Dự Án Cuối Kỳ

Môn: Nhập môn Học máy

Hãy phân tích bài toán đưa ra một giải pháp dựa vào học máy để dự đoán giá trị trong tương lai của một trong loại data sau:

(A) Dự đoán giá các cổ phiếu
(B) Dự đoán giá các hàng hoá
(C) Dự đoán tỉ lệ tăng trưởng GDP của các nước

Đối với giá hàng hoá và cổ phiếu trong tương lai thì dự đoán các mốc thời gian sau:

- Dự đoán giá cổ phiếu theo từng ngày trong 14 ngày tiếp theo (theo đồ thị ngày)
- Dự đoán giá cổ phiếu theo từng tuần trong tuần trong 7 tuần tiếp theo (theo đồ thị tuần)

Đối bài toán dự đoán GDP thì dự đoán các mốc sau:

- Dự đoán 4 quý tiếp theo
- Dự đoán 4 năm tiếp theo

Các nhiệm vụ (câu hỏi) cần làm:

(1) Thu thập dữ liệu các thông tin liên quan của cổ phiếu (hoặc hàng hoá, hoặc GDP) theo thời gian

(2) Phân tích bài toán để xác định thông tin nào được sử dụng làm đầu vào để dự đoán giá tương lai.
(Ví dụ đối với giá cổ phiếu, có thể phụ thuộc vào: các loại giá quá khứ, thông tin ngành nghề, thông tin tài chính của công ty, …)
Lưu ý:
   - Phần 2 và phần 1 liên quan đến nhau, khi phần 2 xác định sử dụng loại thông tin nào thì phần 1 thu thập theo các loại thông tin đó.
   - Nhóm tự quyết định việc sử dụng nhiều hay ít loại thông tin là quyết định riêng của từng nhóm. Về nguyên tắc thì càng nhiều thông tin thì khả năng dự đoán càng chính xác.

(3) Xây dựng các mô hình học máy khác nhau để dự đoán trong đó có 2 thuật toán bắt buộc là: Reccurent Neural Network (RNN) và MultiLayer Perceptron (MLP).
Khuyến khích sử dụng thêm các thuật toán khác: SVM, kNN, DT, Random Forest, …
Đánh giá độ tốt (độ chính xác) của các mô hình và so sánh.
Đánh giá thời gian huấn luyện và thời gian Testing.
Vẽ đồ thị bao gồm giá trị thực và giá trị dự đoán để dễ so sánh.

(4) Sử dụng các phương pháp xử lý vấn đề Overfiting đối với câu hỏi (3) và so sánh với việc không sử dụng Overfiting. Áp dụng trên các models ở câu hỏi (3)

(5) Sử dụng các tập đặc trưng khác nhau (liên quan đến câu hỏi (2)) để tìm ra tập đặc trưng hiệu quả nhất cho bài toán. Tìm đặc trưng quan trọng nhất ảnh hưởng đến sự giá trị của cái phải dự đoán (tức là feature nào ảnh hưởng nhất đến output).

(6) Nghiên cứu áp dụng các mô hình Deep Learning (tự tìm hiểu) để giải quyết bài toán. Trình bày cả phần lý thuyết của mô hình mới này. Ví dụ mô hình LSTM, mô hình CNN, các mô hình kết hợp các phương pháp khác nhau.

Hướng dẫn nộp bài:

1) Nộp file báo cáo, file pdf chứa thông tin thành viên trong nhóm; chứa thông tin trả lời cho 6 câu hỏi trên. Lưu ý tham khảo trên các trang nào thì liệt kê ở cuối tài liệu.
2) Nộp file code python dạng jupyter, chỉ 1 file duy nhất cho toàn bộ các câu hỏi. Chia thành các section khác nhau và đã chạy cho trước kết quả. File data đi kèm.
3) Bản backup: lưu nội dung (1) và (2) ở trên vào Google Driver và nộp đường link

Tất cả các thành viên trong nhóm đều phải nộp giống nhau.

Lưu ý:

- **Không extend thời gian làm bài.**
- **Các nhóm copy của nhau thì sẽ bị 0 điểm.**
- **Báo cáo tiến độ hàng tuần: tuần 1, tuần 2, tuần 3 (nộp bài)**