**TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP. HỒ CHÍ MINH**

**KHOA ĐÀO TẠO CHẤT LƯỢNG CAO**

**NGÀNH CÔNG NGHỆ THÔNG TIN**

**---o0o---**



**ĐỒ ÁN 3**

**TÌM HIỂU VỀ THUẬT TOÁN K-NEAREST-NEIGHBOR**

**Nhóm sinh viên thực hiện**

17110220 Nguyễn Văn Tây

17110217 Trần Nguyên Tài

**GVHD: TS.Lê Văn Vinh**

**Tp. Hồ Chí Minh, tháng 12 năm 2020**

**LỜI CAM ĐOAN**

Nhóm xin cam đoan. Đồ án 3 với đề tài “Tìm hiểu thuật toán K-nearest neighbor” là kết quả tìm hiểu, nghiên cứu của nhóm từ những kiến thức được thầy cô giảng dạy truyền đạt trong thời gian qua và một số tài liệu được nhóm chúng em tìm hiểu sưu tập qua các phương tiện internet.

**MỤC LỤC**

**DANH MỤC HÌNH ẢNH**

**DANH MỤC BẢNG**

**LỜI CẢM ƠN**

Lời đầu tiên cho phép nhóm em gởi lời cảm ơn sâu sắc đến thầy Lê Văn Vinh, người đã hướng dẫn truyền đạt và chỉ dẫn cho nhóm em nhứng kiến thức, bài học và kinh nghiệm kể từ khi bắt đầu nhận đề tài.

Để hoàn thành đồ án này nhóm em cũng xin gởi lời cảm ơn đến những người bạn đã luôn hỗ trợ chia sẻ những kiến thức luôn ở bên động viên khuyến khích để nhóm em có thể hoàn thành được đồ án này

Sau nữa nhóm xin cảm ơn đến bạn bè vì luôn bên cạnh giúp đỡ, động viên trong suốt thời gian thực hiện đề tài để nhóm em có thể hoàn thành tốt đề tài của mình

Trong quá trình nghiên cứu và báo cáo do năng lực, kiến thức, trình độ bản thân còn hạn hẹp nên không tránh những thiếu sót. Nhóm em mong nhận được sự thông cảm và những ý kiến đóng góp của thầy cô và các bạn

Nhóm xin chân thành cảm ơn!

TP.HCM, ngày 30 tháng 12 năm 2020

Sinh viên thực hiện

Nguyễn Văn Tây

Trần Nguyên Tài

**MỞ ĐẦU**

* 1. **Đặt vấn đề.**

Những năm gần đây , trí tuệ nhân tạo nổi lên như một bằng chứng cho cuộc cách mạng công nghệ 4.0. Trí tuệ nhân tạo có thể định nghĩa như một ngành của khoa học máy tính liên quan đến việc tự động hóa các hành vi thông minh. Trí tuệ nhân tạo là một bộ phận khoa học máy tính và do đó nó phải được đặt trên những nguyên lý lý thuyết vững chắc, có khả năng ứng dụng cao.

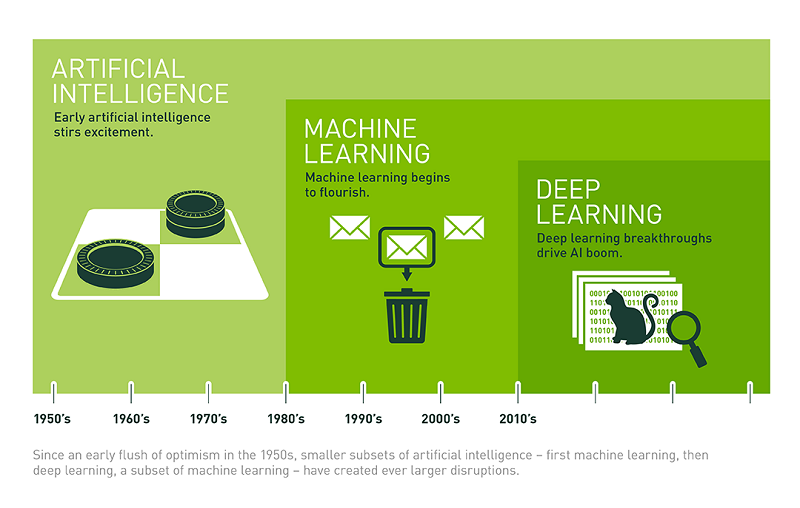
Ở thời điểm hiện tại thuật ngữ này thường được dùng để nói đến các máy tính có mục đích không nhất định và ngành khoa học nghiên cứu về các lý thuyết và các ứng dụng của trí tuệ nhân tạo.

Theo đà phát triển của công nghệ, ứng dụng trí tuệ nhân tạo luôn là xu hướng công nghệ tương lai và các hãng công nghệ trên toàn thế giới đua nhau sáng tạo. Nó là nền tảng cốt lõi của cuộc cách mạng 4.0

Machine Learning là một lĩnh vực của trí tuệ nhân tạo, được sinh ra từ khả năng nhận diện mẫu và từ các lý thuyết các máy tính có thể học mà không cần phải lập trình để xử lý nhiệm vụ cụ thể nào đó.

Hầu hết mọi ngành công nghiệp đang làm việc với lượng dữ liệu lớn đều nhận ra tầm quan trọng của công nghệ Machine Learning. Những cái nhìn sáng suốt từ nguồn dữ liệu này chủ yếu dưới dạng thời gian thực, sẽ giúp các tổ chức vận hành hiệu quả hoặc tạo lợi thế cạch tranh với các đối thủ khác.

Các ứng dụng của Machine Learning đã quá quen thuộc với con người như: Xe tự hành của Google , hệ thông tự tag khuôn mặt trên facebook,hệ thống gợi ý sản phầm của amazone,…. Chỉ là một trong vô vàn những ứng dụng của trí tuệ nhân tạo mang lại



Xu hướng phát triển công nghệ thông tin ngày càng tăng, song song với nó lượng dữ liệu được sinh ra cũng ngày một lớn. vì vậy nhu cầu để xử lý dữ liệu cũng lớn hơn, Machine learning góp phần giải quyết vấn đề này. Một trong những thuật toán thường dùng đó là thuật toán K-nearest neighbor. Ứng dụng của thuật toán này được sử dụng rộng rãi trong các bài toán phân lớp.

* 1. **Mục đích , phạm vi và đối tượng nghiên cứu**
* Mục đích : Nghiên cứu, tìm hiểu thuật toán K-nearest neighbor, đánh giá hiệu quả của thuật toán
* Phạm vi nghiên cứu: Thử nghiệm trên Iris flower dataset
* Đối tượng nghiên cứu: thuật toán K-nearest neighbor và bộ Iris flower dataset.
  1. **Nộ i dung thực hiện**
* Tìm hiểu thuật toán
* Làm quen bộ dữ liệu Iris Flower dataset
* Sử dụng bộ dữ liệu thử nghiệm và đánh giá
  1. **Cấu trúc đồ án**
* **Mở đầu**
* Chương 1: cơ sở lý thuyết
* Chương 2: Thuật toán K-nearest neighbor
* Chương 3 : Thử nghiệm
* Chương 4: Kết luận

1. **Chương 1 : Cơ sở lý thuyết**
   1. **Machine Learning.**
      1. **Định nghĩa**

Là một lĩnh vực của trí tuệ nhân tạo liên quan dến việc nghiên cứu và xây dựng các kỹ thuật cho phép các hệ thống học tự động từ dữ liệu để giải quyết tất cả các vấn đề cụ thể. Ví dụ các máy có thể học cách phân loại thư điện tử có phải thư rác hay không và tự động sắp xếp vào thư mục tương ứng

Machine Learning có liên quan đến thống kê vì cả hai lĩnh vực đều nghiên cứu việc phân tích dữ liệu, nhưng khác với thống kê học máy tập trung vào sự phức tạp của các giải thuật trong việc thực thi tính toán

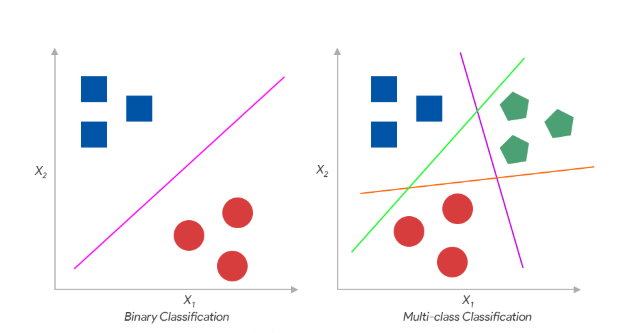
Machine Learning hiện nay áp dụng rộng rãi bao gồm máy truy tìm dữ liệu, máy phân tích thị trường chứng khoán, nhận dạng tiếng nói và chữ viết…

* + 1. **Một số phương thức của Machine Learning**

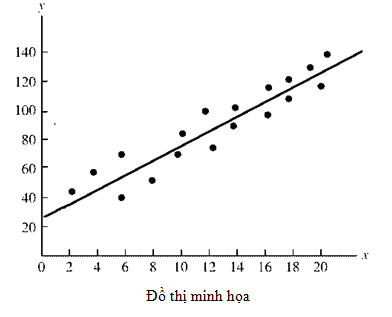
**Học có giám sát**: Thuật toán dự đoán đầu ra của một dữ liệu mới (new input) dựa trên cặp (input, outcome) đã biết từ trước. Cặp dữ liệu này còn gọi là ( data,label), tức (dữ liệu, nhãn). Supervised learning là nhóm phổ biến nhất trong các thuật toán Machine Learning.

Học có giám sát chia làm hai loại:

* Classification (phân lớp) : Là quá trình phân lớp một đối tượng dữ liệu vào hay nhiều lớp như đã cho trước nhờ một mô hình phân lớp (model). Mô hình này được xây dựng dựa trên tập dữ liệu được xây dựng trước đó có gán nhãn ( tập huấn luyện). Quá trình phân lớp là quá trình gắn nhãn cho đối tượng dữ liệu. Có nhiều nhiều bài toán phân lớp như phân lớp nhị phân , phân lớp đa lớp,phân lớp đa trị… Trong số đó phân lớp nhị phân là một loại phân lớp đặc biệt của phân lớp đa lớp. Ứng dụng của bài toán phân lớp được ứng dụng rất nhiều như nhận dạng khuôn mặt , nhận dạng chữ viết, nhận dạng giọng nói….



* Regression (hồi quy) : Nếu không được chia thành các nhóm mà là giá trị thực cụ thể. Đầu ra của diểm dữ liệu sẽ bằng chính đầu ra của điểm dữ liệu đã biết



**Học không giám sát :**  Là một kĩ thuật của máy học nhằm tìm ra một mô hình hay cấu trúc bị ẩn bởi tập dữ liệu không được gán nhãn cho trước. UL khác với SL là không thể xác đị nh trước output từ tập dữ liệu huấn luyện được. Tùy thuộc vào tập huấn luyện kết quả output sẽ khác nhau. Trái ngược với SL, tập dữ liệu huấn luyện của UL không do con người gán nhãn, máy tính sẽ phải tự học hoàn toàn. Có thể nói, học không giám sát thì giá trị đầu ra sẽ phụ thuộc vào thuật toán UL. Ứng dụng lớn phổ biến của học không giám sát là bài toán phân cụm.

**Học bán bám sát**: các bài toán khi có một số lượng lớn dữ liệu lớn nhưng chỉ một phần trong chúng được dán nhãn. Nhưng bài toán này nằm giữa phương thức học giám sát và không giám sát.

* 1. **Bài toán phân lớp dữ liệu**
     1. **Quá trình phân lớp dữ liệu**

Để xây dựng được mô hình phân lớp và đánh giá hiệu quả của mô hình cần phải thực hiện các bước sau:

* + - 1. Chuẩn bị tập dữ liệu huấn luyện và rút trích đặc trưng

Công đoạn này được xem là công đoạn quan trọng nhất của bài toán về machine learning. Vì đây là input cho việc để tìm ra mô hình của bai toán. Chúng ta cần biết phải chọn ra những đặc trung của tập dữ liệu, lược bỏ những đặc trưng không tốt của dữ liệu, gây nhiễu. Ước lượng số chiều của dữ liệu hay nói cách khác là chọn feature. Nếu số lượng dữ liệu nhiễu quá lớn sẽ gây khó khăn cho việc tính toán thì phải giảm số chiều của dữ liệu nhưng vẫn được độ chính xác của dữ liệu đó.

Ở bước này chũng ta cũng chuẩn bị bộ dữ liệu test trên mô hình. Thông thường sẽ sử dụng cross-validation (kiểm tra chéo) để chia tập dataset thành hai phàn, một phần phục vụ cho training và phần còn lại phục vụ cho mục đích testing trên mô hình. Có hai cách thường sử dụng trong cross­validation là splitting và k­fold

* + - 1. Xây dựng mô hình phân lớp

Mục đích của mô hình huấn luyện là tìm ra F(x) và thông qua hàm f tìm được để gán nhãn cho dữ liệu. Bước này thường được gọi là tranning.

F(x) = y.

Trong đó : x là các feature hay input đầu vào của dữ liệu

y là nhãn dán lớp hay output đầu ra.

Thông thường để xây dựng mô hình phân lớp cho bài toán này chúng ta sử dụng các thuật toán học giám sát như: K – nearest neighbor, Support Vector Machine, Decision tree, Navie Bayers.

* + - 1. Kiểm tra dữ liệu với mô hình

Sau khi tìm được mô hình phân lớp ở bước xây dựng mô hình thì bước này ta sẽ đưa dữ liệu mới vào để kiểm tra trên mô hình phân lơp.

* + - 1. Đánh giá mô hình phân lớp và chọn mô hình tốt nhất

Bước cuối cùng chúng ta sẽ dánh giá mức độ lỗi của dữ liệu testing và dữ liệu training thông qua mô hình tìm được. Nếu không đạt được kết quả mong muốn chúng ta phải thay đổi các tham số của thuật toán để tìm ra các mô hình tốt hơn và kiểm tra, đánh giá lại mô hình phân lớp. Cuối cùng chọn ra mô hình phân lớp tốt nhất cho bài toán của chúng ta.

1. **Chương 2: Thuật toán K- Nearest Neighbor**
   1. **Thuật toán k-nearest neighbor**
      1. **Định nghĩa**

K-nearest neighbor là một trong những thuật toán học có giám sát đơn giản nhất trong Machine Learning . Ý tưởng của K- nearest Neighbor là tìm ra output của kiểu dữ liệu dựa trên thông tin của những dữ liệu training gần nó nhất.

* + 1. **Quy trình làm việc của thuật toán K – Nearest Neighbor**

Bước 1 : Xác định tham số k = số láng giềng gần nhất

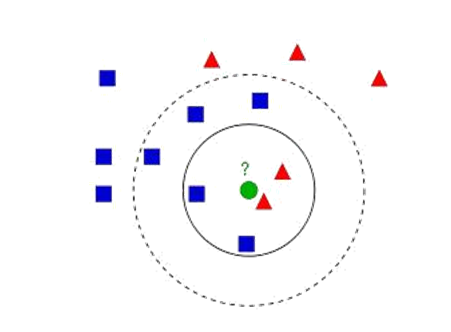
Bước 2: Tính khoảng cách đối tượng cần phân lớp với tất cả các đối tượng trong training data.

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 P 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.

* + 1. **Ví dụ minh họa**



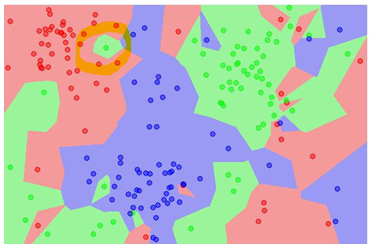
Giả sử bài toán được đặt ra: Mình mới quen một người bạn, tuy nhiên mình là fan của Rap việt, vậy nên mình cần biết người bạn mình là fan của king of rap hay là fan của rap việt. qua thời gian tìm hiểu mình đã thu thập được một số dữ liệu và đã biểu hiện dưới dạng hình vẽ trên

Ta dễ dàng nhận thấy có hai loại hình: hình vuông màu xanh biểu diễn cho những người là fan của king of rap, tam giác màu đỏ biểu diễn cho những người không là fan của King of rap, hình tròn mùa xanh là người bạn của mình mà mình muốn biết có phải là fan của king of rap hay không, khoảng cách giữa chấm tròn màu xanh và các hình còn lại biểu thị cho độ thân thiết của bạn đó và những ngươi khác.

Phương pháp đơn giản nhất là kiểm tra xem bạn đó có chơi thân với người bạn nào nhất, tức là tìm xem chấm xanh thuộc class nào (hình vuông hay là tam giác). Từ hình trên ta dễ dàng nhận thấy điểm gần chấm xanh nhất là hình tam giác màu đỏ, do đó sẽ được phân vào lớp tam giác màu đỏ.

Có một vấn đề ở phương pháp trên là xung quanh chấm xanh xuất hiện rất nhiều hình vuông màu xanh nên việc xét điểm gần nhất là chưa khả thi. Vì vậy ta sẽ xét K điểm gần nhất . Giả sử K = 3 , dựa theo hình ta dễ dàng nhận ra có hai hình tam giác đỏ và một hình vuông xanh có khoảng cách chấm xanh nhất, do đó châm xanh được phân vào lớp tam giác đỏ. Lấy K= 7 ta có năm hình vuông xanh và hai tam giác đỏ, lúc này chấp xanh được xếp vào class hình vuông xanh. Trường hợp K=4 ta thấy sẽ có hai hình vuông xanh và hai hình tam giác đỏ gần với chấm xanh nhất.

Trường hợp này K-Nearest Neighbor sẽ xử lý bằng cách so sánh tổng khoảng cách của hình gần nhất với điểm đang xét. Do xuất hiện điểm bằng nhau , vì vậy người ta thường dùng k số lẻ để xét, đó cũng là ý tưởng của K- Nearest Neighbor.

* + 1.  **Ví dụ về K – Nearest Neighbor nhiễu**

Hình trên là bài toán phân lpows với ba lớp: đỏ , lục, lam. Mỗi điểm dữ liệu mới sẽ được gán nhãn dựa theo màu của điểm đó thuộc về. Trong hình này, Vùng khoanh tròn mày vàng, ta nhận thấy rằng điểm mày lục nằm giữa gai vùng lớn với nhiều dữ liệu đỏ và lam , điểm này rất có thể là nhiễu dẫn đến việc dữ liệu test nếu rơi vào vùng này sẽ có khả năng sai lệch cao

* + 1. **Ưu điểm, nhược điểm của thuật toán**

**Ưu điểm :**

Dễ sử dụng và cài đặt

Việc dự đoán kết quả dữ liệu mới dễ dàng

Độ phức tạp tính toán nhỏ.

**Nhược điểm:**

K- Nearest Neighbor nhiễu sẽ đưa kết quả không chính xác

Cần thời gian lưu training set, khi dữ liệu training và test tăng lên nhiều mất nhiều thời gian tính toán.

* 1. **Khoảng cách trong không gian vector**

Trong không gian một chiều, việc đo khoả ng cách giữa hai điểm đã rất quen thuộc: lấy trị tuyệt đối của hiệu giữa hai giá trị đó. Trong không gian hai chiều, tức mặt phẳng, chúng ta thường dùng khoảng cách Euclid để đo khoảng cách giữa hai điểm

Việc đo khoảng cách giữa hia điểm dữ liệu nhiều chiều, tức hai vector, là rất cần thiết trong Machine Learning. Chúng ta cần đánh giá xem điểm nào là điểm gần nhất của một điểm khác; chúng ta cũng cần đánh giá xem độ chính xác của việc ước lượng và trong rất nhiều ví dụ khác . Đó là lý do mà khái niệm norm ra đơi. Có nhiều loại norm khác nhau

* + 1. **Định nghĩa**

Một số hàm f() ánh xạ một điểm x từ không gian n chiều sang tập số thực một chiều được gọi là norm nếu nó thỏa mãn ba điểu kiện sau:

F(x) >= 0 . dấu bằng chỉ xảy ra khi x=0

F(a,x) = |a|f(x),

F() +f() >= f(+ ),

* + 1. **Một số norm thường dùng**

Giả sử các vector X = [x1; x2; …..; xn], y = [y1; y2;….; yn]

Nhân thấy khoản cách Euclid chính là một norm, norm này thường gòi là norm 2:

||X||2 = (1)

Với p là một số nhỏ hơn 1 bất kỳ, hàm số sau đây:

||X||p =(|x1|p +|x­2|p +…+| xn |p)1/p  (2)

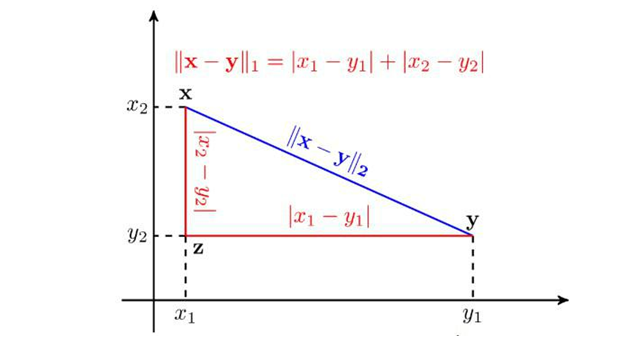
Được chứng minh thỏa mản điều kiện trên , và được gọi là norm p. nhận thấy rằng khi p 🡪0 thì biểu thức bên trên trờ thành phần tử khác 0 của x. hàm số (2) khi p=0 được gọi là giá trị chuẩn (Pseudo -norm) . Nó không phải là norm vù nó không thỏa mãn điều kiện 2 và 3 của norm. giá chuẩn này thường được ký hiệu là ||X||0, khá quan trọng trong machine learning vì nhiều bài toán ta cần “sparse” , một số lượng của x nhỏ, một vài giá trị thường được dùng

Khi P = 2 thường có norm 2

Khi P = 1 chú ta có

||x||1 = |x1| + |x2| +|x3|+ …+|xn| (3)

Là tổng các giá trị tuyệt đối của từng phần tử của x. norm 1 thường được dùng như sấp xỉ norm 0 trong các bài toán có ràng buộc. Dưới đây là một ví dụ so sánh norm 1 và norm 2 trong không gian 2 chiều:



Norm 2 (màu xanh) chính là đường chim bay nối giữ vector x và vector y. khoảng cách norm 1 giữa hai điểm (màu đỏ) có thể diễn giải như là đường từ x đến y trong một thành phố mà thành phố được tạo hình bàn cờ, chúng ta chỉ có thể đi dọc theo bàng cờ chứ không thể đi theo đường thẳng.

Khi p 🡪 vô cùng , ta có norm p chính là giá trị tuyệt đối của phần tử lớn nhất vủa vector đó:

 (4)

1. **Chương 3: Thử nghiệm**
2. **Chương 4: Kết luận**