MACHINE LEARNING: BÀI THỰC HÀNH SỐ 8 – PHẦN 2

MÔ HÌNH MULTI LAYER PERCEPTRON CHO BÀI TOÁN REGRESSION

Trong phần thực hành này, chúng ta sử dụng lại mô hình ANN đã xây dựng trong phần 1, tuy nhiên cải tiến để áp dụng cho bài toán hồi quy. Để đơn giản, chúng ta giả thiết áp dụng cho bài toán hồi quy đơn (đầu ra là dạng numeric $y \in \mathbf{R}$).

Lúc đó có một số điểm chúng ta cần sửa lại:

1) Hàm loss (cost): Sử dụng loss kiểu MSE (hoặc tổng bình phương sai số - tham khảo lại phần Hồi quy tuyến tính đã có code)
2) Số chiều (class) đầu ra C = 1 (chỉ trả về 01 giá trị số)
3) Hàm kích hoạt $\hat{y} = f(z) = SoftMax(z)$ ở đầu ra cần đổi thành $\hat{y} = f(z) = z$. Do đó $f'(z) = 1$ thay cho $f'(z) = (\hat{y} - y)$ như hiện tại.

Cập nhật lại những điểm trên, sau đó thực nghiệm mô hình để dự đoán cho các bài toán sau.

Bài tập 1.

Lấy lại dữ liệu trong tệp SAT_GPA.csv (đính kèm Ví dụ A.1 – Phần Hồi quy tuyến tính – Linear Regression). Nhắc lại tệp có 84 mẫu dữ liệu điểm thi của các sinh viên, mẫu có 02 trường dữ liệu, trong cột thứ nhất chứa trường điểm SAT (Reading + Mathematic + Writing) của các kỳ thi trong bậc phổ thông; cột thứ hai chứa điểm trung bình GPA của sinh viên tương ứng ở bậc học đại học/cao đẳng.

Sử dụng mô hình ANN đã sửa đổi, và từ dữ liệu lấy 54 mẫu bất kỳ đưa vào tập training – 30 mẫu đưa vào tập validation.

- Hãy huấn luyện và dự đoán đầu ra của dữ liệu trên cả hai tập, đánh giá độ chính xác của mô hình thông qua các độ đo R^2 (R – squared) và MSE.
- Sử dụng mô hình hồi quy tuyến tính để thực hiện lại dự đoán. So sánh với mô hình ANN trên các tiêu chí: Thời gian training; Thời gian predict (tính trung bình); độ chính xác.
- Thay đổi số chiều layer ẩn lần lượt là 75, 50. Thực nghiệm lại và đánh giá sự thay đổi kết quả so với các thông số trong đoạn code đã cho. Hãy cho nhận xét về mối liên hệ giữa siêu tham số trên với kết quả dự đoán.

Bài tập 2.

Sử dụng lại dữ liệu Ví dụ B4, phần Hồi quy Tuyến tính về tính toán bề dày lớp nội trung mạc (NTM) – thuộc tính phản ánh một số bệnh lý của cơ thể. Nhắc lại rằng trong thực tế hiện tượng dày lớp NTM động mạch cảnh do nhiều yếu tố như di truyền, chủng tộc, mắc bệnh tim mạch, tuổi, giới, BMI, tăng huyết áp, đái tháo đường.... cùng tác động. Trong ví dụ này ta không đề cập các yếu tố di truyền, chủng tộc, giới, mắc bệnh tim

mạch... mà chỉ lưu ý đến các biến số như: tuổi, cholesterol, glucose, huyết áp tâm thu và BMI tác động lên độ dày NTM.

Hãy áp dụng mô hình ANN đã học và dữ liệu cho trong tệp vidu4_lin_reg.txt (tệp văn bản) để dự đoán bề dày lớp NTM theo các biến số khác. Tham khảo phần đọc dữ liệu từ tệp văn bản đã có trong ví dụ trước. Các trường dữ liệu gồm:

ID	Mã bệnh nhân
TUOI	Tuổi
BIM	chỉ số khối lượng cơ thể (Body Mass Index)
HA	huyến áp tâm thu
GLUCOSE	đường huyết
CHOLESTEROL	độ Cholesterol trong máu
BEDAYNTM (đầu ra)	độ dày NTM

Chia dữ liệu thành: 80 dòng đầu dùng cho training; 20 dòng sau dùng cho testing. Huấn luyện mô hình ANN cải biên với phần training và thực hiện dự đoán trên cả 2 tập. Cho biết kết quả và đánh giá độ chính xác với các độ đo như ở bài tập 1.

So sánh kết quả này với kết quả phương pháp hồi quy tuyến tính đã học.

Bài tập 3. Trong tệp dữ liệu Real_estate.csv đính kèm chứa thông tin các giao dịch mua bán bất động sản. Chúng ta có 414 mẫu dữ liệu, mỗi bản ghi có 8 cột theo thứ tự là

- Cột x1: Số thứ tự (chúng ta sẽ bỏ qua trường này)
- Cột x2: Ngày giao dịch mua bán (ta chỉ lấy phần nguyên là năm)
- Cột x3: Tuổi của căn nhà (theo năm)
- Cột x4: Khoảng cách tới ga MRT (phương tiện công cộng nội đô) gần nhất
- Cột x5: Số cửa hàng tiện ích gần đó
- Cột x6: Kinh độ căn nhà; Cột X7: Vĩ độ căn nhà;
- Cột Y (đầu ra dự báo): Giá của căn nhà

Hãy chia dữ liệu thành phần training với 350 mẫu đầu tiên, phần validation với số mẫu còn lại. Hãy tham khảo các bài phần hồi quy tuyến tính và sử dụng mô hình ANN đã có để dự đoán Y đầu ra theo các cột từ X2 đến X6. Sau đó hãy chạy dự đoán cho phần dữ liệu validation và đưa ra tổng bình phương sai số của dự đoán. So sánh với phương pháp hồi quy tuyến tính.