Bài tập 1: Phân loại hoa Iris

Mô tả: Sử dụng tập dữ liệu Iris (có sẵn trong thư viện scikit-learn) để xây dựng một mô hình phân loại nhằm dự đoán loại hoa Iris dựa trên các đặc trưng của nó.

Yêu cầu:

1. Tải tập dữ liệu Iris và chia thành hai phần: tập huấn luyện (80%) và tập kiểm tra (20%).
2. Tiền xử lý dữ liệu nếu cần thiết.
3. Áp dụng thuật toán K-Nearest Neighbors (KNN) để xây dựng mô hình. Thử nghiệm với các giá trị khác nhau của tham số k.
4. Đánh giá mô hình trên tập kiểm tra và báo cáo độ chính xác.
5. Vẽ biểu đồ so sánh độ chính xác của mô hình với các giá trị khác nhau của k.
6. Giải thích về sự lựa chọn giá trị k tốt nhất.

Điểm số: 10 điểm

Bài tập 2: Dự đoán giá nhà ở Boston

Mô tả: Sử dụng tập dữ liệu giá nhà ở Boston để xây dựng một mô hình hồi quy nhằm dự đoán giá trị trung bình của các căn nhà.

Yêu cầu:

1. Tải tập dữ liệu giá nhà ở Boston (có sẵn trong thư viện scikit-learn) và chia thành hai phần: tập huấn luyện (80%) và tập kiểm tra (20%).
2. Thực hiện các bước tiền xử lý cần thiết, bao gồm việc chuẩn hóa dữ liệu.
3. Áp dụng thuật toán hồi quy tuyến tính (Linear Regression) để xây dựng mô hình dự đoán.
4. Đánh giá mô hình dựa trên các chỉ số: Mean Absolute Error (MAE), Mean Squared Error (MSE), và R-squared (R^2).
5. So sánh với mô hình Decision Tree Regression.
6. Viết nhận xét về hiệu quả của hai mô hình và đề xuất cải tiến.

Điểm số: 10 điểm

Bài tập 3: Phân loại bệnh tiểu đường

Mô tả: Sử dụng tập dữ liệu tiểu đường để xây dựng mô hình phân loại nhằm dự đoán nguy cơ mắc bệnh tiểu đường dựa trên các chỉ số sức khỏe.

Yêu cầu:

1. Tải tập dữ liệu và chia thành tập huấn luyện (80%) và tập kiểm tra (20%).
2. Tiền xử lý dữ liệu, bao gồm chuẩn hóa các đặc trưng.
3. Áp dụng thuật toán Logistic Regression để xây dựng mô hình phân loại.
4. Đánh giá mô hình bằng các chỉ số: độ chính xác, độ nhạy, độ chính xác, và F1-score.
5. Thử nghiệm với thuật toán SVM và so sánh kết quả.
6. Giải thích sự khác biệt giữa các thuật toán và đưa ra lựa chọn tối ưu.

Điểm số: 10 điểm

Bài tập 4: Phân loại thư rác

Mô tả: Sử dụng tập dữ liệu email để xây dựng một mô hình phân loại nhằm xác định xem một email là thư rác hay không.

Yêu cầu:

1. Tải tập dữ liệu và chia thành tập huấn luyện (80%) và tập kiểm tra (20%).
2. Thực hiện các bước tiền xử lý văn bản (tokenization, stop words removal).
3. Biến đổi văn bản thành dạng số bằng phương pháp TF-IDF.
4. Áp dụng thuật toán Naive Bayes để phân loại.
5. Đánh giá độ chính xác và vẽ biểu đồ ma trận nhầm lẫn.
6. Đề xuất cải tiến mô hình dựa trên kết quả.

Điểm số: 10 điểm

Bài tập 5: Dự đoán lượng mưa

Mô tả: Sử dụng tập dữ liệu thời tiết để xây dựng mô hình hồi quy nhằm dự đoán lượng mưa dựa trên các yếu tố khí hậu.

Yêu cầu:

1. Tải tập dữ liệu và chia thành tập huấn luyện (80%) và tập kiểm tra (20%).
2. Tiền xử lý dữ liệu, bao gồm xử lý giá trị thiếu và chuẩn hóa.
3. Áp dụng thuật toán Random Forest Regression để xây dựng mô hình.
4. Đánh giá mô hình dựa trên các chỉ số: MAE, MSE, và R-squared.
5. So sánh với mô hình hồi quy tuyến tính.
6. Nhận xét về mô hình tốt nhất và giải thích lý do.

Điểm số: 10 điểm

Bài tập 6: Phân loại chữ số viết tay (MNIST)

Mô tả: Sử dụng tập dữ liệu MNIST để xây dựng một mô hình phân loại nhằm dự đoán các chữ số từ 0 đến 9.

Yêu cầu:

1. Tải tập dữ liệu MNIST và chia thành tập huấn luyện (80%) và tập kiểm tra (20%).
2. Tiền xử lý dữ liệu, chuẩn hóa hình ảnh.
3. Áp dụng thuật toán SVM để xây dựng mô hình phân loại.
4. Đánh giá mô hình bằng độ chính xác và vẽ ma trận nhầm lẫn.
5. Thử nghiệm với một thuật toán khác (ví dụ: KNN) và so sánh kết quả.
6. Giải thích lý do chọn mô hình cuối cùng.

Điểm số: 10 điểm

Bài tập 7: Dự đoán giá cổ phiếu

Mô tả: Sử dụng tập dữ liệu giá cổ phiếu để dự đoán giá cổ phiếu của một công ty dựa trên các dữ liệu lịch sử.

Yêu cầu:

1. Tải tập dữ liệu và chia thành tập huấn luyện (80%) và tập kiểm tra (20%).
2. Tiền xử lý dữ liệu (điều chỉnh giá trị thiếu, chuẩn hóa).
3. Áp dụng mô hình hồi quy tuyến tính để dự đoán giá cổ phiếu.
4. Đánh giá mô hình qua các chỉ số: MAE, MSE, và R-squared.
5. So sánh với mô hình LSTM.
6. Đưa ra kết luận về hiệu quả dự đoán.

Điểm số: 10 điểm

Bài tập 8: Dự đoán điểm số sinh viên

Mô tả: Dựa trên các đặc trưng như số giờ học và mức độ tham gia các hoạt động, dự đoán điểm số của sinh viên.

Yêu cầu:

1. Tải và chia tập dữ liệu thành tập huấn luyện (80%) và tập kiểm tra (20%).
2. Tiền xử lý dữ liệu, bao gồm mã hóa các biến phân loại nếu cần.
3. Áp dụng mô hình hồi quy tuyến tính để dự đoán điểm số.

4. Đánh giá mô hình bằng các chỉ số: MAE, MSE, và R-squared.
5. So sánh với mô hình Decision Tree Regression.
6. Viết nhận xét về kết quả và đề xuất cải thiện.

Điểm số: 10 điểm

Bài tập 9: Dự đoán sự hài lòng của khách hàng

Mô tả: Sử dụng tập dữ liệu khảo sát khách hàng để dự đoán mức độ hài lòng dựa trên các đặc điểm về dịch vụ.

Yêu cầu:

1. Tải tập dữ liệu và chia thành tập huấn luyện (80%) và tập kiểm tra (20%).
2. Tiền xử lý dữ liệu và mã hóa các biến phân loại.
3. Áp dụng thuật toán phân loại Random Forest để xây dựng mô hình.
4. Đánh giá hiệu quả của mô hình với độ chính xác và F1-score.
5. So sánh với thuật toán SVM.
6. Giải thích sự khác biệt và lý do chọn mô hình tối ưu.

Điểm số: 10 điểm

Bài tập 10: Phân tích cảm xúc trên đánh giá sản phẩm

Mô tả: Sử dụng tập dữ liệu đánh giá sản phẩm để phân loại cảm xúc thành tích cực và tiêu cực.

Yêu cầu:

1. Tải tập dữ liệu và chia thành tập huấn luyện (80%) và tập kiểm tra (20%).
2. Thực hiện tiền xử lý văn bản (xóa dấu câu, mã hóa văn bản).
3. Sử dụng phương pháp TF-IDF để chuyển đổi văn bản thành dạng số.
4. Áp dụng thuật toán Naive Bayes để xây dựng mô hình phân loại.
5. Đánh giá mô hình bằng độ chính xác và F1-score.
6. Đề xuất phương pháp cải thiện dự đoán.

Điểm số: 10 điểm

Bài tập 11: Dự đoán bệnh tim

Mô tả: Sử dụng dữ liệu bệnh lý để dự đoán nguy cơ mắc bệnh tim.

Yêu cầu:

1. Tải tập dữ liệu và chia thành tập huấn luyện (80%) và tập kiểm tra (20%).
2. Tiền xử lý các giá trị thiếu và chuẩn hóa dữ liệu.
3. Áp dụng thuật toán Logistic Regression để xây dựng mô hình phân loại.
4. Đánh giá mô hình bằng độ chính xác và F1-score.
5. Thử nghiệm với thuật toán khác (ví dụ: Random Forest).
6. Giải thích sự khác biệt giữa các thuật toán.

Điểm số: 10 điểm

Bài tập 12: Dự đoán số lượt đặt phòng khách sạn

Mô tả: Sử dụng dữ liệu lịch sử đặt phòng để dự đoán số lượt đặt phòng của khách sạn.

Yêu cầu:

1. Tải tập dữ liệu và chia thành tập huấn luyện (80%) và tập kiểm tra (20%).
2. Tiền xử lý và mã hóa các biến phân loại.
3. Áp dụng thuật toán hồi quy tuyến tính để dự đoán.
4. Đánh giá mô hình bằng các chỉ số: MAE, MSE, và R-squared.
5. So sánh với mô hình khác (ví dụ: Random Forest Regression).
6. Đưa ra nhận xét về mô hình tốt nhất.

Điểm số: 10 điểm
