Đề thi:

MACHINE LEARNING WITH PYTHON

Thời hạn nộp bài: 23h59 ngày 12/08/2023

**** HV submit bài làm qua Google Classroom , mục bài thi cuối khóa hoặc gửi bài thi qua email eric.le2131@gmail.com (khuyến khích cách 1) ****

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

**** HV phải gửi bài làm đúng hạn nộp bài, sau hạn nộp bài nếu HV không gửi thì sẽ không được chấm điểm ****

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 được làm trong 1 file jupyter notebook riêng biệt. Trong trường hợp câu hỏi có yêu cầu giải thích/ trả lời bằng từ ngữ, học viên tạo 1 mark down cell để ghi lại câu trả lời.
- 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.

Text Classification with Naïve Bayes and NLP Techniques

Cho tập dữ liệu MBTI Kaggle, dùng để phân loại 16 tính cách (personalities). Dựa vào mô hình Naive Bayes, hãy thực hiện các yêu cầu sau :

1. Load tập data train từ file raw_train.csv và tập test từ file raw_test.csv
2. Tiến hành pre-process tập data train và test theo các phương pháp đã học (tokenizer, word lemmatizer…)
3. Xây dựng mô hình Naive Bayes dựa trên tập data đã preprocessed. Tiến hành dự đoán trên tập test.
4. Load label ground truth cho các mẫu trong tập test từ file solution.csv. Tính toán performance của model trên tập test. Nhận xét.

Supervised Learning: Classification Task, Parameter Search, Cross Validation and Boosting

Cho tập dữ liệu Pizza.csv, cột brand chứa tên các nhãn hang pizza (target), các cột còn lại chứa thông số của các loại pizza: mois, prot, fat, ash, sodium carb, cal và cột id chứa ID của các mẫu pizza. Tiến hành xây dựng mô hình phân loại bánh pizza dựa trên bộ data trên.

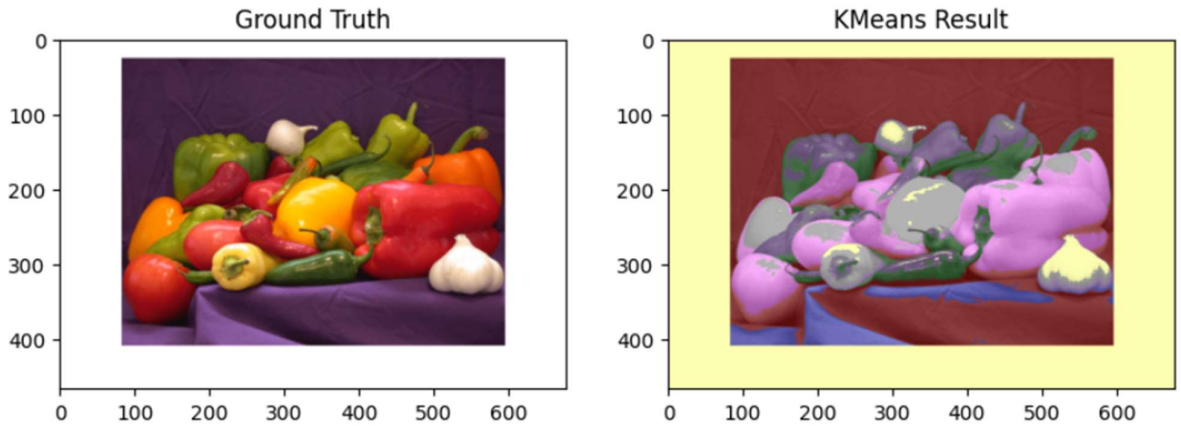
Yêu cầu:

1. Chia bộ dữ liệu ra thành 2 tập train/test với tỉ lệ 70:30
2. Dựa vào tập train, tiến hành tìm ra bộ thông số phù hợp cho các mô hình sau: SVM, kNN, Decision Tree, Random Forest và XGBoost, với cross validation = 5. Nhận xét các kết quả thu được, trực quan hóa và vẽ biểu đồ.
3. Chọn mô hình tốt nhất từ (2) để fit trên tập Train và dự đoán trên tập test. Report performance của model và nhận xét về performance của model.
4. Lựa chọn 1 trong các model Boosting đã học, fit trên tập train và dự đoán trên tập test. Nhận xét performance của model boosting vừa chọn và model từ (3) trên tập test.

Image Segmentation with Kmeans and GMM

Cho hình ảnh peppers.png, thực hiện các yêu cầu sau :

1. Thực hiện việc Segment hình ảnh trên bằng phương pháp KMeans với số lượng clusters thích hợp, trong không gian màu LAB.
2. Thực hiện việc Segment hình ảnh trên bằng phương pháp GMM với số lượng components thích hợp.
3. So sánh hiệu quả của 2 phương pháp.
   Lưu ý : Output của (1) và (2) nên là 1 hình ảnh tương tự như hình ví dụ bên dưới :



Association Rules Analysis

Học viên **chỉ chọn 1 trong 2 câu hỏi A HOẶC B**. Yêu cầu : Xử dữ liệu (nếu cần) và **dùng thuật toán** Association Rules Mining để thực hiện việc **tìm mức độ kết hợp giữa các items** dựa trên tập data trên và trả lời các câu hỏi.

A. Movie View Association

Cho tập dữ liệu trong file ratings.csv, gồm các thông tin sau:

userId	Chứa User ID của các thành viên
movieId	Chứa Id của các bộ phim được xem

Mỗi dòng trong dataset đại diện cho một bộ phim (movieId) đã được xem bởi một người dùng cụ thể (userId).

1. Áp dụng thuật toán **APRIORI**. In kết quả Item sets và giá trị support tương ứng, ví dụ như hình dưới:

userId	movieId
1	31
1	1029
1	1061

2. Chọn 1 metric thích hợp trong các metric sau: Confidence / Support / Lift / Conviction, sắp xếp các item set theo thứ tự giảm dần của metric vừa chọn. Cho biết top 5 item sets có giá trị metric vừa chọn cao nhất?

3. Cho biết movie có Id = 10 thường được xem chung với các movie nào?
4. Tìm top 10 movie được nhiều users xem nhất. Vẽ biểu đồ.

B. Basket Association Learning

Cho tập dữ liệu market.csv

1. Áp dụng thuật toán **ECLAT**. In kết quả Item sets và giá trị support tương ứng, ví dụ như hình dưới:

Item	Support
spaghetti & mineral water	0.059725
mineral water & chocolate	0.052660
mineral water & eggs	0.050927
milk & mineral water	0.047994

2. Cho biết item "tomatoes" thường được mua chung với các item nào?
3. Tìm top 5 item được mua nhiều nhất. Vẽ biểu đồ.
4. Theo bạn, ECLAT mặc định có hỗ trợ việc tạo ra các association rule có dạng X->Y hay không? Lợi ích của việc sử dụng association rules so với việc chỉ sử dụng frequent item sets là gì?

Time Series Forecasting

Cho tập dữ liệu daily_birth.csv gồm 2 cột chứa các thông tin sau:
Cột Date, chứa ngày-tháng-năm của từng mẫu dữ liệu
Cột Births: số ca sinh nở vào 1 ngày bất kỳ

Yêu cầu: Áp dụng thuật toán thích hợp để thực hiện các yêu cầu sau:

1. Trực quan hóa dữ liệu Time Series bằng 1 biểu đồ.
2. Thực hiện việc decomposition trên tập data, trực quan hóa kết quả bằng các biểu đồ, bạn có nhận xét gì về Trend, Seasonality?
3. Tiến hành chia bộ data như sau: tập train bao gồm từ mẫu ngày 01 tháng 01 năm 1959 đến ngày 31 tháng 7 năm 1959, tập test gồm các mẫu còn lại.
4. Xây dựng model ARIMA trên tập train, dự đoán trên tập test, vẽ biểu đồ so sánh prediction trên tập test với data thực tế.
5. Tiến hành dự đoán giá trị cho 1 tháng tiếp theo sau tập test (từ ngày 01 tháng 01 năm 1960 tới ngày 31 tháng 01 năm 1960). Vẽ biểu đồ trực quan kết quả dự đoán

Clustering

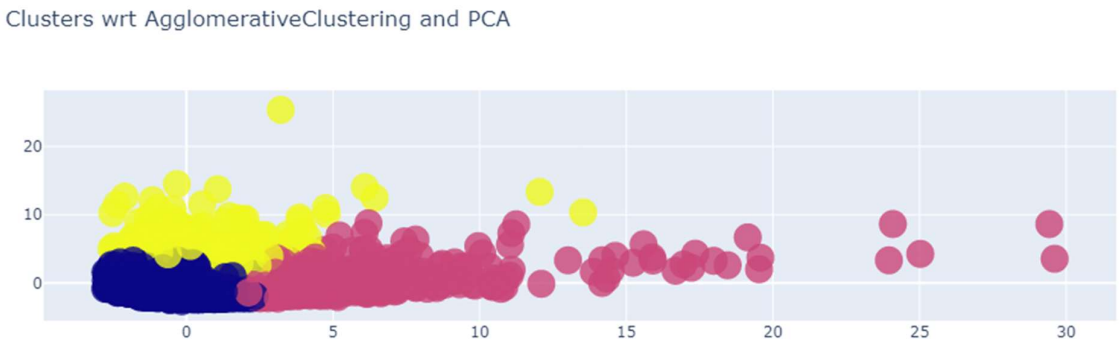
Cho tập dữ liệu card_data.csv gồm các cột sau :

country	Name of the country
child_mort	Death of children under 5 years of age per 1000 live births
exports	Exports of goods and services per capita. Given as %age of the GDP per capita
health	Total health spending per capita. Given as %age of GDP per capita
imports	Imports of goods and services per capita. Given as %age of the GDP per capita
Income	Net income per person
Inflation	The measurement of the annual growth rate of the Total GDP
life_expec	The average number of years a new born child would live if the current mortality patterns are to remain the same

total_fer	The number of children that would be born to each woman if the current age-fertility rates remain the same.
gdpp	The GDP per capita. Calculated as the Total GDP divided by the total population.

Yêu cầu :

1. Scale data với phương pháp thích hợp.
2. Xây dựng 3 mô hình clustering : KMeans, Agglomerative Clustering, GMM trên tập data trên. Lựa chọn số cluster phù hợp cho mỗi mô hình, kèm theo giải thích. Vẽ biểu đồ minh họa.
3. PCA với n_components = 2 trên tập data đã qua xử lý. Tiến hành chọn 1 trong 3 thuật toán ở câu 2, kèm số lượng cluster đã chọn ở câu 2 để fit trên tập data vừa được giảm chiều bởi PCA.
4. Visualize tập data đã qua xử lý PCA, vẽ 1 đồ thị 2D, trong đó các điểm cùng thuộc 1 cluster sẽ có cùng 1 màu (ví dụ như hình bên dưới). **Giải thích ý nghĩa của từng cluster dựa trên các thông tin thu được.**



--- Chúc các bạn làm bài tốt ---