1. **Bài toán phân loại màu ảnh**

Phân loại màu ảnh là một nhiệm vụ quan trọng trong xử lý ảnh số, có ứng dụng rộng rãi trong nhiều lĩnh vực như y tế, công nghiệp, nông nghiệp, v.v. Mục tiêu của bài toán là tự động phân loại các pixel trong ảnh thành các lớp màu khác nhau, ví dụ như màu đỏ, xanh lá, xanh lam, v.v.

* Phân tích hình ảnh y tế: ví dụ như hình ảnh X-quang, để phát hiện các bệnh lý.
* Kiểm soát chất lượng trong công nghiệp: ví dụ như để phát hiện các lỗi trong sản phẩm..
* Phân loại nông sản: ví dụ như để phân loại trái cây theo độ chín.

Hiện nay, nghiên cứu về ứng dụng các thuật toán học máy cho bài toán phân loại màu ảnh vẫn đang được tiếp tục để cải thiện độ chính xác và hiệu quả của các phương pháp phân loại.

Trong đề tài này, thì nhóm 3 chúng em sẽ ứng dụng thuật toán k-láng giềng gần nhất cho bài toán phân lại màu ảnh. Đây là một thuật toán học có giám sát dễ hiểu và dễ triển khai.

1. **Thuật toán KNNs cho bài toán phân lớp.**

Thuật toán KNN hay còn gọi là học lười, học dựa trên ví dụ. Đây là một thuật toán không làm gì trong quá trình học, chỉ đơn giản là nó ghi nhớ những ví dụ mà nó đã được học.

**Ý tưởng:** Trong quá trình học, thuật toán chỉ lưu lại những mẫu huấn luyện được cung cấp. Khi cần dự đoán nhãn cho mẫu mới, thuật toán tìm k mẫu huấn luyện gần nhất và xác định nhãn phân loại dựa k mẫu dữ liệu gần nhất.

* + Với một tập các ví dụ học (huấn luyện)
    - Đơn giản là lưu lại các ví dụ học
  + Đối với một ví dụ cần phân loại/dự đoán
    - Xét quan hệ giữa ví dụ đó với tất cả các ví dụ học để gán giá trị của hàm mục tiêu (một nhãn lớp, hoặc một giá trị thực)

**Thuật toán KNN đối với bài toán phân lớp:**

* + Mỗi ví dụ x được biểu diễn bởi hai thành phần:
    - Vecto đặc trưng của ví dự: x = (x1, x2, . . . , xn), trong đó xi ∈ R
    - Nhãn lớp của ví dụ
  + Giai đoạn học (huấn luyện): Lưu lại các ví dụ trong tập dữ liệu huấn luyện
  + Giai đoạn phân lớp: Để phân lớp cho ví dụ mới z:
    - Tính khoảng cách giữa z và tất cả ví dụ x trong tập
    - Xác định tập láng giềng gần nhất của z => Xác định k ví dụ học trong tập dữ liệu huấn luyện gần nhất với z theo một hàm khoảng cách d
    - Nhãn của z sẽ là nhãn chiếm đa số trong tập k láng giềng kia.

A red and blue dots with a question mark

Description automatically generated

Phân loại màu sử dụng KNN

**Ưu, nhược điểm của thuật toán**

* Ưu điểm
  + Đơn giản, dễ hiểu, dễ triển khai.
  + Độ phức tạp tính toán của quá trình training là bằng 0.
* Nhược điểm
  + KNN có thể bị ảnh hưởng nhiều bởi các nhiễu hoặc dữ liệu không chuẩn.
  + KNN yêu cầu lưu trữ toàn bộ tập dữ liệu huấn luyện trong bộ nhớ 🡪 điều này có thể là một vấn đề với các tập dữ liệu lớn
  + Tốc độ chậm trong giai đoạn phân loại, do phải so sánh ví dụ mới với toàn bộ tập học.

1. **Ứng dụng thuật toán KNN vào bài toán phân loại màu ảnh**
2. Sơ đồ phân loại

Trong bài tập lớn này, màu sắc được phân loại bằng cách sử dụng thuật toán phân loại Machine Learning K-Nearest Neighbors.

Trong sơ đồ hình 3.1:

• Giai đoạn tiền xử lý dữ liệu: đây là giai đoạn đặc trưng hóa dữ liệu, hay biểu diễn mẫu ví dụ x thành n các đặc trưng của nó để đem đi phân loại

• Giai đoạn phân loại: sử dụng thuật toán KNN

1. Mô tả tập dữ liệu

Tập dữ liệu của mô hình bao gồm 2 loại: tập Training, tập Test.

Tập Training là file csv chứa 3 đặc trưng được phân tích theo biểu đồ màu RGB và nhãn của 5053 màu được chia làm 11 nhóm màu: đen, xanh dương, xanh lá, nâu, xám, cam, hồng, tím, đỏ, trắng và vàng với số lượng được mô tả trong hình 3.2.

Mỗi màu là một dòng trong file csv, cụ thể trong hình 3.3.

Tập Test gồm 307 file ảnh với đầy đủ 11 nhóm màu kể trên, chưa được trích xuất đặc trưng.

Tập Training được sử dụng để huấn luyện mô hình còn tập Test được sử dụng để đánh giá độ chính xác sau khi huấn luyện.

1. Đặc trưng hóa dữ liệu

**Đối với mẫu dữ liệu dùng để Training/Test**

Để lấy các đặc trưng về màu sắc của ảnh, ta sử hàm color\_histogram\_of\_imag để tính toán các histogram của 1 ảnh tên img\_nam theo biểu đồ màu RGB, sau đó đặc trưng của mẫu dữ liệu sẽ được ghi vào tệp testing.data cùng với nhãn mô tả. Nhãn sẽ được lấy theo tên của thư mục chứ tệp

**Đối với mẫu dữ liệu đem đi dự đoán**

Sử dụng hàm color\_histogram\_of\_test\_image với chức năng và cách làm tương tựhàm color\_histogram\_of\_imag, tuy nhiên được bỏ đi phần gán nhãn

1. Phân loại sử dụng KNN

Xây dựng hàm tính khoảng cách: dùng khoảng cách O-clit

Chọn k láng giềng gần nhất

Xác định nhãn chiếm đa số

Dự đoán nhãn: mang nhãn chiếm đa số trong k láng giềng

1. Xây dựng mô hình hoàn chỉnh

File ảnh được truyền vào, các đặc trưng được trích xuất bằng cách sử dụng hàm color\_histogram\_of\_test\_image và đưa vào giải thuật KNN để phân loại màu sắc chủ đạo của ảnh.

1. Dự đoán nhãn cho ảnh

Sau khi đã xây dựng các hàm cần thiết, ta áp dụng chúng để dự đoán màu chủ đạo (nhãn) cho 1 file ảnh bất kỳ. Đặc trưng của file ảnh lưu trong tệp có đường đẫn là testing\_path. Các đặc trưng này được đưa qua KNN để phân loại, nhãn của ảnh sẽ là nhãn chiếm đa số của k láng giềng gần nhất.

1. Demo (Chạy tay)