Đồ án thực hành cuối kỳ

Bài 1. (5.0đ)

Sử dụng dữ liệu thực tế để xây dựng mô hình dự đoán giá nhà tại Stockton, California. Sinh viên cần khám phá dữ liệu và tự đề xuất mô hình hồi quy phù hợp (đơn biến hoặc đa biến, tuyến tính hoặc log biến đổi) nhằm đạt được hiệu quả dự đoán tối ưu. Đồ án khuyến khích tư duy phản biện, thử nghiệm nhiều hướng tiếp cận và đánh giá mô hình dựa trên các tiêu chí phù hợp.

Tệp stockton.csv chứa thông tin về các giao dịch bán nhà ở Stockton, CA giữa năm 2005. Các đặc trưng bao gồm:

sprice: giá bán nhà, tính bằng đô la

livarea: diện tích ở, tính bằng trăm feet vuông (1 trăm feet vuông ≈ 9.29 mét vuông)

beds: số phòng ngủ

baths: số phòng tắm

lgelot: =1 nếu diện tích lô đất > 0.5 mẫu Anh (acre), 0 nếu ngược lại

age: tuổi của ngôi nhà tại thời điểm bán, tính bằng năm

pool: =1 nếu nhà có hồ bơi, 0 nếu ngược lại

a) Khám phá dữ liệu: Đọc dữ liệu và phân tích tương quan giữa các đặc trưng với sprice qua hình vẽ phù hợp.

b) Xây dựng mô hình đơn đặc trưng: sử dụng sprice làm đặc trưng phụ thuộc và lần lượt các đặc trưng còn lại làm đặc trưng độc lập qua mô hình: .

Đánh giá mô hình theo chỉ số độ lớn vector phần dư.

Vẽ đồ thị so sánh giá trị thực tế và giá trị dự đoán.

c) Xây dựng mô hình đa đặc trưng: sử dụng giá nhà làm đặc trưng phụ thuộc và các đặc trưng còn lại làm đặc trưng độc lập qua mô hình: $\text{sprice} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots.$

Đánh giá mô hình theo chỉ số độ lớn vector phần dư. Vẽ đồ thị so sánh giá trị thực tế và giá trị dự đoán.

d) Biến đổi các đặc trưng để có thể tìm một mô hình có kết quả tốt hơn qua các mô hình đã xây dựng ở câu b, c nhằm dự đoán sprice bằng cách lấy log, bình phương, …

Bài 2. (5.0đ)

Xích Markov là một mô hình toán học dùng để mô tả các quá trình ngẫu nhiên mà trạng thái hiện tại quyết định hoàn toàn đến trạng thái kế tiếp, mà không phụ thuộc vào lịch sử

các trạng thái trước đó. Tập hợp trạng thái là hữu hạn và quá trình chuyển đổi giữa các trạng thái được mô tả bằng ma trận xác suất chuyển trạng thái.

Chúng ta cần dự đoán tình trạng giao thông ngày hôm sau dựa vào lịch sử giao thông của một tuyến đường chính trong thành phố, ví dụ đường Nguyễn Văn Cừ. Dữ liệu sẽ gồm chuỗi trạng thái giao thông được ghi nhận trong vòng nhiều ngày, phân loại thành:

K – Kẹt xe, B – Bình thường, T – Thoáng

Mục tiêu: dựa trên chuỗi trạng thái đã có trong quá khứ, xây dựng ma trận chuyển trạng thái, sử dụng ma trận đó để dự đoán xác suất xảy ra của từng trạng thái cho ngày tiếp theo.

a) Tạo dữ liệu mô phỏng: Sinh viên tạo chuỗi dữ liệu giả lập mô tả tình trạng giao thông trong 60 ngày, lưu trữ dưới dạng danh sách các giá trị dạng ký hiệu: "K", "B", "T" với khả năng xảy ra các tình trạng là như nhau ở mỗi ngày.

b) Xây dựng ma trận đếm số lần chuyển trạng thái: Từ dữ liệu, sinh viên cần đếm số lần chuyển từ mỗi trạng thái này sang trạng thái khác (VD: từ K → K, K → B, ...).

c) Tính toán ma trận xác suất chuyển: mỗi dòng trong ma trận được chuẩn hóa bằng cách chia số lần chuyển sang từng trạng thái cho tổng số lần xuất phát từ trạng thái đó.

d) Dự đoán trạng thái các ngày tiếp theo: sinh viên viết chương trình với tên *prop_distribution* để tính xác suất xảy ra các phân loại giao thông tại **ngày thứ k** sau ngày hôm nay, giả định rằng ngày hôm nay khả năng của các trạng thái là như nhau.

Kiểm tra lại chương trình đã viết bằng cách tính xác suất sau 1, 2, 3, 10 ngày.

e) Tìm phân phối dừng của xích markov: sinh viên viết chương trình tìm phân phối dừng với tên *prop_sta* của xích markov đã xây dựng và xuất kết quả.

Kiểm tra lại chương trình đã viết bằng cách xuất phân phối dừng đã mô phỏng ở trên.

YÊU CẦU

- Thực hiện toàn bộ bài làm trên 1 tập tin Jupyter Notebook (. ipynb).
- SV nộp tập tin gồm. ipynb và .pdf lên Moodle, không nén 2 file.
- Đặt tên file: MSSV-hoten(không dấu, chữ thường). VD: 23202122-tranvanan

Ở phần đầu của file cần có phần giới thiệu thông tin cá nhân: Họ tên, MSSV, lớp. Ghi chú ý nghĩa của từng hàm (mỗi hàm 1 lần, tại nơi đầu tiên xuất hiện hàm).

QUY ĐỊNH: SV bị 0 điểm trong các trường hợp sau:

- Nộp sai qui định. Thực thi mã nguồn báo lỗi.
- Chép bài của thí sinh khác hoặc cho thí sinh khác chép bài.
- Và các hình thức kỷ luật khác theo qui chế thi.

Sinh viên chỉ sử dụng các thư viện hỗ trợ cho việc đọc dữ liệu, vẽ hình, biểu diễn ma trận, phát sinh dữ liệu mô phỏng.