Yêu cầu phần báo cáo cuối HK _ Machine learning – K 16

A. YÊU CẦU

Report in 2 mặt, nội dung gồm có:

- [Tất cả các nhóm] lời cam kết không đạo văn, ko đạo code
- [Dành cho những nhóm có presentation]
    - Trình bày các nội dung đã tìm hiểu được về việc tìm hiểu thư viện Scikit-learn cho các giải thuật cuối HK đã bốc thăm
- [Tất cả các nhóm] làm bài tập cho 6 giải thuật cuối HK
    - Trình bày cách xử lý data, code, cách làm , ….
    - Đánh giá độ hiệu quả của mô hình học máy
    - Đồ thị, plot, chart,…. nếu có
- 6 giải thuật cuối HK gồm có:
    - Multivariable Regression (MR)
        - Trình bày cách làm, tiền xử lý data nếu cần, ….
        - Các nhóm tự lựa chọn dùng independent variables nào để đưa vào giải thuật
        - Trình bày kết quả đánh giá
    - Polynomial Regression
        - (trình bày tương tự MR)
    - Logistic Regression
        - (trình bày tương tự MR, các nhóm tự chọn threshold)
    - PCA
        - Trước khi dùng PCA: chọn 1 giải thuật X (tự chọn) và thực hiện classification, đánh giá độ hiệu quả classification với X khi không dùng PCA
        - Dùng PCA:
            - Trình bày covariance matrix hoặc correlation matrix
            - Eigen values, eigen vectors đã được sắp xếp
            - Chọn số P (tự chọn)
            - P chiều dữ liệu mới chiếm bao nhiêu % lượng thông tin của dataset cũ
            - Chụp hình 1 phần ví dụ dataset mới sau khi dùng PCA
            - Chạy giải thuật X cho dataset mới này
            - Đánh giá độ hiệu quả classification với X khi dùng PCA
    - Kernel PCA
        - (làm tương tự như PCA, nhưng chạy thực nghiệm với những kernel khác nhau)
    - LDA
        - (làm tương tự PCA, cần trình bày thêm Within-class và Between-class scatter matrix)

Deadline nộp bài:

- 18h ngày 20/6/2019

- Địa chỉ: phòng 12.04 , tầng 12, Citilight tower, 45 Võ Thị Sáu, Q1, Tp. HCM
- Nộp sớm không cộng điểm
- Từ ngày 21/6 bị tính là nộp trễ
    - Mỗi ngày nộp trễ trừ 0.5 điểm cho toàn bộ nhóm
    - Nếu nộp sau ngày 25/6 thì bị 0 điểm vì lúc đó phòng ĐT đã khóa điểm

B. HIỂU VỀ DATASET

dataset_for_MultiLinear_regression.csv: dataset về các chỉ số con người, giải thích:

- Cột FAT_PER: cho biết mức độ mỡ (fat) của cơ thể, đây là biến phụ thuộc cần hồi quy
- Các cột khác: cho biết các số đo về con người (tuổi, cân nặng, chiều cao, …) là biến độc lập

dataset_for_Poly_regression.csv, giải thích:

- Rating: số sao rating cho 1 mobile app trên Google Play Store
- Reviews: số lần app được review
- Size: size của app tính theo KB hoặc MB
- Biến phụ thuộc cần hồi quy: "Installs" là số lần mà app được người dùng tải xuống và install
- Sinh viên preprocessing data nếu cần ;)

dataset_for_Logistic_regression.csv: dataset về dữ liệu hành khách đi tàu Titanic, giải thích:

- Survived: là label cần phân loại
- Các cột khác: là data features
- Gợi ý: vì các data features là dạng categorical variable nên cần tạo thêm các dummy variables ;)

dataset_for_PCA_LDA.csv: dataset về các chỉ số đo lường của 1 file source code của phần mềm

- Cột cuối cùng TEST_RESULT chính là label, có 2 label cho biết file source code là PASS hay FAIL 1 phép test (sinh viên ko cần quan tâm phép test này là gì)
- Tất cả những cột khác: là những thông số đo lường trên file source code, ví dụ :
    - LOC_BLANK: số lượng line of code để trống (xuống dòng)
    - LOC_COMMENTS: số lượng line of code dùng cho comment
    - LOC_TOTAL: tổng số lượng line of code của 1 file source code
    - ….
    - (sinh viên ko cần trình bày ý nghĩa của từng data features này)

dataset_for_Kernel_PCA.csv, giải thích:

- Cột cuối cùng "Purchased" chính là label, có 2 label cho biết khách hàng (UserID) sẽ mua hay không mua 1 món hàng
- Các cột khác (không tính UserID) là data features