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 với 2 phương pháp khác nhau (% split, k-cross validation với k=10) 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 TestData vào để kiểm tra xem kết quả phân lớp có đúng hay không? Trình bày và nhận xét.

Câu 2:

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,…Lưu ý là có 1 số vùng thông tin trùng tên nhau lý do người ta lưu lại thông tin ở dạng chi tiết cùng với dạng đã chia nhóm (ví dụ có 2 vùng Age – tuổi, 1 vùng Age lưu tuổi cụ thể của từng nhân viên, 1 vùng Age lưu tên nhóm tương ứng với tuổi cụ thể từng nhân viên)

Trước tiên anh chị:

- Chuyển tất cả các dòng có gán nhãn là PerformanceRating=1 và PerformanceResult=Does not Meet Minimum thành PerformanceRating=2 và PerformanceResult= Meet Expectation
- Chuyển tất cả các dòng có gán nhãn là PerformanceRating=5 và PerformanceResult=Outstanding thành PerformanceRating=4 và PerformanceResult= Exceed Expectation

Như vậy tập tin dữ liệu lúc này có 3 lớp (classes)

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 khi tách thì cũng cần tách theo tỷ lệ số phần tử từng lớp.

Dùng tập tin huấn luyện (1000 nhân viên) với thuật toán KNN để huấn luyện sau đó thử kiểm tra xếp loại các nhân viên trong tập tin kiểm tra (200 nhân viên) xem có đúng/tốt hay không?

Tương tự câu trên nhưng dùng thuật toán Naïve Bayes.