UỶ BAN NHÂN DÂN THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC SÀI GÒN
KHOA TOÁN - ỨNG DỤNG

BÁO CÁO
HỌC PHẦN: KHAI PHÁ DỮ LIỆU (858016)

Giảng viên hướng dẫn: Đỗ Như Tài

Sinh viên thực hiện: Trần Quốc Hoàng (3123580015)

Trần Chí Vỹ (3123580065)

Phan Trần Hữu Tấn (3123580042)

Lớp: DDU1231

Thành phố Hồ Chí Minh, ngày 14 tháng 11 năm 2025

QUÁ TRÌNH THAM GIA

Tên - MSSV	Phân công	Tham gia (%)
Trần Quốc Hoàng - 3123580015 (Nhóm trưởng)	Câu 1 của KNN, viết báo cáo	100%
Trần Chí Vỹ - 3123580065	Câu 2 của KNN	100%
Phan Trần Hữu Tấn - 3123580042	Câu 1, 2 của Naive Bayes	100%

BÁO CÁO

I. Báo cáo thực hành: K-Nearest Neighbors (KNN)

Thuật toán K-Nearest Neighbors (KNN) là một phương pháp phân lớp dựa trên sự tương tự, phân loại điểm dữ liệu mới dựa trên nhãn của K điểm dữ liệu gần nhất trong không gian đặc trưng.

1. Câu 1: Phân loại loài hoa Iris (Toy Example)

Tiêu chí	Chi tiết	Người thực hiện
Mục tiêu	Cài đặt chương trình demo thuật toán KNN từ đầu (from scratch) để phân loại hoa Iris.	Trần Quốc Hoàng (Nhóm trưởng)
Dữ liệu	UCI Iris (150 mẫu, 4 đặc trưng, 3 lớp). Tỷ lệ chia: 67% huấn luyện (100 mẫu), 33% kiểm thử (50 mẫu).	
Phương pháp	Sử dụng hàm tính Khoảng cách Euclidean để đo lường độ tương tự.	

Kết quả thực thi	Độ chính xác
KNN cơ bản (K = 3, Majority Vote)	98.00%
Tối ưu K (Majority Vote) (K = 1 đến K = 19)	K = 1 đạt 98.00%
Weighted KNN (K = 1, $1/d^2$ Vote)	98.00%

- **Ví dụ dự đoán sai (K = 3):** Một mẫu có 4 đặc trưng là [5.9, 3.2, 4.8, 1.8] có nhãn thực tế là **Iris-versicolor**, nhưng được dự đoán là **Iris-virginica** vì cả 3 láng giềng gần nhất đều là ['Iris-virginica', 'Iris-virginica', 'Iris-virginica'].

---

2. Câu 2: Nhận dạng ký tự A-Z

Tiêu chí	Chi tiết	Người thực hiện
Mục tiêu	Áp dụng thuật toán KNN để giải quyết bài toán nhận dạng ký tự viết tay, sử dụng thư viện scikit-learn cho hiệu năng tính toán.	Trần Chí Vỹ
Dữ liệu	UCI Letter-Recognition Data (20,000 mẫu, 26 lớp, 16 đặc trưng). Tỷ lệ chia: 16,000 huấn luyện, 4,000 kiểm tra.	
Thách thức	Quy mô và độ phức tạp tính toán lớn hơn hẳn so với Iris (26 lớp so với 3, 16 chiều).	

Kết quả thực thi	Độ chính xác
KNN cơ bản (K = 5, Uniform/Majority Vote)	95.33%
Tối ưu K (Uniform Vote) (K = 1 đến K = 11)	K = 1 đạt 96.03%
Weighted KNN (K = 1, $1/d^2$ Vote)	96.03%

- **Nhận xét:** Độ chính xác cao **96.03%** cho thấy KNN hoạt động hiệu quả trên bộ dữ liệu lớn, mặc dù có độ phức tạp tính toán cao.

II. Báo cáo thực hành: Naïve Bayes (NB)

Thuật toán Naïve Bayes là phương pháp phân loại dựa trên Định lý Bayes, hoạt động nhanh và hiệu quả nhờ giả định "ngây thơ" rằng các đặc trưng là độc lập khi đã biết lớp của đối tượng.

1. Câu 1 & Câu 2: Phân loại hoa Iris và Nhận dạng ký tự

Tiêu chí	Chi tiết	Người thực hiện
Mục tiêu kép	Cài đặt và áp dụng thuật toán Naïve Bayes cho cả hai bài toán: Phân loại hoa Iris và Nhận dạng ký tự.	Phan Trần Hữu Tấn
Bài toán Iris	Sử dụng mô hình Gaussian Naïve Bayes (phù hợp với đặc trưng số thực, liên tục).	
Bài toán Ký tự	So sánh hiệu năng của GaussianNB và MultinomialNB từ sklearn.	

Kết quả thực thi

Bộ dữ liệu	Mô hình Naïve Bayes	Độ chính xác	Nhận xét
Iris (Toy Example)	GaussianNB (from scratch)	96.00%	Thấp hơn KNN (98.00%) nhưng vẫn rất tốt cho một mô hình đơn giản.
Nhận dạng Ký tự (4,000 mẫu test)	GaussianNB (sklearn)	65.22%	Tốt hơn so với MultinomialNB.
Nhận dạng Ký tự (4,000 mẫu test)	MultinomialNB (sklearn)	55.53%	Thấp hơn đáng kể so với GaussianNB.

- **Ví dụ dự đoán sai trên Iris**: Một mẫu có 4 đặc trưng [6.7, 3.0, 5.0, 1.7] có nhãn thực tế là **Iris-versicolor** nhưng được dự đoán là **Iris-virginica**. Xác suất hậu nghiệm cho *Iris-virginica* (2.22×10^{-2}) cao hơn *Iris-versicolor* (6.68×10^{-5}).
- **Nhận xét trên bài toán Ký tự**: Naïve Bayes cho thấy ưu điểm về tốc độ so với KNN trên bộ dữ liệu lớn, do chỉ tính toán tham số xác suất

thay vì tính khoảng cách đến mọi điểm huấn luyện. Tuy nhiên, độ chính xác của Naïve Bayes (65.22%) lại thấp hơn đáng kể so với KNN (96.03%).

III. Tổng kết chung

- **KNN** cho kết quả phân loại chính xác hơn (Iris: 98.00%, Ký tự: 96.03%) nhưng có chi phí tính toán lớn trên bộ dữ liệu lớn.
- **Naïve Bayes** cho kết quả phân loại nhanh hơn (đặc biệt trên tập lớn) nhưng có độ chính xác thấp hơn (Iris: 96.00%, Ký tự: 65.22%) do giả định độc lập.