Bài thực hành số 6

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:
- Cho biết số lượng dữ liệu trong tập tin, các giá trị Mean/Median/Mod/Phương sai/Độ lệch chuẩn của các vùng tuổi và thu nhập.
- Vẻ 2 biểu đồ phân bổ tần số theo tuổi và thu nhập.
- Xóa vùng mã định danh để bảo đảm thông tin cá nhân.
- Cho biết Mod của các vùng tình trạng hôn nhân và số con.
- Lưu lại kết quả.
- Sao chép thành 1 tập tin mới đặt tên là Bank-Data-Clustering,csv

.

Câu 2:

- Mở tập tin Bank-Data-Clustering,csv và tách thành 2 tập tin Bank-Data-Clustering-Train,csv và Bank-Data-Clustering-Test,csv trong đó tập tin Bank-Data-Clustering-Test,csv chứa 50 dòng dữ liệu.
- Sử dụng tập tin Bank-Data-Clustering-Train,csv (loại bỏ vùng quyết định cho vay) để gom thừ thành 2 cụm, 3 cụm, 4 cụm.
- Sử dụng tập tin Bank-Data-Clustering-Test,csv (loại bỏ vùng quyết định cho vay) để kiểm tra lại kết quả với mỗi thực nghiệm gom 2 cụm, 3 cụm, 4 cụm.
- Nhận xét các kết quả.

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

Câu 3:

- Với tập tin ban đầu (ở câu 1) : Tách thành 2 tập tin Train (80%) và Test (20%)
- Dùng tập tin Train để xây dựng mô hình phân lớp dựa vào thuật toán KNN cho phép đưa ra quyết định cho vay/không cho vay (Lần lượt chọn K=3, 5, 7 và 9)
- Dùng tập tin Test để kiểm tra lại kết quả mô hình đã tìm ra (ứng với từng mô hình của mỗi giá trị K).
- Vẽ 1 đồ thị kết quả như sau: Trục hoành thể hiện các giá trị k tăng dần (3/5/7/9). Trục tung thể hiện giá trị sai số. Trên đồ thị này: vẽ đường cong biểu diễn sai số của giai đoạn huấn luyên (train) và đường cong biểu diễn sai số trong giai đoạn thử nghiệm (test). Dựa vào đồ thị hãy cho biết ta nên chọn k là bao nhiêu?

Câu 4:

- Dùng tập tin Train để xây dựng mô hình phân lớp dựa vào **vài thuật toán Cây QĐ** (Decision Tree, REP Tree, Random Forest) cho phép đưa ra quyết định cho vay/không cho vay (thử nghiệm lại với việc qui định chiều cao cây tối đa lần lượt là 4,5 và 6).
- Dùng tập tin Test để kiểm tra lại kết quả mô hình đã tìm ra.
- Thực nghiệm lại câu trên nhưng trước đó sử dụng chức năng Chọ thuộc tính (Select Features/Attributes) để chọn 7 thuộc tính quan trọng nhất.
- Nhận xét các thử nghiệm và kết quả.

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.