**TIỂU LUẬN MÔN CHUYÊN ĐỀ - HK1 2021-2022**

Lớp IT…..

Họ Tên: Nguyễn Thị Diễm My

MSSV: 1851050091

Lớp IT…..

**Câu 1 (2.5đ): Hãy phân tích và so sánh thuật toán k-NN và K-means một cách chi tiết.**

……. Trình bày…….

k-NN (k-Nearest Neighbors):

* Đặc điểm:
* Là một trong những thuật toán học máy (Machine Learning) đơn giản nhất dựa trên kỹ thuật học có giám sát (Supervised Learning)
* Giả định sự giống nhau giữa trường hợp mới và những trường hợp có sẵn và đặt trường hợp mới vào danh mục gần giống nhất vói danh mục có sẵn
* Lưu trữ tất cả dữ liệu có sẵn và phân loại một điểm dữ liệu mới dựa trên sự tương đồng
* Thường sử dụng cho các bài toán phân loại (Classification) và hồi quy (Regression)
* Không đưa ra bất kỳ giả định nào về dữ liệu cơ bản
* Không học từ tập huấn luyện (Training) ngay mà chỉ lưu trữ tập dữ liệu, khi nhận được tập dữ liệu mới thì sẽ phân loại thành một loại gần giống
* Các bước thực hiện:
* Bước 1: Chọn số lượng k lân cận
* Bước 2: Tính toán khoảng cách Euclidean giữa các điểm dữ liệu, cụ thể là giữa điểm dữ liệu Test với các điểm dữ liệu Training
* Bước 3: Chọn k điểm dữ liệu Training gần nhất theo khoảng cách Euclidean được tính
* Bước 4: Trong số k lân cận đó, đếm số điểm dữ liệu của mỗi lớp
* Bước 5: Chọn lớp có nhiều điểm dữ liệu điểm
* Ưu điểm:
* Thực hiện đơn giản
* Thích hợp và hiệu quả hơn với dữ liệu Training lớn
* Nhược điểm:
* Luôn luôn cần xác định giá trị của k
* Phải tính toán khoảng cách nhiều

K-means:

* Đặt điểm:
* Là một trong những thuật toán Machine Learning cơ bản nhất trong học không giám sát (Unsupervised Learning)
* Không biết nhãn (Label) của từng điểm dữ liệu
* Phân dữ liệu thành các cụm (Cluster) khác nhau sao cho dữ liệu trong cùng một cụm có tính chất giống nhau
* Các bước thực hiện:
* Bước 1: Chọn ngẫu nhiên K tâm (Centroid) cho K Cluster, mỗi Cluster được đại diện bằng các Centroid của Cluster
* Bước 2: Dùng Euclidean tính khoảng cách giữa các đối tượng (Objects) đến K Centroid
* Bước 3: Gán cho mỗi Object vào các Cluster gần nhất
* Bước 4: Tính tập hợp các điểm mới dựa vào Centroid của từng Cluster theo sự phân chia hiện tại
* Bước 5: Trở lại bước 3, thuật toán kết thúc khi không có sự thay đổi về Cluster
* Ưu điểm:
* Đơn giản, dễ hiểu, tương đối hiệu quả
* Thường được tối ưu cục bộ
* Nhược điểm:
* Phức tạp tuyến tính
* Bị ảnh hưởng nhiều bởi số lượng Cluster, độ tương đồng, loại dữ liệu

Bảng so sánh k-NN và K-means

|  |  |
| --- | --- |
| k-NN | K-means |
| Học có giám sát | Học không giám sát |
| Phân loại, hồi quy | Phân cụm |
| k trong k-NN là số lân cận gần nhất | K trong K-means là số nhóm |
| Tính k điểm dữ liệu gần nhất từ điểm dữ liệu X, sử dụng những điểm này để xác định X thuộc lớp nào | Sử dụng khoảng cách từ điểm dữ liệu đến K tâm để phân cụm dữ liệu thành K nhóm |
| Điểm dữ liệu được phân loại vẫn là giống nhau | Cập nhật tâm mỗi lần trôi qua tính toán trên tất cả dữ liệu trong một lớp |
| Chỉ yêu cầu k tính toán khoảng cách | Phải lặp lại dữ liệu cho đến trung tâm điểm không di chuyển |
| Phải cung cấp nhãn dữ liệu | Không biết nhãn của từng điểm dữ liệu |

…………………….

**Câu 2(2.5đ):**

**a. Feature selection (lựa chọn đặc trưng) đóng vai trò như thế nào trong các bài toán phân loại có giám sát ?**

**b. FS có phải là một bài toán của machine learning hay không ?**

**c. Trình bày một hệ số lựa chọn đặc trưng cho trường hợp Unsupervised hoặc Supervised.**

……. Trình bày…….

1. Vai trò Feature selection trong các bài toán phân loại có giám sát là:

* Xác định các tính năng liên quan để đạt được mục tiêu tốt nhất của mô hình
* Cải thiện hiệu suất: tốc độ, sức mạnh dự đoán, tính đơn giản của mô hình
* Trực quan hóa dữ liệu để lựa chọn mô hình
* Giảm kích thước và loại bỏ tiếng ồn
* Loại bỏ dữ liệu không liên quan
* Tăng độ chính xác dự đoán của các mô hình đã học
* Giảm chi phí dữ liệu
* Nâng cao hiệu quả học tập: giảm yêu cầu lưu trữ, chi phí tính toán
* Giảm độ phức tạp của mô hình kết quả mô tả, nâng cao hiểu biết về dữ liệu và mô hình

1. Feature selection là quá trình giảm số lượng biến đầu vào khi phát triển mô hình dự đoán. Feature seclection không phải là một bài toán của Machine Learning mà là một nhánh trong Machine Learning và phụ thuộc vào ba ngữ cảnh: Supervised, Unsupervised và Semi-Supervised.

Feature selection:

* Dựa vào tính sẵn có của dữ liệu được gán nhãn
* Cho phép thuật toán Machine Learning train nhanh hơn
* Làm giảm độ phức tạp của một mô hình và làm cho mô hình đó dễ hiểu hơn
* Cải thiện độ chính xác của một mô hình nếu tập hợp con phù hợp được chọn

1. Một hệ số lựa chọn đặc trưng cho trường hợp Supervised: Fisher score

* Fisher score là một trong những phương pháp Supervised feature selection được sử dụng rộng rãi nhất
* Nhãn của dữ liệu được sử dụng
* Gọi biểu thị tất cả các trường hợp trên feature, là giá trị trung bình và là bình phương của lớp tương ứng với feature. Cụ thể, cho tới sẽ chạy từ các lớp tương ứng và là số lượng phần tử tương ứng của từng lớp. Chia nhỏ ra từng lớp tương ứng là , là bộ đặc trưng của cột đầu tiên. là giá trị trung bình của variance class của lớp thứ tương ứng

…………………….

**Câu 3(2.5đ): Trình bày các hệ số ràng buộc trong trường hợp semi-supervised. Có bao nhiêu trường hợp nhãn dữ liệu có thể cung cấp trong trường hợp này. Cho ví dụ minh họa cụ thể dựa trên dữ liệu Wine.**

……. Trình bày…….

Có 2 hệ số ràng buộc trong trường hợp semi-supervised: và

:

* Được tính là tổng của các cặp đặc trưng nằm trong nhóm chia cho tổng các cặp đặc trưng nằm trong nhóm

:

* Được tính là tổng của các cặp đặc trưng nằm trong nhóm trừ cho tổng các cặp đặc trưng nằm trong nhóm theo một hệ số

Có 2 trường hợp nhãn dữ liệu có thể cung cấp: (must – link) và (cannot -link)

Ví dụ minh họa: Dữ liệu Wine có 178 bình chia ra 3 lớp. Lớp 1 có 59 bình (2->60), lớp 2 có 71 bình (61->131) và lớp 3 có 48 bình(132->179). Trong trường hợp (must – link) gồm các cặp (5, 10) (70, 80) (135, 140),…Trong trường hợp (cannot -link) gồm các cặp (4, 78) (4, 178) (25, 168)…

…………………….

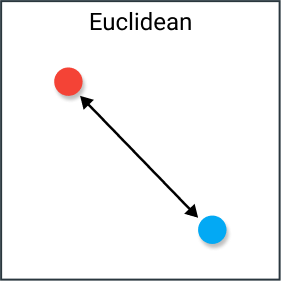
**Câu 4 (2.5đ): Trình bày ít nhất 5 độ đo khoảng cách khác nhau một cách chi tiết.**

……..Trình bày…….

Các độ đo khoảng cách đóng một vai trò quan trọng trong học máy. Đó là một điểm số khách quan tóm tắt sự khác biệt tương đối giữa hai đối tượng trong một miền vấn đề. 5 độ đo khoảng cách thường được sử dụng: Euclidean, Matnhattan, Cosine

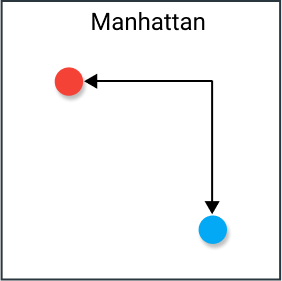
1. Euclidean:





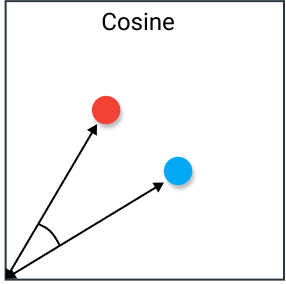
* Là phương pháp phổ biến nhất
* Là một thước đó khoảng cách tốt nhất, thường được giải thích là độ dài của một đoạn nối 2 điểm
* Hoạt động hiệu quả khi có dữ liệu chiều thấp và độ lớn của vectơ
* Trực quan để sử dụng, đơn giản để thực hiện và cho kết quả tuyệt vời trong nhiều trường hợp sử dụng
* Có thể bị ảnh hưởng bởi đơn vị của feature nên phải normalize trước khi tính toán
* Khi số chiều vectơ tăng thì trở nên kém hiệu quả

1. Manhattan:



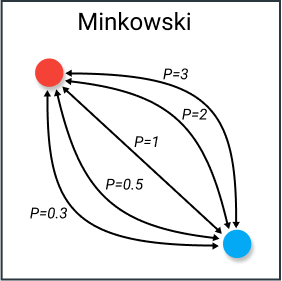
* Hay còn gọi là City Block
* Tính toán khoảng cách giữa các vectơ có giá trị thực
* Giá trị về độ lớn của các vectơ được tận dụng
* Hữu ích hơn với các vectơ mô tả các đối tượng trên một lưới đồng nhất
* Được sử dụng như so sánh string
* Khá phù hợp với dữ liệu chiều cao
* Kém trực quan hơn Euclidean
* Có nhiều khả năng cho một giá trị khoảng cách cao hơn

1. Cosine



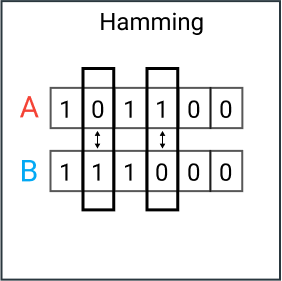
* Thường được sử dụng để giải quyết vấn đề của khoảng cách Euclidean ở không gian nhiều chiều
* Tính góc tạo thành giữa hai vectơ
* Không tận dụng được độ lớn của vectơ, chỉ tính theo hướng
* Làm mất mát thông tin so sánh

1. Minkowski



* Là một biện pháp khá phức tạp
* Là một số liệu được sử dụng trong không gian vectơ định mức mà khoảng cách có thể biểu diễn dưới dạng vectơ độ dài
* Có ba yêu cầu: vectơ zero – vectơ, hệ số vô hướng, bất đẳng thức tam giác
* Thước đo khoảng cách này sử dụng tham số p để điều khiển chỉ số khoảng cách để gần giống với những người khác
* Có thể kém hiệu quả về mặt tính toán trong một số trường hợp sử dụng

1. Hamming



* Là số giá trị khác nhau giữa hai vectơ
* Sử dụng để so sánh hai chuỗi nhị phân có độ dài bằng nhau
* Sử dụng cho các chuỗi để so sánh để so sánh mức độ tương tự bằng cách tính số ký tự khác nhau
* Có thể được sử dụng để xác định số lượng bit bị bóp méo trong một từ nhị phân
* Khó sử dụng khi hai vectơ không có độ dài bằng nhau
* Không tính đến giá trị thực tế

…………………….

**Yêu cầu:**

* Tất cả text phải tự gõ, không chấp nhận copy/paste hình công thức (vi phạm 0 điểm)
* SV tự làm cá nhân, không sao chép (các bài giống nhau 0 điểm)
* Nội dung viết đầy đủ, vừa phải (không quá ngắn cũng như quá dài)