29.375, 3.0]] thì mỗi mẫu sẽ lần lượt thuộc cụm nào?
    6. Vẽ hình (với mỗi cụm là một màu), trên hình có biểu diễn luôn kết quả của X_new.

5. Groceries dataset (1.5 điểm)

- 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 càng nhỏ, gợi ý: min_support = 0.001). 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.

6. Pen-Based Recognition of Handwritten Digits (1.5 điểm)

- 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

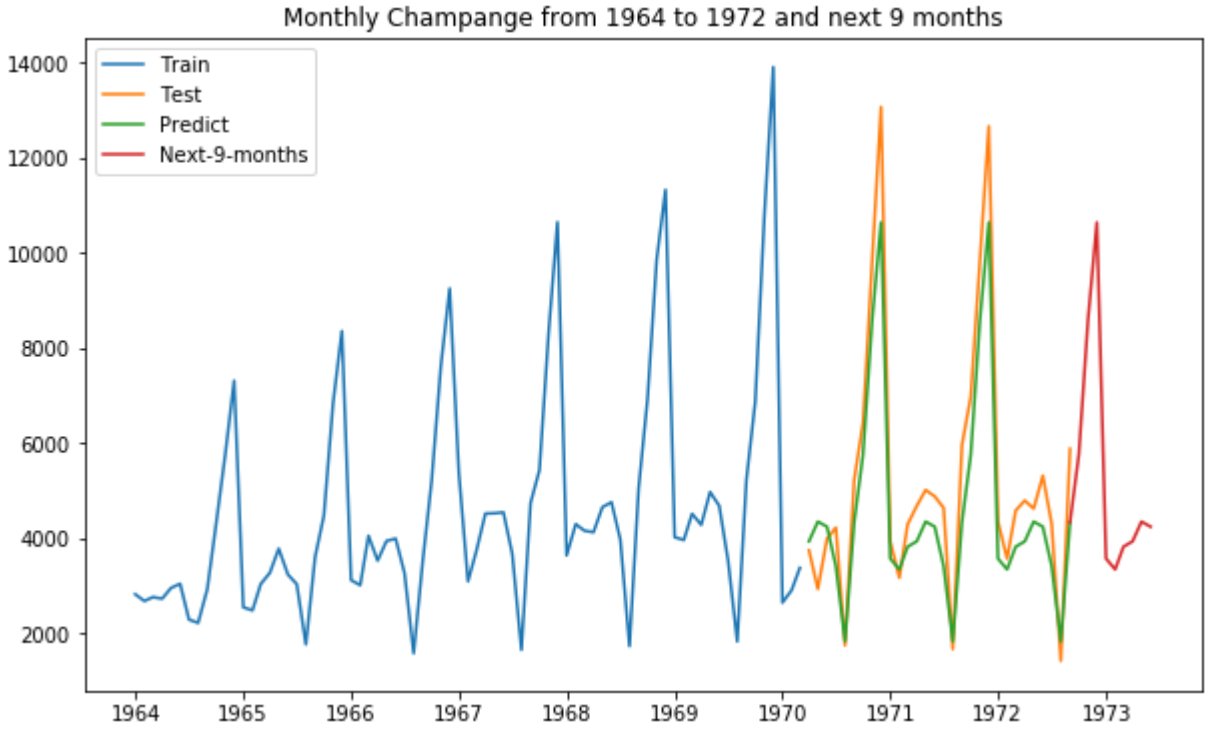
Attribute	Domain	Attribute	Domain
At1	[0, 100]	At9	[0, 100]
At2	[0, 100]	At10	[0, 100]
At3	[0, 100]	At11	[0, 100]
At4	[0, 100]	At12	[0, 100]
At5	[0, 100]	At13	[0, 100]
At6	[0, 100]	At14	[0, 100]
At7	[0, 100]	At15	[0, 100]
At8	[0, 100]	At16	[0, 100]
Class	{0,1,2,3,4,5,6,7,8,9}		

- 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 nào (trong các loại 0, 1, 2, 3, 4, 5, 6, 7, 8, 9) dựa trên các thông tin được cung cấp.
    1. Tạo X_train, X_test, y_train, y_test từ dữ liệu đọc và được chuẩn hóa, với tỷ lệ dữ liệu test là 0.3
    2. Áp dụng thuật toán thích hợp để xây dựng model. Dựa trên cơ sở nào để cho rằng thuật toán này thích hợp?
    3. Đánh giá model dựa trên train/test.
    4. 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 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**.

7. Monthly champagne sales millions (1.5 điểm)

- 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 ☺ ---