**KẾ HOẠCH THỰC HIỆN TIỂU LUẬN,BÀI TẬP LỚN,ĐỒ ÁN**

Tên lớp :KHMT1 Khóa : 12

Họ và tên sinh viên : Vũ Văn Hùng, Phạm Hoàng Tiến, Phạm Văn Yên

Tên nhóm : Nhóm 3

Tên chủ đề : Nhận dạng biển số xe trên ảnh sử dụng KNN

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Tuần** | **Người thực hiện** | **Nội dung công việc** | **Kết quả đạt được** | **Phương pháp thực hiện** |
| 1 | Cả nhóm | Tìm hiểu đề tài | Đã xác định được đầu vào và đẩu ra của bài toán. | Cả nhóm cùng thảo luận |
| 1 | Cả nhóm | Tìm hiểu về thuật toán KNN | - Đã có biết cơ bản bản chất của thuậ toán và quy trình hoạt động  - Xác đinh các ưu và nhược điểm của thuật toán từ đó xác định cách làm tiếp theo của trong btl | Đọc tài liệu:  [Thuật toán KNN](https://www.slideshare.net/trongthuy3/luan-van-nghien-cuu-thuat-toan-knearest-neighbor-haym-9d)  [Uu+nhược KNN](https://machinelearningcoban.com/2017/01/08/knn/) |
| 2 | Cả nhóm | Tìm hiểu về dataset , tìm kiếm một số bộ dataset | Có được dataset và hiểu được thuật toán KNN |  |
| 3 | Cả nhóm | Code data trainning cho KNN | Có được 2 file trainning data: classifications.txt và flattened\_images.txt | Tham khảo tài liệu:  Trainning data on kaggle.com  (Người báo cáo Vũ Văn Hùng) |

1. **TÌM HIỂU THUẬT TOÁN KNN**

**1. Giới thiệu về thuật toán KNN**

KNN (K-Nearest Neighbors) là một trong những thuật toán học có giám sát đơn giản nhất được sử dụng nhiều trong khai phá dữ liệu và học máy. Ý tưởng của thuật toán này là nó không học một điều gì từ tập dữ liệu học (nên KNN được xếp vào loại lazy learning), mọi tính toán được thực hiện khi nó cần dự đoán nhãn của dữ liệu mới. Lớp (nhãn) của một đối tượng dữ liệu mới có thể dự đoán từ các lớp (nhãn) của k hàng xóm gần nó nhất. Hay ngắn gọn ý tưởng là tìm ra output của dữ liệu dựa trên thông tin của những dữ liệu training gần nó nhất

### \* Đầu vào và đầu ra

### Thuật toán có 2 đầu vào, một là tập các dữ liệu đã biết trước kiểu(loại) của từng dữ liệu(hay còn gọi là tập huấn luyện - training set), đầu vào thứ 2 là dữ liệu, chúng ta chưa biết kiểu(loại) dữ liệu đó. Đầu ra của thuật toán kNN là kiểu dữ liệu của đầu vào thứ 2.

2.1.2. **Quy trình làm việc của thuật toán KNN**

- Bước 1: xác định tham số K= số láng giềng gần nhất.

-Bước 2: tính khoảng cách đối tượng cần phân lớp với tất cả các đối tượng trong training data.

-Bước 3: sắp xếp khoảng cách theo thứ tự tăng dần và xác định K láng giềng gần nhất với đối tượng cần phân lớp

- Bước 4: lấy tất cả các lớp của K láng giềng gần nhất.

- Bước 5: dựa vào phần lớn lớp của K để xác định lớp cho đối tượng cần phân lớp.

### 3.Ưu, nhược điểm của KNN

### -Ưu điểm

+Độ phức tạp tính toán của quá trình training là bằng 0.

+Việc dự đoán kết quả của dữ liệu mới rất đơn giản.

+Không cần giả sử gì về phân phối của các class.

### -Nhược điểm của KN

+KNN rất nhạy cảm với nhiễu khi K nhỏ.

+Như đã nói, KNN là một thuật toán mà mọi tính toán đều nằm ở khâu test. Trong đó việc tính khoảng cách tới từng điểm dữ liệu trong training set sẽ tốn rất nhiều thời gian, đặc biệt là với các cơ sở dữ liệu có số chiều lớn và có nhiều điểm dữ liệu. Với K càng lớn thì độ phức tạp cũng sẽ tăng lên. Ngoài ra, việc lưu toàn bộ dữ liệu trong bộ nhớ cũng ảnh hưởng tới hiệu năng của KNN.

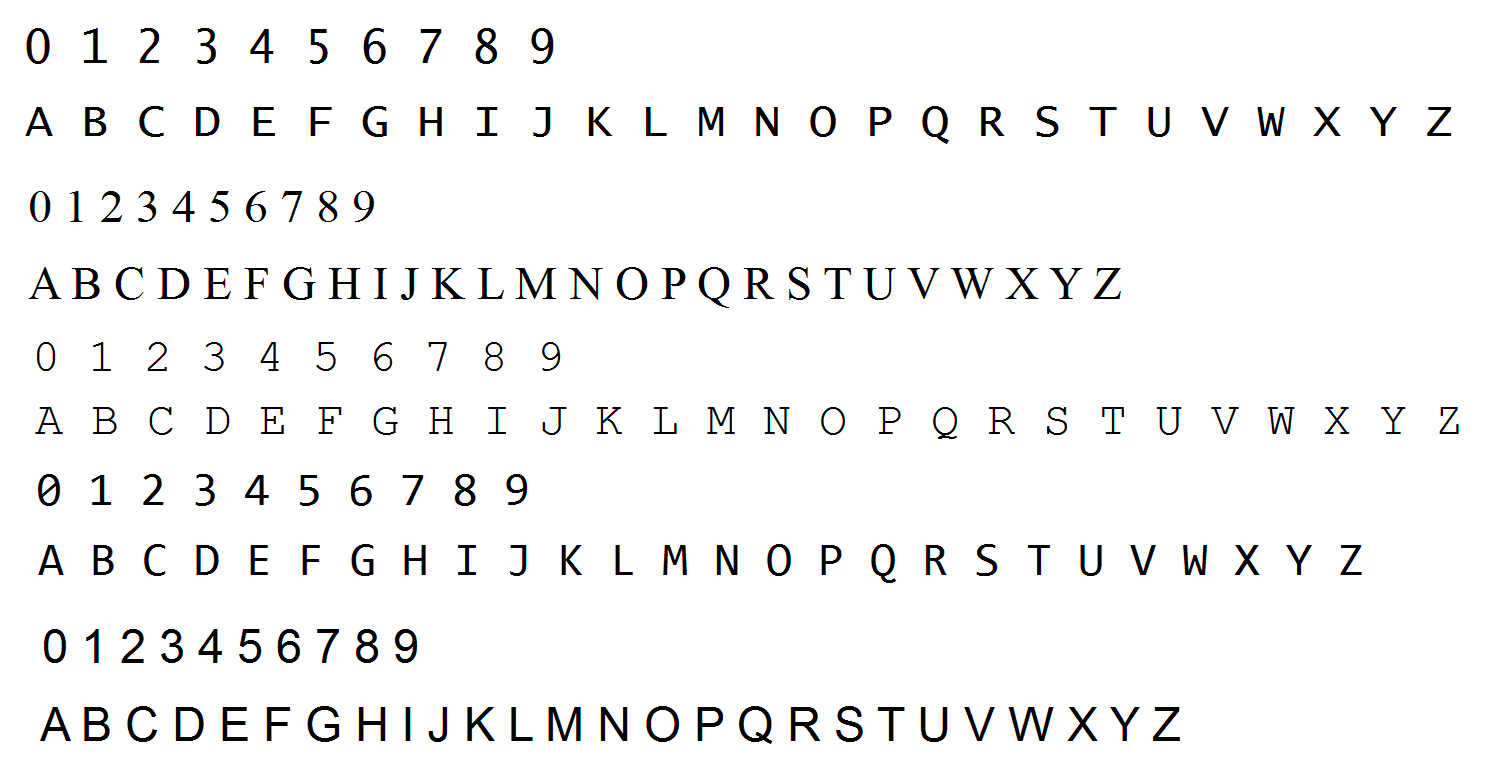
+Với K nhỏ dễ gặp nhiễu dẫn tới kết quả đưa ra không chính xác

+Cần nhiều thời gian để thực hiện do phải tính toán khoảng cách với tất cả các đối tượng trong tập dữ liệu.

+Cần chuyển đổi kiểu dữ liệu thành các yếu tố định tính.

1. **Module trainning data**

* Thực hiện code module dùng để trainning các tập dữ liệu với các chữ cái từ A-Z và các chữ số từ 0-9



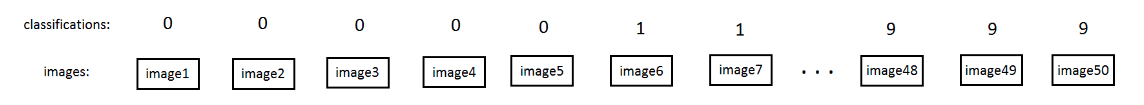
***Hình 1: Các tập dữ liệu đặc trưng dùng để trainning***

Với các kiểu chữ đặc trưng cho biển số xe. Mỗi chữ cái hay chữ số đều có 5 trặc trưng khác nhau trong bộ dữ liệu trainning này.

* Giả sử, chúng ta nhận ra các chữ số từ 0-9 và mỗi chữ số đều có 5 ảnh đặc trưng vậy ta có tất cả: 5 (đặc trưng) \* 10 (chữ số) = 50 hình ảnh để đào tạo
* Giả sử kích thước mỗi hình ảnh là 10x10 pixel, tức là mỗi ảnh sẽ có tổng kích thước tối đa là 100px. Sau khi đào tạo xong sẽ tạo ra 2 cấu trúc dữ liệu song song:
* Group/Classification: là một bộ số cho biết hình ảnh tương ứng thuộc phân loại nào. Một đối tượng phân loại bằng một phiếu bầu nhiều người láng giềng của nó với đối tượng đặc trưng phổ biến nhất trong số k cụm láng giềng gần nhất với nó
* A set of images/Flattened\_images: là một bộ hình ảnh đã được đào tạo,

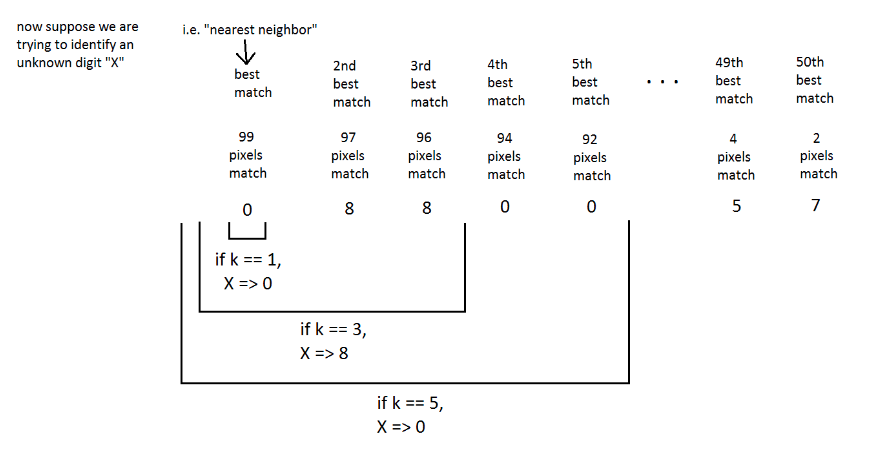
Là những giá trị thuộc tính của đối tượng, giá trị này được coi là giá trị trung bình của k lân cận gần nhất

* Mô tả thuật toán để tìm ra các tập dữ liệu đặc trưng



Ví dụ ta có 1 bộ dữ liệu như trên:

* Khi quá trình trainning hoàn tất



Giả dụ chúng ta đang cố gắng xác định 1 chữ số không xác định X thì:

* Khi số cụm k = 1 thì giá trị hàng xóm gần nhất của X = 0
* Khi số cum k = 3 thì giá trị hàng xóm gần nhất của X = 8
* Khi số cụm k = 5 Thì giá trị hàng xóm gần nhất của X = 0
* Tương tự với các trường hợp cụm còn lại (trong ví dụ này k\_min = 1 và k\_max = 50)

Với dữ liệu đầu vào như hình 1 ta tạo ra những tập đặc trưng cho lớp và bộ thuộc tính của từng đối tượng với 5 đặc trưng khác nhau

(Classification/Set of image – Xem thêm ở file classifications.txt và flattened\_images.txt)