
Trí tuệ nhân tạo

# Tìm hiểu và ứng dụng phương pháp KNN

*Nhóm 3*

Ngày 15 tháng 4 năm 2024

**Sinh viên thực hiện**
106200284 - Hồ Đức Vũ - 20KTMT2
106200241 - Nguyễn Minh Phương - 20KTMT1
106200240 - Huỳnh Vũ Đình Phước - 20KTMT1

**Giáo viên hướng dẫn**
TS. Hoàng Lê Uyên Thục

**Môn học**
Trí tuệ nhân tạo

**Đề tài**
Tìm hiểu và ứng dụng phương pháp KNN

**Xuất bản**
Đà Nẵng, Ngày 15 tháng 4 năm 2024

**Số trang**
5

# Mục lục

# Nội dung báo cáo

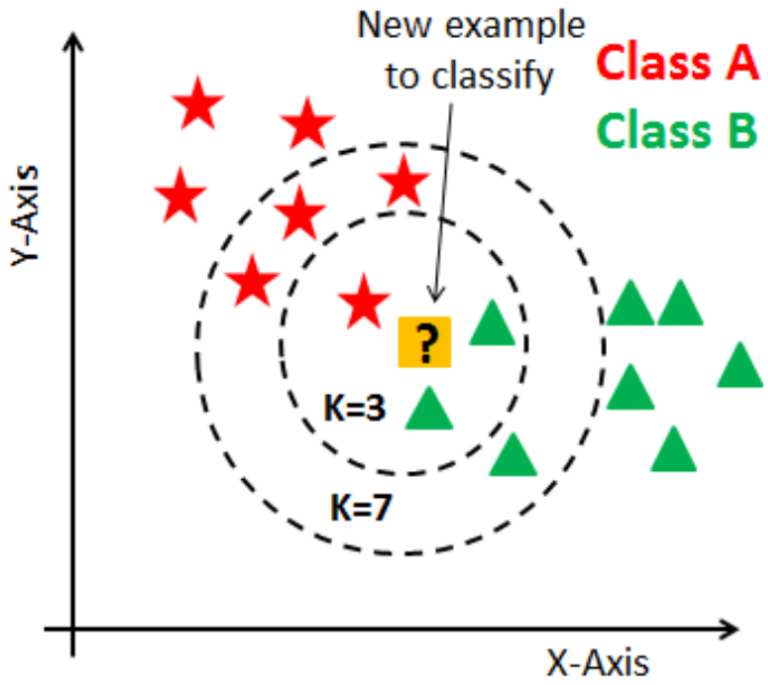
## 1 Giới thiệu

Máy học đã thay đổi cách chúng ta tiếp cận và giải quyết các vấn đề phức tạp trong thế giới số ngày nay, và bài toán phân loại không phải là ngoại lệ. Trong thực tế, phân loại là một trong những lĩnh vực quan trọng và hấp dẫn nhất của máy học, xuất hiện rộng rãi trong đời sống hàng ngày. Từ việc lọc email vào hòm thư đến nhận dạng ảnh khuôn mặt trong các ứng dụng an ninh, phân loại đóng vai trò không thể phủ nhận trong việc giúp tự động hóa các quy trình và tối ưu hóa công việc của con người.

Một ví dụ cụ thể về bài toán phân loại là việc phân loại hoa diên vĩ dựa trên các đặc trưng. Điều này không chỉ mang lại giá trị thực tế trong việc hiểu biết về các loài hoa mà còn là một bước tiến quan trọng trong việc áp dụng Machine Learning vào thế giới tự nhiên. Bằng cách sử dụng dữ liệu cụ thể từ cơ sở dữ liệu Iris, chúng ta có thể xây dựng một mô hình có khả năng dự đoán loài hoa dựa trên các đặc trưng được cung cấp.

Ở đây, nhóm sẽ sử dụng một trong những thuật toán phân loại cơ bản nhất K-Nearest Neighbors (KNN) để xây dựng một mô hình có khả năng dự đoán loài hoa dựa trên các đặc trưng được cung cấp và áp dụng phương pháp k-fold cross-validation để đánh giá mô hình. Bằng cách này, chúng ta sẽ không chỉ tiến xa hơn trong việc hiểu và dự đoán về loài hoa diên vĩ, mà còn làm sâu rộng hơn về lĩnh vực học máy và ứng dụng của nó trong thế giới thực.

## 2 Phương pháp KNN




Hình 2.1: Thuật toán KNN

Trong lĩnh vực học máy, thuật toán KNN (K-Nearest Neighbors) đã trở thành một công cụ vô cùng quan trọng và hiệu quả. Dù đơn giản, nhưng KNN thuộc loại thuật toán học có giám sát, tức là nó học từ dữ liệu đã được gán nhãn để dự đoán nhãn cho dữ liệu mới.

Tầm quan trọng của KNN không chỉ dừng lại ở độ phức tạp thấp mà còn ở tính linh hoạt cao. Đây là một trong những lựa chọn phổ biến cho nhiều vấn đề phân tích dữ liệu, từ phân loại đến hồi quy và hệ thống gợi ý.

KNN hoạt động dựa trên nguyên tắc rằng các điểm dữ liệu tương tự nhau thường nằm gần nhau. Khi cần dự đoán nhãn cho một điểm dữ liệu mới, KNN xem xét 'K' điểm dữ liệu gần nhất và dựa trên đa số nhãn của chúng để quyết định nhãn cho điểm dữ liệu mới đó.

Giá trị 'K' là một siêu tham số quan trọng trong KNN, quy định số lượng hàng xóm gần nhất sẽ được xem xét. Lựa chọn K phù hợp rất quan trọng, với K quá nhỏ có thể dẫn đến mô hình bị ảnh hưởng bởi nhiễu dữ liệu, trong khi K quá lớn có thể làm mô hình không nhạy với các đặc điểm cụ thể của dữ liệu.

Việc tính toán khoảng cách giữa các điểm dữ liệu là một phần quan trọng của KNN. Tùy thuộc vào bản chất của dữ liệu mà ta lựa chọn phương pháp tính khoản cách phù hợp. Trong bài báo cáo này, ta sẽ sử dụng các chuẩn norm-1, norm-2.

Manhattan distance(norm-1): $\sum_{i=1}^{n} |x_i - y_i|$

Euclidean distance(norm-2): $\sqrt{\sum_{i=1}^{n} |x_i - y_i|^2}$

Trên hết, KNN không chỉ là một thuật toán đơn giản mà còn là một công cụ mạnh mẽ và linh hoạt trong học máy, đóng góp vào việc giải quyết nhiều vấn đề phức tạp trong thế giới thực.

# 3 Thử nghiệm và ứng dụng

## 3.1 CSDL

Tập dữ liệu Iris là một trong những tập dữ liệu phổ biến và kinh điển trong lĩnh vực học máy và thống kê. Được giới thiệu bởi nhà thống kê người Anh Ronald Fisher vào năm 1936, tập dữ liệu này thường được sử dụng để minh họa các thuật toán phân loại và phân tích dữ liệu.

| Id | f1 | f2 | f3 | f4 | Class | folder |
|---|---|---|---|---|---|---|
| 1 | 5.1 | 3.5 | 1.4 | 0.2 | 0 | 1 |

Hình 3.1: Các đặc trưng của cơ sở dữ liệu

Trong hình 3.1 là các đặc trưng của tập dữ liệu Iris bao gồm thông tin về loài hoa iris: Iris Setosa, Iris Versicolor. Đối với mỗi loài hoa, có bốn đặc trưng được đo lường:

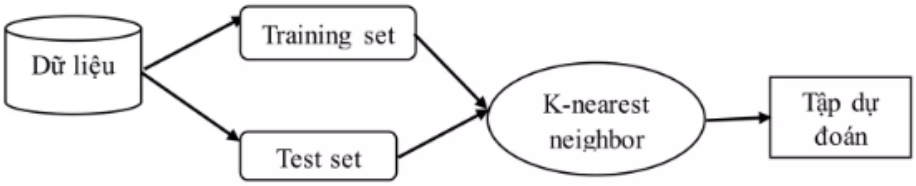
f1 - Chiều dài đài hoa (Sepal Length) - Đo lường độ dài của lá đài của hoa iris.

f2 - Chiều rộng đài hoa (Sepal Width) - Đo lường chiều rộng của lá đài của hoa iris.

f3 - Chiều dài cánh hoa (Petal Length) - Đo lường độ dài của cánh hoa của hoa iris.

f4 - Chiều rộng cánh hoa (Petal Width) - Đo lường chiều rộng của cánh hoa của hoa iris.

Trong tập dữ liệu ta có 100 mẫu, cứ mỗi 10 mẫu ta gộp thành 1 folder, ta chia dữ liệu với tỉ lệ 90-10 lần lượt để train và test.




Hình 3.2: Quá trình thực hiện bài toán KNN

## 3.2 Kịch bản

Trong phần này, từ 100 dữ liệu trong tập Iris flower, ta sẽ tách nó thành hai phần gồm training set và test set. Thuật toán sẽ dựa vào thông tin ở phần training set để dự đoán xem mỗi dữ liệu ở phần test set sẽ tương ứng với loại hoa nào. Dữ liệu được chuẩn đoán này sẽ được đối chiếu với loại hoa thật của mỗi dữ liệu trong tập test set để đánh giá độ hiệu quả của KNN.

Bước 1: Khai báo các thư viện cần thiết
Bước 2: Load dữ liệu. Bước này sẽ đọc tất cả dữ liệu có trong file csv và lưu vào các tập dữ liệu cần thiết.
Bước 3: Tách các tập training set và test set.
Giả sử ta có 10 tập dữ liệu, 1 tập dữ liệu gồm 10 dữ liệu. Ta sẽ chia tập thứ nhất làm tập test còn 9 tập còn lại làm tập training. Và làm tương tự với các lần thực hiện sau.
Bước 4: Dùng KNN để dự đoán.
Bước 5: Đánh giá độ chính xác.
Để đánh giá độ chính xác của thuật toán KNN, ta sẽ xem thử có bao nhiêu điểm trong tập test được dự đoán đúng. Lấy số lượng đó chia cho tổng số lượng có trong tập test thì sẽ cho ra độ chính xác của mô hình.

## 3.3 Kết quả - nhận xét


Hình 3.3: Kết quả thử nghiệm phương pháp KNN

Nhận xét: nhận thấy rằng các kết quả đánh giá mô hình KNN trên tập dữ liệu Iris cho cả hai giá trị K=3 và K=5, với cả hai loại khoảng cách Manhattan (norm-1) và Euclidean (norm-2), đều đạt độ chính xác là 100% trong tất cả các cấu hình. Điều này cho thấy rằng mô hình đã thể hiện sự hiệu quả đáng kể trong việc phân loại các loài hoa diên vĩ trong bộ dữ liệu này.

# 4 Kết luận

KNN là một phương pháp đơn giản và dễ triển khai. Không cần có quá nhiều tiền xử lý dữ liệu hay giả định về phân phối của dữ liệu. KNN tuy có độ phức tạp thấp nhưng vẫn cho ra kết quả rất tốt trên tập dữ liệu. việc lựa chọn đối số k - số lượng láng giềng gần nhất cần xem xét khi dự đoán. Quá ít hoặc quá nhiều láng giềng có thể ảnh hưởng đến hiệu suất của mô hình. KNN có thể là một lựa chọn hợp lý, nhưng cần phải được thử nghiệm và điều chỉnh một cách cẩn thận để đạt được hiệu suất tốt nhất.

# 5 Tài liệu tham khảo

[1] Dataset Iris. https://www.kaggle.com/datasets/uciml/iris
[2] Slide bài giảng.