

CÁC PHƯƠNG PHÁP ĐỊNH LƯỢNG 2**Bài thi giữa kỳ**

Ngày phát: 31/03/2022

Hạn nộp: 8h20, 02/04/2022

Học viên được yêu cầu chỉ nộp bản điện tử trên Microsoft Teams

Dùng bộ dữ liệu Boston.csv và lấy cột dữ liệu ‘crim’ làm biến phụ thuộc và các cột dữ liệu khác làm biến độc lập, Anh/Chị thực hiện các yêu cầu sau.

Câu 1: Anh/Chị xác định những biến độc lập có ảnh hưởng lớn nhất đến biến phụ thuộc. Cụ thể, Anh/Chị hãy:

- a) Chuẩn hóa các biến độc lập;
- b) Xây dựng 1 mô hình hồi quy OLS của biến phụ thuộc với tất cả các biến độc lập đã được chuẩn hóa (sử dụng toàn bộ các hàng trong bộ dữ liệu);
- c) Xác định những biến độc lập có ảnh hưởng lớn nhất đến biến phụ thuộc;
- d) Giải thích vì sao việc chuẩn hóa các biến độc lập có thể giúp trả lời câu hỏi trên.

Câu 2: Anh/Chị chia ngẫu nhiên bộ dữ liệu thành 2 phần theo tỷ lệ 70:30 và thực hiện các yêu cầu sau:

- a) Xây dựng mô hình hồi quy Lasso (sử dụng phần dữ liệu 70%) với các giá trị lambda khác nhau và tính giá trị test MSE tương ứng với mỗi giá trị lambda này (dùng phần dữ liệu 30%);
- b) Vẽ đồ thị biểu diễn sự biến thiên của các hệ số hồi quy của mô hình Lasso theo các giá trị của lambda;
- c) Vẽ đồ thị biểu diễn sự biến thiên của giá trị test MSE theo các giá trị lambda;
- d) Xác định giá trị lambda tương ứng với giá trị test MSE nhỏ nhất;
- e) Dùng giá trị lambda trên để xây dựng một mô hình hồi quy Lasso mới sử dụng toàn bộ các hàng trong bộ dữ liệu. Trình bày các hệ số hồi quy của mô hình Lasso này.

Câu 3: Anh/Chị xây dựng mô hình hồi quy Lasso của biến ‘crim’ với các biến khác. Tuy nhiên, thay vì xác định giá trị của lambda theo phương pháp validation set như trong Câu 2, Anh/Chị xác định giá trị của lambda bằng phương pháp 10-fold cross validation. Dùng giá trị lambda này để xây dựng một mô hình hồi quy Lasso mới sử dụng toàn bộ các hàng trong bộ dữ liệu.

Anh/Chị có nhận xét gì về cách làm của 2 phương pháp ở câu 2 và câu 3?

Anh/Chị có thể dùng phần mềm và/hoặc công cụ tính toán phù hợp nhất với bản thân.

Anh/Chị cần trình bày chi tiết từng bước của bài làm cũng như các giải thích nhận định trong một báo cáo (ví dụ bằng file Word). Việc chấm điểm sẽ chỉ dựa trên vào các thông tin được trình bày trong báo cáo.

Anh/Chị vẫn phải nộp đầy đủ các file tính toán (Excel hoặc các đoạn code) nhưng chúng chỉ có giá trị tham khảo bổ sung trong quá trình chấm bài.

---Hết---