RÚ T TRÍ CH DỮ LIỆU VÀTI ỀN XỬ LÝ

1. Mục Tiêu

- Rút trích dữ liệu và tiền xử lý dữ liệu cho bài toán phân lớp

2. Bài tập thực hành

Một dự án máy học thông thường được chia thành 6 bước cơ bản sau:

1. Định nghĩa vấn đề
2. Phân tích dữ liệu
3. Chuẩn bị dữ liệu
4. Lượng giá thuật toán
5. Cải thiện kết quả
6. Trình bày kết quả

Trong đó, các bước cụ thể trong rút trích và tiền xử lý của một dự án bao gồm:

1. Định nghĩa vấn đề

- Phân tích yêu cầu cụ thể của bài toán là gì?
- Xác định đầu vào (Input): các tính chất (features) dùng để phân loại mẫu (pattern)
- Xác định đầu ra (Output): dữ liệu phân lớp (classification) hay hồi qui (regression)

2. Chuẩn bị vấn đề

- Khai báo thư viện
- Tải cơ sở dữ liệu

3. Phân tích dữ liệu

- Thống kê mô tả về dữ liệu:
  - Các thông tin cơ bản
  - Tính toàn vẹn dữ liệu (trùng lặp, giá trị thiếu)
  - Xác định đầu vào, đầu ra trong cơ sở dữ liệu
  - Các tính chất thống kê cơ bản (count, mean, min, max, 25%, 50% (median), 75%)
  - Phân bố dữ liệu phân lớp hay dữ liệu danh mục
  - Mối tương quan giữa các tính chất
- Hiển thị dữ liệu
  - Hiển thị dữ liệu cho các thuộc tính (features) đơn với dạng đồ thị box plot, histogram
  - Hiển thị tương quan của nhiều tính chất

4. Chuẩn bị dữ liệu

- Làm sạch dữ liệu
  - Tạo bảng dữ liệu làm sạch (loại bỏ các trường dữ liệu không cần thiết)
  - Xóa dữ liệu trùng nhau
  - Xử lý giá trị thiếu (loại bỏ feature, loại bỏ samples, hay điền giá trị hằng số hoặc tính toán như median)
- Chuyển đổi dữ liệu
  - Chuyển đổi dữ liệu danh mục sang số

  - Chuyển đổi dữ liệu danh mục sang dạng one-hot
- Chuẩn hóa dữ liệu: Min-Max, Standard
- Chia dữ liệu thực nghiệm

Bài 1. Thực hiện bước rút trích và tiền xử lý dữ liệu cho bài toán Phân lớp hoa Iris trong dữ liệu ***iris.csv***. Tóm tắt và trình bày kết quả đạt được dùng trình diễn.

Bài 2. Thực hiện bước rút trích và tiền xử lý dữ liệu cho bài toán Dự đoán bệnh tiểu đường trong dữ liệu ***pima-indians-diabetes.csv***. Tóm tắt và trình bày kết quả đạt được dùng trình diễn.

--- Hết ---