GV: Phạm Thị Xuân Hiền

Bài tập Naive Bayes

1. Phân loại hoa Iris

Đề bài: Sử dụng bộ dữ liệu Iris, hãy xây dựng một mô hình Gaussian Naive Bayes để phân loại các loài hoa Iris. Tính độ chính xác của mô hình trên tập kiểm tra.

Hướng dẫn:
1. Sử dụng thư viện sklearn để tải bộ dữ liệu Iris.
2. Chia dữ liệu thành tập huấn luyện và tập kiểm tra (thường là 70% - 30%).
3. Khởi tạo mô hình Gaussian Naive Bayes từ sklearn.naive_bayes.
4. Huấn luyện mô hình trên tập huấn luyện.
5. Dự đoán trên tập kiểm tra.
6. Sử dụng accuracy_score để tính độ chính xác.

2. Xử lý dữ liệu thiếu

Đề bài: Tạo một bộ dữ liệu giả lập có chứa các giá trị thiếu, sử dụng Gaussian Naive Bayes để xử lý dữ liệu này và đánh giá hiệu suất của mô hình.

Hướng dẫn:
1. Sử dụng numpy để tạo một bộ dữ liệu giả lập.
2. Thêm các giá trị NaN vào dữ liệu để tạo giá trị thiếu.
3. Sử dụng SimpleImputer từ sklearn.impute để điền các giá trị thiếu.
4. Chia dữ liệu thành tập huấn luyện và kiểm tra.
5. Áp dụng Gaussian Naive Bayes và đánh giá hiệu suất.

3. So sánh với các phương pháp cụ thể Decision Tree và Random Forest.

Đề bài: Sử dụng bộ dữ liệu về chất lượng rượu vang, hãy so sánh hiệu suất của Gaussian Naive Bayes với các thuật toán khác như Decision Tree và Random Forest.

Hướng dẫn:
1. Tải bộ dữ liệu về chất lượng rượu vang từ sklearn hoặc UCI repository.
2. Chia dữ liệu thành tập huấn luyện và kiểm tra.
3. Huấn luyện ba mô hình: Gaussian Naive Bayes, Decision Tree, và Random Forest.
4. Dự đoán trên tập kiểm tra với cả ba mô hình.
5. So sánh độ chính xác của ba mô hình.

4. Xử lý dữ liệu không cân bằng

Đề bài: Tạo một bộ dữ liệu không cân bằng và sử dụng SMOTE (Synthetic Minority Over-sampling Technique) kết hợp với Gaussian Naive Bayes để cải thiện hiệu suất phân loại.

Hướng dẫn:
1. Sử dụng make_classification từ sklearn để tạo dữ liệu không cân bằng.

2. Chia dữ liệu thành tập huấn luyện và kiểm tra.
3. Áp dụng SMOTE từ thư viện imbalanced-learn trên tập huấn luyện.
4. Huấn luyện Gaussian Naive Bayes trên dữ liệu đã cân bằng.
5. Đánh giá hiệu suất trên tập kiểm tra, sử dụng các metric phù hợp cho dữ liệu không cân bằng như balanced accuracy hoặc F1-score.

5. Trực quan hóa kết quả phân loại

Đề bài: Sử dụng bộ dữ liệu Iris, hãy xây dựng mô hình Gaussian Naive Bayes và vẽ ma trận nhầm lẫn (confusion matrix) để trực quan hóa kết quả phân loại.

Hướng dẫn:
1. Tải và chia dữ liệu Iris như trong bài tập 1.
2. Huấn luyện mô hình Gaussian Naive Bayes.
3. Dự đoán trên tập kiểm tra.
4. Sử dụng confusion_matrix từ sklearn.metrics để tạo ma trận nhầm lẫn.
5. Sử dụng ConfusionMatrixDisplay để vẽ ma trận nhầm lẫn.
6. Phân tích kết quả: các ô trên đường chéo chính cho biết số lượng mẫu được phân loại đúng, các ô khác cho biết số lượng mẫu bị phân loại sai.