Bài thực hành số 5

Lưu ý: Chủ đề trong diễn đàn nhóm tạo ra dùng để nộp các bài tập, các bài báo cáo, các kết quả thảo luận,.. trong suốt học kỳ. Các nhóm có thể tham khảo nội dung của các nhóm khác nhưng tuyệt đối không sao chép cho các bài kết quả của nhóm mình (đề cao tính trung thực).

Trong các buổi làm việc: yêu cầu làm việc theo nhóm, các nhóm có thể trao đổi, hỗ trợ nhau khi cần thiết. Nhóm có thể mang máy tính cá nhân cùng làm việc.

Câu 1:

Tập tin bank-data.csv trong LMS/Bộ dữ liệu thức hành chứa thông tin về các cá nhân vay tiền ngân hàng. Thông tin bao gồm: mã định danh, tuổi, giới tính, khu vực sinh sống, thu nhập (USD) 1 năm, tình trạng hôn nhân, số con, có xe hơi?, có mở tài khoản tiết kiệm?, có mở tài khoản thanh toán? Có nợ tiền mua nhà?, quyết định cho vay (Yes/No).

Sử dụng dữ liệu trong tập tin này để thực hiện các yêu cầu sau:

Xóa vùng định danh và xem như tập tin này được phân thành 2 lớp (nhóm): cho vay tiền (YES) và không cho vay (NO).

a) Sử dụng phần mềm/ứng dụng để khảo sát thuật toán phân lớp Knn. Khi trắc nghiệm mô hình (test), thử nhiều phương pháp khác nhau (test data=train data, % split, k-cross validation) và quan sát các kết quả (confusion matrix, TP rate/FP rate/accuracy/precision/..) . Cho biết mô hình mà anh chị xem là tốt nhất ứng với các tham số nào, các giá trị đánh giá,…

b) Từ tập tin ban đầu: tách thành 2 tập tin. Tập tin 1: TrainData sẽ dùng để huấn luyện (chiếm 90% dữ liệu) và tập tin 2:TestData (chiếm 10%) dữ liệu dùng để kiểm tra. Lưu ý khi phân chia cũng phải bảo đảm tỷ lệ từng lớp/nhóm trong 2 tập tin (nghĩa là tập tin 1 phải chứa 90% dữ liệu loại YES, 90% dữ liệu loại NO). Thực hiện lại câu a để tìm ra mô hình tốt nhất trong lúc huấn luyện.

c) Sau đó đưa tập tin 2:TestData vào để kiểm tra xem kết quả phân lớp có đúng hay không? Trình bày và nhận xét.

d) Thực hiện lại b và c với thuật toán Naive-Bayes.

Câu 2:

- Sử dụng lại tập tin kết quả thi THTHPT 2020 của TpHCM có kết quả xếp loại, nếu không có tập tin này thì thực hiện:
    - Bổ sung 2 cột: Điểm xếp loại và Xếp loại TN, sau đó tính tóan để điền thông tin cho 2 cột mới theo qui định dưới đây
      Điểm xếp loại = Tổng số điểm các môn thi /Tổng số môn thi.
    - Cách xếp loại tốt nghiệp THPT như sau:
      Tất cả bài thi và các môn thi thành phần của bài thi tổ hợp đăng ký dự thi để xét công nhận tốt nghiệp đều đạt trên 1 điểm, và có điểm xếp loại từ 5 điểm trở lên được công nhận tốt nghiệp THPT.
        + Loại giỏi: điểm xếp loại từ 8,0 điểm trở lên; không bài thi nào có điểm dưới 7,0.
        + Loại khá: điểm xếp loại từ 6,5 điểm trở lên; không bài thi nào bị điểm dưới 6,0.
        + Loại trung bình: các trường hợp còn lại.

      Các trường hợp còn lại: Không tốt nghiệp!!
- Xóa các thông tin cá nhân để bảo mật và tôn trọng quyền riêng tư. Chỉ giữ lại các vùng điểm Toán-Văn-Ngoại Ngữ, 3 môn thành phần trong khối thi, điểm xếp loại và Xếp loại TN.
- Tách thành 2 tập tin KetQuaTNTHPT-Train và KetQuaTNTHPT-Test với số lượng thí sinh trong 2 tập tin theo tỷ lệ 80(Train)/20(Test) .

- Sử dụng thuật toán KNN để xem kết quả xếp loại trong tập tin Test có đúng/tốt hay không?
- Sử dụng thuật toán Naive-Bayes để xem kết quả xếp loại trong tập tin Test có đúng/tốt hay không?

Kết quả câu 1 và câu 2 sẽ nộp trong buổi thực hành.

Câu 3:

Tập tin Collected_Hr_data_performances.xls (trên LMS) chứa các thông tin đánh giá năng lực làm việc của khoảng 1200 nhân viên dựa trên các thông tin về gia đình, học vấn, kinh nghiệm,…

Tách tập tin này thành 2 tập tin: 1 dùng để huấn luyện (khoảng 1000 nhân viên) và 1 dùng để kiểm tra kết quả (khỏang 200 nhân viên). Lưu ý có nhiều vùng trong tập tin chứa thông tin thuộc kiểu định danh.

Dùng thuật toán KNN để kiểm tra xếp loại các nhân viên trong tập tin Test có đúng/tốt hay không?

Dùng thuật toán Naïve Bayes để kiểm tra xếp loại các nhân viên trong tập tin Test có đúng/tốt hay không?

Câu 4: Phần thu hoạch

Phân công thực hiện, thực hiện ở nhà, thời hạn 1 tuần và gửi vào chủ đề trong diễn đàn.

- Báo cáo các kết quả của câu 3.
- Thực hiện câu 3 với chương trình Python (sữ dụng thư viện liên quan đến Machine Learning – sklearn) .

Lưu ý: Nhóm có thể dùng bất kỳ ứng dụng để thực hiện báo cáo (Power Point, Word, Clip/Video,…). Phần thu hoạch này cần nộp lên diễn đàn trước buổi học lý thuyết kế tiếp: GV có thể mời 1 nhóm lên trình bày.