ĐẠI HỌC KHOA HỌC TỰ NHIÊN TP HỒ CHÍ MINH
KHOA CÔNG NGHỆ THÔNG TIN

So sánh mô hình học Support Vector Machine và Linear Regression trong gán nhãn đa lớp

Nguyễn Đình Thúc, Lương Việt Thắng

Ngày 22 tháng 5 năm 2016

1 MỤC TIÊU

- Cài đặt hoặc sử dụng thư viện có các thuật toán Linear Regression và Support Vector Machine
- Hiểu được các đặc điểm riêng của mô hình học Linear Regression và Support Vector Machine

2 NỘI DỤNG

Các em sẽ phải áp dụng SVM trên lớp để gán nhãn 1 tập dữ liệu được sinh ra bằng hàm sinh.

2.1 SINH DỮ LIỆU

Tập mẫu được sinh ra theo cách sau:

- Chọn 1 hàm 2 biến $(x, y) \in [0,1] : f(x, y) = ax + by + cxy$ (1) với a, b, c là các hằng số trong $[0, 1]$
- Giá trị của a, b, c bằng giá trị của 3 số cuối cùng khác 0 của mã số sinh viên của bạn chia cho 10.
- Ví dụ: MSSV: 1312223 thì $a = 2/10$, $b = 2/10$, $c = 3/10$. Nếu MSSV: 1312530 thì $a = 2/10$, $b = 5/10$, $c = 3/10$

- Sinh ra 200 mẫu $W = < x, y, label >_i$ với $i = 1, \cdots, 200$ trong đó $label_i = f(x, y) mod 3$ hay $label_i =$ (int) f(x+y)%3 trong C++ và x, y được sinh ra ngẫu nhiên theo điều kiện $x, y \in [0, 1]$ và $f(x, y)$ được tính theo (1).

2.2 YÊU CẦU

Các em thực hiện các yêu cầu sau:

- Biểu diễn dữ liệu: cung cấp các thông tin thống kê về dữ liệu như trung bình, độ lệch chuẩn, vẽ phân phối của dữ liệu, khoảng $[minmax]$
- Cài đặt mô hình Linear Regression cho tập dữ liệu được sinh
- Cài đặt mô hình SVM cho tập dữ liệu được sinh
- Huấn luyện và kiểm tra 2 mô hình học trên theo phương pháp 10-fold cross validation.
- So sánh và phân thích kết quả của 2 mô hình học, từ đó nêu ra ưu nhược điểm của từng thuật toán đối với dữ liệu này.
- Viết báo cáo giải thích những điểm quan trọng trong mã nguồn.

2.3 LƯU Ý

Các em được phép tham khảo các tài liệu trên mạng và sử dụng thư viện có sẵn nhưng phải nắm vững được cách làm bài tập vì sẽ có chấm vấn đáp nội dung thực hành trong kiểm tra lý thuyết cuối kỳ.

3 NỘP BÀI

Các em nộp bài lên hệ thống Moodle gồm các thư mục sau:

- Báo cáo
- Mã nguồn

Nén lại với tên file là mã số sinh viên.