**Họ và tên: Phạm Văn Hiệu**

**Lớp: 22CNTT2**

**MỤC LỤC**

***Trang***

BÀI TẬP SỐ 1 4

PHẦN I. LÝ THUYẾT 4

I. Khai phá dữ liệu 4

1. Khái niệm khai phá dữ liệu. 4

2. Các lĩnh vực liên quan tới Khai phá dữ liệu: giải thích sự liên hệ các lĩnh vực này có liên quan như thế nào tới Khai phá dữ liệu 4

3. Các nhiệm vụ của Khai phá dữ liệu, nêu tóm tắt từng nhiệm vụ 4

4. Các ứng dụng thực tế của Khai phá dữ liệu: kể tên ít nhất 3 ứng dụng, giải thích sự ứng dụng này. 5

5. Vai trò của Khai phá dữ liệu trong nền công nghiệp 4.0 6

6. Phân biệt được Khai phá dữ liệu với tìm kiếm thông thường. Cho ví dụ 7

II. Quy trình khai phá dữ liệu 8

1. Vẽ mô hình tổng quan Quy trình Khai phá dữ liệu 8

2. Trình bày nhiệm vụ của mỗi bước trong Quy trình khai phá dữ liệu. Cho ví dụ minh họa tại mỗi bước. 8

3. Tại bước tiền xử lí dữ liệu, làm rõ nhiệm vụ của tiền xử lí dữ liệu giải quyết những vấn đề gì? Trình bày cụ thể cho vấn đề feature extraction gồm Feature selection và Feature reduction. Khái niệm Feature selection và Feature reduction. Sự khác nháy giữa hai cơ chế này. Cho ví dụ minh họa. Trình bày thuật toán SVD/PCA cho feature reduction. Định nghĩa và nhiệm vụ của SVC/PCA là gì? Cho ví dụ minh họa cụ thể để giảm số chiều dữ liệu như thế nào cho thuật toán PCA? 9

III. Kho dữ liệu 11

1. Khái niệm Kho dữ liệu. 11

2. Phân biệt được Kho dữ liệu với Cơ sở dữ liệu tác nghiệp. Cho ví dụ 11

3. Trình bày các đặc tính của Kho dữ liệu 13

4. Phát biểu khái niệm OLAP. Phân biệt OLTP và OLAP. Cho ví dụ 13

5. Trình bày kiến trúc của Kho dữ liệu 14

6. Trình bày quy trình xây dựng Kho dữ liệu 15

7. Phát biểu khái niệm mô hình hóa Kho dữ liệu. Phân biệt được sơ đồ ngôi sao, bông tuyết và chòm sao. Cho ví dụ minh họa. 15

IV. Học máy 18

1. Khái niệm học máy 18

2. Các cơ chế học máy 18

3. Sự liên hệ của Học máy với Khai phá dữ liệu 18

4. Ứng dụng của Học máy 19

V. Học có giám sát (Supervised learning) 19

1. Khái niệm 19

2. Các ứng dụng thực tiễn cho cơ chế học có giám sát, giải thích. 19

VI. Phân lớp (Classification) 20

1. Khái niệm 20

2. Quy trình tổng quan cho việc phân lớp. Giải thích từng bước. Vẽ mô hình tổng quan cho quy trình Classification. 20

3. Giải thích rõ các thuật ngữ liên quan tới bài toán Phân lớp: Instance/Sample, Label, Training dataset, Testing dataset, Feature, Feature extraction, Feature Selection, Feature Reduction, Ground truth, Validation, visualization. 21

4. Các phương pháp đánh giá độ chính xác của mô hình: khái niệm, giải thích các tham số, cách thực hiện, cho ví dụ minh họa. (Note: Trình bày các phương pháp, tuy nhiên bắt buộc nắm rõ và phải trình bày được phương pháp confusion matrix). 22

5. Cho ví dụ và vẽ mô hình tổng quan cho bài toán ví dụ gồm: 23

VII. Thuật toán K-NN 24

1. Khái niệm/ Ý tưởng 24

2. Giải thích các tham số 24

3. Quy trình thực hiện Phân lớp dựa trên thuật toán KNN 24

4. Viết mã giả 25

5. Ví dụ minh họa: Có thể lấy bài tập ở Phần II để thực hiện phần này. 25

PHẦN II: THỰC HÀNH 32

Ex1: Cho một tập dữ liệu 32

Ex2: Given a collection dataset as following: Using KNN method 34

# BÀI TẬP SỐ 1

# PHẦN I. LÝ THUYẾT

# Khai phá dữ liệu

## Khái niệm khai phá dữ liệu.

Khai phá dữ liệu là quá trình tìm kiếm và khám phá thông tin tiềm ẩn, mẫu và quy luật từ tập dữ liệu lớn Các nguồn dữ liệu có thể bao gồm cơ sở dữ liệu, kho dữ liệu, web, kho thông tin khác hoặc dữ liệu liên quan đến máy học ( machine learning).

## Các lĩnh vực liên quan tới Khai phá dữ liệu: giải thích sự liên hệ các lĩnh vực này có liên quan như thế nào tới Khai phá dữ liệu

+ Các lĩnh vực liên quan đến khai phá dữ liệu như tổ chức dữ liệu

* Trí tuệ nhân tạo: Giúp AI học từ dữ liệu để xây dựng các mô hình thông minh, đặc biệt là học máy, phụ thuộc vào khai phá dữ liệu để tìm ra các mẫu có ý nghĩa và cải thiện dự đoán.
* Học máy: Sử dụng các kỹ thuật khai phá dữ liệu để học các mô hình dự đoán từ dữ liệu. Tìm ra các biến số quan trọng, tạo mô hình và cải thiện kết quả.
* Thống kê: Khai phá dữ liệu và thống kê đều tập trung vào tìm hiểu và phân tích các mẫu. Các phương pháp thống kê sẽ đưa ra mức độ tin cậy và phân bố dữ liệu
* Cơ sở dữ liệu: Phụ thuộc vào cơ sở dữ liệu để thu thập thông tin và lưu trữ, quản lí khối lượng dữ liệu lớn. Giúp chúng ta truy xuất và tổ chức dữ liệu nhanh hơn.
* Kinh doanh thông minh: Cung cấp nền tảng giúp doanh nghiệp khai thác thông tin từ dữ liệu, hỗ trợ chiến lược.

## Các nhiệm vụ của Khai phá dữ liệu, nêu tóm tắt từng nhiệm vụ

Có thể phân thành 2 loại chính đó là Dự đoán (Predictive) và Mô tả (Descriptive)

Predictive:

* Classification - phân lớp: Phân loại đối tượng vào các nhóm hoặc lớp đã xác định từ trước dựa trên các đặc điểm của đối tượng.

VD: Phân loại email rác hoặc không.

* Regression - hồi quy: Dự đoán giá trị liên tục của một biến dựa trên các biến đầu vào. Giúp tìm ra mối quan hệ giữa các biến số.

VD: Dự đoán giá nhà, diện tích, vị trí …

* Deviation Detection - phát hiện độ lệch: Tìm các mẫu hoặc điểm dữ liệu khác thường trong tập dữ liệu.

VD: Phát hiện giao dịch thẻ tín dụng gian lận.

Descriptive:

* Clustering - phân cụm: Nhóm các đối tượng có các đối tượng tương tự nhau thành các cụm mà không cần biết trước về các nhóm cụ thể. Các đối tượng trong cùng một cụm có nhiều điểm chung hơn so với các đối tượng ở cụm khác.

VD: Phân nhóm khách hang thành các phân khúc thị trường dựa trên mua hàng.

* Association Rule Discovery - phát hiện luật kết hợp: Tìm ra các mẫu hoặc mối liên hệ giữa các mục trong dữ liệu.

VD: Các khách hàng mua bánh mì thường mua thêm sữa, đề xuất trưng bày 2 sản phẩm cạnh nhau.

## Các ứng dụng thực tế của Khai phá dữ liệu: kể tên ít nhất 3 ứng dụng, giải thích sự ứng dụng này.

* **Tiếp thị và Bán hàng (Marketing & Sales)**
* **Ứng dụng:** Khai phá dữ liệu giúp các doanh nghiệp phân tích hành vi mua sắm của khách hàng, từ đó phát hiện các mẫu và xu hướng tiềm năng.
* **Giải thích:** Doanh nghiệp có thể sử dụng các thuật toán khai phá dữ liệu để phân tích lịch sử mua hàng của khách hàng, phân loại khách hàng thành các nhóm dựa trên hành vi tiêu dùng. Điều này giúp họ phát triển các chiến lược marketing cá nhân hóa, quảng cáo chính xác hơn và tăng hiệu quả bán hàng.
  + - **Phát hiện gian lận (Fraud Detection)**
* **Ứng dụng:** Ngành tài chính sử dụng khai phá dữ liệu để phát hiện các hành vi gian lận trong các giao dịch tài chính như thanh toán thẻ tín dụng hoặc bảo hiểm.
* **Giải thích:** Các mô hình khai phá dữ liệu có thể học từ các dữ liệu giao dịch để nhận diện các hành vi bất thường có khả năng là gian lận. Ví dụ, khi một thẻ tín dụng có các giao dịch không thường xuyên hoặc ở những địa điểm khác thường, hệ thống có thể cảnh báo hoặc tự động ngăn chặn các giao dịch đó.
  + - **Y tế và Chẩn đoán (Healthcare & Diagnosis)**
* **Ứng dụng:** Khai phá dữ liệu được sử dụng để dự đoán và chẩn đoán bệnh dựa trên dữ liệu y tế của bệnh nhân, giúp phát hiện sớm các vấn đề sức khỏe và cải thiện quá trình điều trị.
* **Giải thích:** Các hệ thống khai phá dữ liệu y tế có thể phân tích dữ liệu bệnh án, xét nghiệm và chẩn đoán hình ảnh để tìm ra các mô hình hoặc yếu tố rủi ro cho các bệnh như ung thư, tiểu đường. Nhờ đó, các bác sĩ có thể có được những thông tin quan trọng để đưa ra quyết định điều trị chính xác và kịp thời.

## Vai trò của Khai phá dữ liệu trong nền công nghiệp 4.0

* + Tối ưu hoá quy trình
* Các quy trình sản xuất có thể được tối ưu hóa để tiết kiệm năng lượng, giảm chi phí và tăng sản lượng.
  + Phát triển sản phẩm
* Phân tích dữ liệu về thói quen và hành vi của khách hàng giúp doanh nghiệp phát triển các sản phẩm mới phù hợp với thị trường.
  + Tự động hoá và ra quyết định thông minh
* Giảm sự can thiệp của con người, tăng tốc độ và độ chính xác của các quyết định.
  + Dự đoán và bảo trì tiên đoán
* Giảm chi phí bảo trì, tránh hỏng hóc bất ngờ và tối ưu hóa vòng đời của thiết bị.
  + Phân tích dữ liệu lớn
* Giúp doanh nghiệp đưa ra các quyết định chiến lược dựa trên dữ liệu thực, cải thiện hiệu quả kinh doanh.
  + Quản lí và an ninh dữ liệu
* Tối ưu hóa lưu trữ và truy xuất dữ liệu một cách an toàn.

## Phân biệt được Khai phá dữ liệu với tìm kiếm thông thường. Cho ví dụ

|  |  |  |
| --- | --- | --- |
|  | Khai phá dữ liệu | Tìm kiếm thông thường |
| Mục tiêu | Tìm ra xu hướng, mô hình và quan hệ mà chưa được biết trước. | Truy xuất thông tin chính xác mà người dùng yêu cầu từ 1 nguồn dữ liệu đã được tổ chức. |
| Phương pháp | Thuật toán phức tạp, phân tích thống kê và các tính toán khác để phân tích và mô hình hoá dữ liệu | Dựa trên việc so khớp từ khoá hoặc truy vấn cụ thể với nội dung có sẵn. |
| Kết quả | Các phát hiện mới, mô hình hoặc các quy luật ẩn mà không rõ ràng. | Các tài liệu hoặc thông tin trực tiếp mà bạn tìm kiếm. |

VD: Trong cùng một việc tìm kiếm lịch sử phát triển của bánh mì thì:

Tìm kiếm thông thường: Chỉ đưa ra các lịch sử của bánh mì và không thêm thông tin gì khác.

Khai phá dữ liệu: Giúp phát hiện các cách sử dụng, nguồn gốc, số lượng tiêu thụ với các loại bánh mì qua từng giai đoạn.

# **Quy trình khai phá dữ liệu**

## Vẽ mô hình tổng quan Quy trình Khai phá dữ liệu

A diagram of a diagram

Description automatically generated

Dữ liệu thô - Trích chọn dữ liệu - Tiền xử lí - Biến đổi dữ liệu - Khai phá dữ liệu - Đánh giá và giải thích - Biểu diễn tri thức.

## Trình bày nhiệm vụ của mỗi bước trong Quy trình khai phá dữ liệu. Cho ví dụ minh họa tại mỗi bước.

Bước 1- Trích chọn dữ liệu (data selection): Là bước trích chọn những tập dữ liệu cần được khai phá từ các tập dữ liệu lớn (databases,data warehouses).

VD: Chọn dữ liệu mua hàng từ năm 2023 2024 bao gồm các cột thời gian, sản phẩm, hoá đơn, mã khách hàng.

Bước 2 - Tiền xử lý dữ liệu (data preprocessing): Là bước làm sạch dữ liệu (xử lý dữ liệu không đầy đủ, dữ liệu nhiễu, dữ liệu không nhất quán,...), rút gọn dữ liệu (sử dụng các phương pháp thu gọn dữ liệu, histograms, lấy mẫu,...), rời rạc hoá dữ liệu (dựa vào histograms, entropy, phân khoảng,…). Sau bước này, dữ liệu sẽ nhất quán, đầy đủ, được rút gọn và được rời rạc hoá.

VD: Loại bỏ các dữ liệu thiếu thông tin như hoá đơn không phù hợp.

Bước 3- Biến đổi dữ liệu (data transformation): Là bước chuẩn hoá và làm mịn dữ liệu để đưa dữ liệu về dạng thuận lợi nhất nhằm phục vụ cho các kỹ thuật khai phá ở bước sau.

VD: Chuyển hoá giá trị tiền tệ.

Bước 4- Khai phá dữ liệu (data mining): Đây là bước quan trọng và tốn nhiều thời gian nhất của quá trình khám phá tri thức, áp dụng các kỹ thuật khai phá (phần lớn là các kỹ thuật của machine learning) để khai phá, trích chọn được các mẫu (pattern) thông tin, các mối liên hệ đặc biệt trong dữ liệu.

VD: Áp dụng các thuật toán để tìm ra nhu cầu của khách hàng.

Bước 5- Đánh giá và biểu diễn tri thức (knowledge representation & evaluation): Dùng các kỹ thuật hiển thị dữ liệu để trình bày các mẫu thông tin (tri thức) và mối liên hệ đặc biệt trong dữ liệu đã được khai phá ở bước trên biểu diễn theo dạng gần gũi với người sử dụng như đồ thị, cây, bảng biểu, luật,… Đồng thời, bước này cũng đánh giá những tri thức khai phá được theo những tiêu chí nhất định.

VD: Biểu diễn các thói quen của khách hàng để dễ dàng quản lí.

## Tại bước tiền xử lí dữ liệu, làm rõ nhiệm vụ của tiền xử lí dữ liệu giải quyết những vấn đề gì? Trình bày cụ thể cho vấn đề feature extraction gồm Feature selection và Feature reduction. Khái niệm Feature selection và Feature reduction. Sự khác nháy giữa hai cơ chế này. Cho ví dụ minh họa. Trình bày thuật toán SVD/PCA cho feature reduction. Định nghĩa và nhiệm vụ của SVC/PCA là gì? Cho ví dụ minh họa cụ thể để giảm số chiều dữ liệu như thế nào cho thuật toán PCA?

* + - Tiền xử lý dữ liệu (data preprocessing): Là bước làm sạch dữ liệu (xử lý dữ liệu không đầy đủ, dữ liệu nhiễu, dữ liệu không nhất quán,...), rút gọn dữ liệu (sử dụng các phương pháp thu gọn dữ liệu, histograms, lấy mẫu,...), rời rạc hoá dữ liệu (dựa vào histograms, entropy, phân khoảng,…).
* Feature Reduction: Là quá trình biến đổi đặc trưng ban đầu thành các đặc trưng nhưng có kích thước nhỏ hơn, thường là tổ hợp của các đặc trưng cũ nhưng chứa nhiều thông tin tổng hợp.
  + - * VD: Biến ma trận 10x10 thành ma trận 4x4.
    - Feature Selection: Là một phương pháp giảm số lượng các biến thông tin đầu vào trong mô hình Machine Learning của bạn, bằng cách chỉ sử dụng những dữ liệu liên quan, có ý nghĩa và loại bỏ các dữ liệu nhiễu
* VD: Có thể loại bỏ đi thuộc tính của các hoá đơn mua hàng như về độ tuổi, chữ kí, địa chỉ…
  + - Sự khác nhau giữa Feature Selection và Feature Reduction
* Selection chỉ làm gọn những dữ liệu đã có sẵn và không tạo ra các dữ liệu mới.
* Reduction thì làm biến đổi đặc trưng ban đầu thành nhỏ hơn và trong đó các dữ liệu thuộc tính mới là tổ hợp của các thuộc tính cũ.
  + - PCA
* Là một kĩ thuật giảm chiều dữ liệu bằng cách biến đổi các đặc trưng gốc thành đặc trưng mới, tuy số chiều của tập dữ liệu nhỏ hơn nhưng giữ phần lớn thông tin.
* Nhiệm vụ: Giảm chiều dữ liệu không làm mất nhiều thông tin, giúp giảm độ phức tạp của mô hình tính toán.
* VD: Có tập dữ liệu 3D thì PCA sẽ giảm xuống còn 2D nhưng vẫn giữ được thông tin quan trọng.
* VD cụ thể trong bộ dữ liệu Iris gồm 150 mẫu và 4 thuộc tính.

1.Độ dài cánh hoa

2.Độ rộng cánh hoa

3. Độ dài đài hoa

4.Độ rộng đài hoa

PCA sẽ làm giảm số chiều của dữ liệu từ 4D xuống 2 hoặc 3D mà vẫn giữ lại phần lớn thông tin.

Bước 1: Chuẩn hoá dữ liệu để thuộc tính có trung bình bằng 0 và phương sai bằng

Bước 2: TÍnh toán ma trận hiệp phương sai giữa các thuộc tính

Bước 3: Tính các giá trị riêng và vector riêng từ ma trận hiệp phương sai.

Bước 4: Chọn các thành phần chính chứa phần lớn thông tin.

Bước 5: Chiếu dữ liệu ban đầu.

Bước 6: Trực quan hoá dữ liệu.

Chúng ta sẽ nhận được kết quả số chiều giảm từ 4 xuống 2 mà vẫn giữ lại phần lớn thông tin.

* SVD
  + - * Phân rã giá trị suy biến là một kỹ thuật ma trận phân rã giúp phân tích và giảm chiều dữ liệu bằng cách tạo ra các ma trận đơn giản.
      * Nhiệm vụ: Phân rã ma trận dữ liệu ban đầu thành các ma trận đơn giản hơn, từ đó dễ dàng giảm số chiều dựa trên các giá trị riêng.
      * VD: Nếu có ma trận 4x3 gồm 4 thành phần thì SVD sẽ giảm thành các ma trận nhỏ hơn nhưng không làm mất thông tin quan trọng.

# Kho dữ liệu

## Khái niệm Kho dữ liệu.

+ Là một kho lưu trữ thông tin trung tâm, có thể được phân tích để đưa ra các lựa chọn tốt hơn.

+ Là một hệ thống lưu trữ dữ liệu từ nhiều nguồn, nhiều môi trường khác nhau như phần mềm bán hàng, kế toán hay hệ thống lõi ngân hàng.

## Phân biệt được Kho dữ liệu với Cơ sở dữ liệu tác nghiệp. Cho ví dụ

|  |  |  |
| --- | --- | --- |
|  | Cơ sở dữ liệu | Kho dữ liệu |
| Mục đích | Ghi và truy vấn dữ liệu | Xử lí và phân tích dữ liệu |
| Chức năng | Hỗ trợ các hoạt động hàng ngày | Hỗ trợ quyết định mang tính chiến lược |
| Phương pháp xử lí | Cơ sở dữ liệu sử dụng xử lí giao dịch trực tuyến(OLTP) | Kho dữ liệu sử dụng xử lí phân tích trực tuyến(OLAP) |
| Các bảng và phép nối | Có độ phức tạp cao vì chúng được chuẩn hóa (cho RDMS) để giảm dữ liệu thừa, tối ưu hóa dung lượng lưu trữ | Bảng và phép nối rất dễ dàng trong kho dữ liệu vì chúng không được chuẩn hóa. |
| Tính chất dữ liệu | Chi tiết, được cập nhật thường xuyên | Có tính lịch sử và thống kê, được thêm mới chứ không cập nhật |
| Lưu trữ dữ liệu | Phương pháp tiếp cận quan hệ phẳng, nhiều dữ liệu khác nhau được tích hợp vào một nguồn | Phương pháp tiếp cận đa chiều và chuẩn hóa, nhiều nguồn dữ liệu khác nhau được tích hợp và định dạng lại |
| Sử dụng | Thường xuyên | Những trường hợp đặc biệt |
| Đơn vị công việc | Giao dịch đơn giản, ngắn | Các câu truy vấn phức tạp |
| Độ đo | Thông lượn giao dịch, có thể thực hiện nhiều giao dịch cùng lúc | Thông lượng truy vấn và trả lời. |
| Sự định hướng | Định hướng ứng dụng | Định hướng chủ đề. |
| Mô hình sử dụng | Mô hình quan hệ – thực thể | Mô hình dữ liệu đa chiều. |
| Loại truy vấn | Những truy vấn giao dịch đơn giản được sử dụng. | Những truy vấn phức tạp được áp dụng cho mục đích phân tích. |
| Hiệu suất truy vấn phân tích | Thấp | Cao |
| Ví dụ | Lưu lại tình hình buồn bán, doanh thu của cửa hàng. | Phân tích tình hình kinh doanh của cửa hàng trong nửa năm hoặc trong vòng một quý. |

## Trình bày các đặc tính của Kho dữ liệu

Kho dữ liệu có 4 đặc tính quan trọng

* + - Hướng chủ đề( Subject-oriented): Kho dữ liệu được tổ chức theo chủ đề, tập trung vào các vấn đề và quá trình kinh doanh cụ thể của doanh nghiệp. Điều này giúp truy xuất thông tin nhanh chóng và dễ dàng hơn.
    - Nguồn đa dạng(Integrated): Tích hợp nhiều nguồn khác nhau, đảm bảo tính nhất quán và đúng đắn của dữ liệu.
    - Dữ liệu được gán thời gian(Time variant): Chứa các dữ liệu lịch sử và cập nhật, giúp cho người dùng phân tích dữ liệu theo thời gian, tìm các xu hướng và quy luật phát triển.
    - Dữ liệu không biến đổi(Non-volatile): Dữ liệu trong kho không bị thay đổi hay xoá bỏ, ngoại trừ khi bảo trì và cập nhât. Điều này giúp kho dữ liệu đáng tin cậy hơn về phần bảo mật.

## Phát biểu khái niệm OLAP. Phân biệt OLTP và OLAP. Cho ví dụ

OLAP: Hay còn được biết đến là công nghệ phần mềm mà bạn có thể sử dụng để phân tích dữ liệu kinh doanh từ các quan điểm khác nhau. Các tổ chức sẽ thu thập và lưu trữ dữ liệu từ nhiều nguồn dữ liệu, chẳng hạn như trang web, ứng dụng, và hệ thống nội bộ.

|  |  |  |
| --- | --- | --- |
| Tiêu chí | OLAP | OLTP |
| Mục đích | Giúp bạn phân tích khối lượng dữ liệu lớn để hỗ trợ quá trình ra quyết định | Giúp bạn quản lí và xử lí các giao dịch thời gian thực. |
| Nguồn dữ liệu | Sử dụng dữ liệu lịch sử và dữ liệu tổng hợp từ nhiều nguồn | Sử dụng dữ liệu thời gian thực và dữ liệu giao dịch từ một nguồn duy nhất. |
| Cấu trúc dữ liệu | Sử dụng cơ sử dự liệu đa chiều hoặc quan hệ | Sử dụng cơ sở dữ liệu quan hệ |
| Mô hình dữ liệu | Sử dụng lược đồ ngôi sao, bông tuyết, hoặc các mô hình phân tích khác | Sử dụng các mô hình chuẩn hoá hoặc phi chuẩn hoá |
| Khối lượng dữ liệu | Yêu cầu lưu trữ lớn.(TB,PB) | Yêu cầu lưu trữ tương đối nhỏ(GB) |
| Thời gian phản hồi | Thời gian phản hồi dài hơn, tính bằng giây hoặc phút | Tính bằng mili giây |
| Ví dụ | Phân tích xu hướng, dự đoán hành vi của khách hàng và xác định khả năng sinh lời | Phù hợp với xử lý thanh toán, quản lí dữ liệu khách hàng và xử lí đơn hàng. |

## Trình bày kiến trúc của Kho dữ liệu

Kiến trúc kho dữ liệu gồm 3 phần chính

* Kiến trúc một tầng: Bao gồm một lớp lưu trữ trực tiêp trong một cơ sở dữ liệu trung tâm, giúp truy cập dữ liệu nhanh chóng nhưng không tối ưu cho các hệ thống lớn.
* Kiến trúc hai tầng: Bao gồm một lớp lưu trữ dữ liệu và một lớp truy vấn dữ liệu. Đây là cách tiếp cận đơn giản nhưng không hỗ trợ mở rộng quá tốt cho hệ phân tán.
* Kiến trúc ba tầng: Kiến trúc phổ biến nhất, chia hệ thống thành 3 thầng chính để dễ dàng quản lí, bảo trì và mở rộng.

## Trình bày quy trình xây dựng Kho dữ liệu

* Phân tích nhu cầu: Các doanh nghiệp cần phải phân tích nhu cầu và yêu cầu của người dùng, bao gồm các yêu cầu về dữ liệu, truy vấn dữ liệu và báo cáo.
* Lập kế hoạch: Xác định mục tiêu, phạm vi, nguồn dữ liệu, cấu trúc, công nghệ, ngân sách và thời gian cần thiết cho việc xây dựng kho dữ liệu.
* Thiết kế kiến trúc kho dữ liệu: Bao gồm các lớp dữ liệu, bảng dữ liệu, khoảng thời gian lưu trữ, quy trình đồng bộ và quy trình xử lí, quản lí.
* Thu thập và chuẩn hoá dữ liệu: Thu nhập từ nhiều nguồn khác nhau và được xử lí trùng lập, kiếm tra và đưa về dạng chuẩn trước khi lưu trữ.
* Xây dung ETL(Extract, Transform, Load): Tích hợp dữ liệu từ các nguồn dữ liệu khác nhau vào kho dữ liệu.
  + Extract: xác định và trích xuất các dữ liệu cần thiết.
  + Load: Tải các dữ liệu được trích xuất sẽ được lên các database xác định.
  + Transform: chuyển đổi các dữ liệu từ hình thức cũ lên hình thức mới.
* Phát triển các ứng dụng phân tích dữ liệu: Ứng dụng kho dữ liệu vào phân tích trong các doanh nghiệp.
* Triển khai và bảo trì.

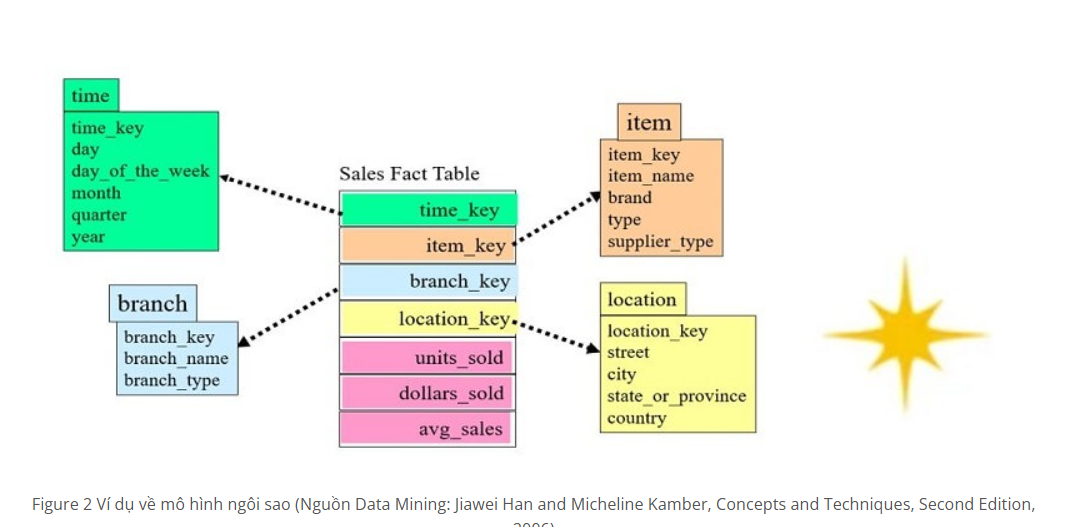
## Phát biểu khái niệm mô hình hóa Kho dữ liệu. Phân biệt được sơ đồ ngôi sao, bông tuyết và chòm sao. Cho ví dụ minh họa.

* Mô hình hoá kho dữ liệu là quá trình thiết kế các lược đồ thông tin chi tiết và tóm của kho dữ liệu. Đây kà một giai đoạn thiết yếu của việc xây dựng một kho dữ liệu vì 2 lí do chính.
  + Thứ nhất: Thông qua lược đồ, khách hàng của kho dữ liệu kho để sử dụng dễ dàng.
  + Thứ hai: Lược đồ thiết kế tốt cho phép xuất hiện cấu trúc kho dữ liệu hiệu quả, giảm chi phí phát triển và nâng cấp.

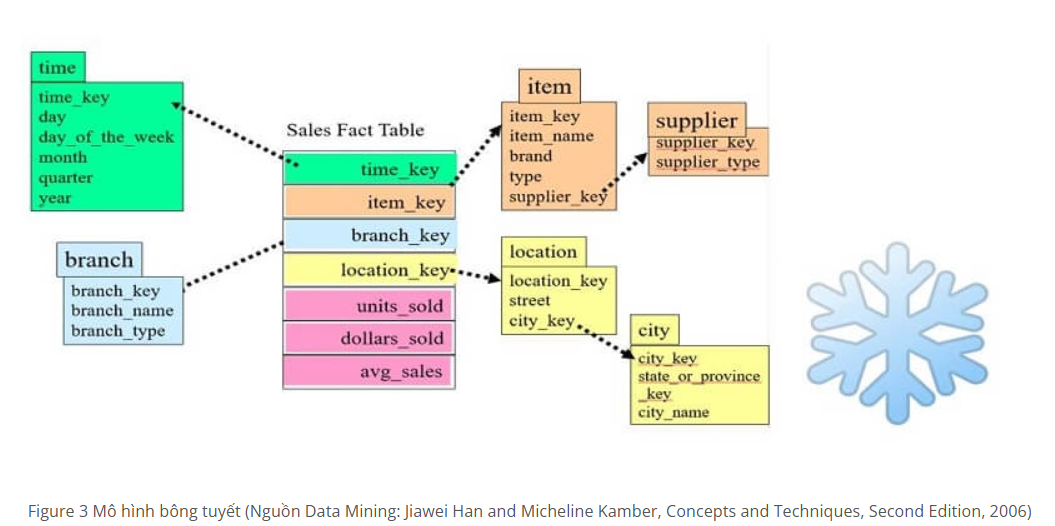
|  |  |  |  |
| --- | --- | --- | --- |
| Tiêu chí | Lược đồ bông tuyết | Lược đồ ngôi sao | Mô hình chòm sao |
| Cấu trúc | Bảng dữ liệu trung tâm được bao quanh bởi nhiều bảng chiều, một số được chia thành các bảng kích thước. | Bảng dữ kiện trung tâm được kết nối trực tiếp với một số bảng kích thước không chuẩn hoá. | Gồm nhiều bảng kết nối với các bảng chiều khác nhau. Một số bảng được dùng chung giữa các bảng. |
| Chuẩn hoá | Chuẩn hoá.  Các bảng chiều được chia thành nhiều bảng có liên quan. | Phi chuẩn hoá.  Chiều được hợp nhất thành các bảng đơn, rộng hơn. | Kết hợp cả chuẩn hoá và phi chuẩn hoá. Bảng sự kiện thường phi chuẩn hoá còn bảng chiều chuẩn hoá |
| Hiệu suất truy vấn | Chậm hơn so với ngôi sao | Nhanh vì cần ít liên kết. | Chậm hơn so với ngôi sao |
| Tính linh hoạt | Linh hoạt vì mô hình dữ liệu có thể dễ dàng được điều chỉnh hơn bằng cách sửa đổi các bảng kích thước. | Cứng nhắc hơn vì những thay đổi thường yêu cầu cập nhật bảng dữ liệu trung tâm. | Linh hoạt cao, dễ dàng mở rộng hệ thống bằng cách them các bảng sự kiện hoặc bảng chiều mà không ảnh hưởng nhiều đến cấu trúc hiện có. |
| Độ phức tạp | Phức tạp. | Đơn giản. | Cao |
| Kết luận | Nhấn mạnh vào tính chuẩn hoá và linh hoạt. | Ưu tiên hiệu suất truy vấn và tính đơn giản của mô hình dữ liệu. | Phù hợp với các hệ thống phức tạp và yêu cầu độ linh hoạt cao. |

**Ví dụ:**

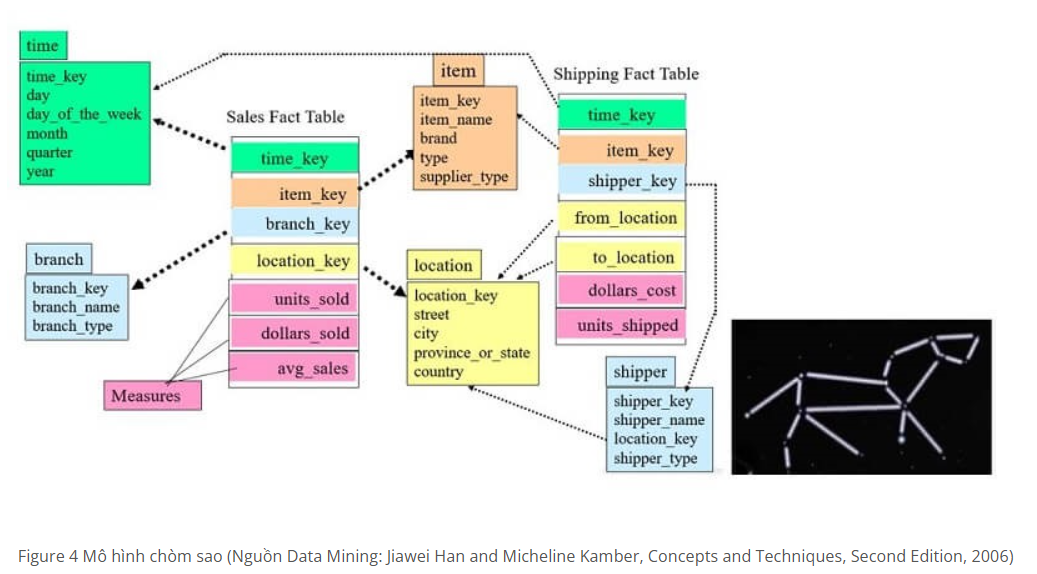
**Lược đồ ngôi sao**: Ở trong ví dụ trong Figure 2, bảng sự kiện chứa thông tin bán hàng với các độ đo: số lượng bán được (unit sold), số tiền thu được (dollars sold) và trung bình doanh thu (average sale). Bảng sự kiện này có liên kết với 4 bảng chiều: thời gian (time), chi nhánh cửa hàng (branch), mặt hàng (item) và vị trí cửa hàng (location).



**Lược đồ bông tuyết**: Figure 3 minh họa mô hình bông tuyết trong đó bảng location đã được chuẩn hóa, do thiếu không gian nên ở hình này chỉ vẽ được đến bảng city (trong đó có chứa khóa state\_or\_province\_key) để liên kết với bảng state\_or\_province, tiếp tục bảng state\_or\_province lại có quan hệ với bảng country (nếu muốn chuẩn hóa đến chuẩn 3).



**Mô hình chòm sao:** Trong ví dụ ở Figure 4, khi bổ sung thêm vào bảng sự kiện Giao hàng (shipping), khi đó bảng sự kiện này sẽ có nhu cầu sử dụng chiều location, item, time.



# Học máy

## Khái niệm học máy

Học máy (Machine Learning) là một lĩnh vực của trí tuệ nhân tạo, tập trung vào việc phát triển các thuật toán và mô hình cho phép máy tính học từ dữ liệu và cải thiện hiệu suất theo thời gian mà không cần lập trình rõ ràng.

## Các cơ chế học máy

Học máy có ba cơ chế chính:

* **Học có giám sát (Supervised Learning):** Máy học từ các dữ liệu đã được gán nhãn. Ví dụ: Dự đoán giá nhà dựa trên các đặc điểm như diện tích, số phòng.
* **Học không giám sát (Unsupervised Learning):** Máy học từ các dữ liệu chưa được gán nhãn. Ví dụ: Phân cụm khách hàng dựa trên hành vi mua sắm.
* **Học tăng cường (Reinforcement Learning):** Máy học thông qua việc thử và sai, nhận phản hồi từ môi trường để tối ưu hóa hành động. Ví dụ: Máy chơi cờ tự học cách chơi tốt hơn qua từng ván cờ.

## Sự liên hệ của Học máy với Khai phá dữ liệu

Học máy và khai phá dữ liệu (Data Mining) đều liên quan đến việc phân tích dữ liệu để tìm ra các mẫu và thông tin hữu ích. Tuy nhiên, học máy tập trung vào việc phát triển các mô hình dự đoán và cải thiện hiệu suất theo thời gian, trong khi khai phá dữ liệu chủ yếu tập trung vào việc khám phá và trích xuất thông tin từ dữ liệu.

## Ứng dụng của Học máy

Học máy có rất nhiều ứng dụng trong thực tế, bao gồm:

* **Y tế**: Chẩn đoán bệnh, phân tích hình ảnh y khoa.
* **Tài chính**: Dự đoán thị trường chứng khoán, phát hiện gian lận, đánh giá rủi ro, giao dịch tự động..
* **Thương mại** điện tử: Gợi ý sản phẩm, phân tích hành vi khách hàng.
* **Giao thông**: Xe tự lái, tối ưu hóa lộ trình.
* **Giải trí:** Gợi ý phim, âm nhạc dựa trên sở thích người dùng
* **Xử lý ngôn ngữ tự nhiên:** Dịch máy, chatbot, phân tích cảm xúc.
* **Xử lý hình ảnh:** Nhận dạng khuôn mặt, phân loại đối tượng, tạo ảnh.

# Học có giám sát (Supervised learning)

## Khái niệm

* Học có giám sát (Supervised Learning) là một phương pháp trong lĩnh vực học máy (Machine Learning), thuật toán nhằm dự đoán đầu ra của một dữ liệu mới trên các cặp (input, output) đã biết từ trước. Về cơ bản, dữ liệu đầu vào là dữ liệu đã qua đào tạo máy bằng cách sử dụng dữ liệu được gắn nhãn hoặc cho kết quả rõ ràng. Kỹ thuật này thường được áp dụng cho các bài toán phân lớp (Classification).

## Các ứng dụng thực tiễn cho cơ chế học có giám sát, giải thích.

**Dự đoán rủi ro**

* Học có giám sát có thể được dùng để dự đoán và đánh giá rủi ro trong các dịch vụ tài chính hoặc lĩnh vực bảo hiểm, nhằm mục đích giảm thiểu tối đa rủi ro của công ty.

**Nhận dạng và phân loại hình ảnh**

* Phân loại hình ảnh là một trong những trường hợp điển hình để minh họa cho kỹ thuật học máy có giám sát.

**Ví dụ :**

* Các hệ thống an ninh sử dụng học có giám sát để nhận diện khuôn mặt trong video hoặc hình ảnh.
* Các ứng dụng như Google Photos sử dụng học có giám sát để phân loại hình ảnh thành các danh mục như “con người”, “động vật”, "phong cảnh".

**Phát hiện gian lận**

* Được sử dụng để xác định xem các giao dịch được thực hiện bởi người dùng thực tế không, hay do hệ thống máy tính spam thực hiện. Điều này giúp các công ty tránh được gian lận.

**Ví dụ:**

* Các ngân hàng sử dụng học có giám sát để phát hiện các giao dịch gian lận dựa trên các mẫu giao dịch bất thường

**Nhận dạng trực quan**

* Mô hình học có giám sát sẽ giúp hệ thống có thể xác định các đối tượng, địa điểm, con người, hành động hoặc hình ảnh một cách trực quan.

**Ví dụ:**

* Các trợ lý ảo như Siri và Google Assistant sử dụng học có giám sát để chuyển đổi giọng nói thành văn bản và hiểu lệnh của người dùng.

# Phân lớp (Classification)

## Khái niệm

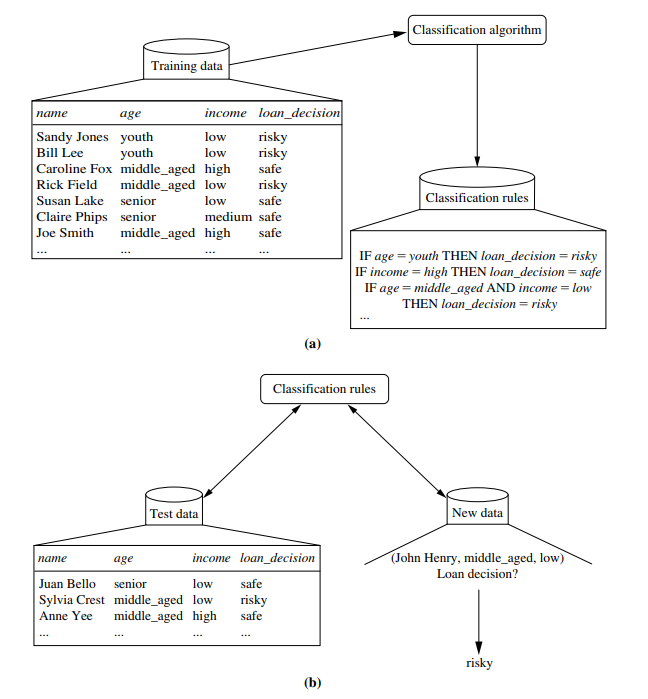
Phân lớp (classification): là tiến trình khám phá các luật phân loại hay đặc trưng cho các tập dữ liệu đã được xếp lớp. Tập dữ liệu học bao gồm tập đối tượng đã được xác định lớp sẽ được dùng để tạo mô hình phân lớp dựa trên đặc trưng của đối tượng trong tập dữ liệu học

## Quy trình tổng quan cho việc phân lớp. Giải thích từng bước. Vẽ mô hình tổng quan cho quy trình Classification.

gồm 2 bước :

+ Học tập : Bước này bao gồm việc thu thập và phân tích dữ liệu đầu vào. Bạn cần hiểu rõ về dữ liệu mà bạn đang làm việc, bao gồm các thông tin như đặc điểm dữ liệu, nguồn gốc …

+ Phân loại : Bước này liên quan đến việc áp dụng các thuật toán phân lớp vào dữ liệu đã chuẩn bị để đưa ra dự đoán. Các bước chính trong bước này bao gồm chọn mô hình, huấn luyện , dự đoán đánh giá



## Giải thích rõ các thuật ngữ liên quan tới bài toán Phân lớp: Instance/Sample, Label, Training dataset, Testing dataset, Feature, Feature extraction, Feature Selection, Feature Reduction, Ground truth, Validation, visualization.

Instance/Sample : là 1 đối tượng

Label : nhãn dán của dữ liệu

Training dataset : tập dữ liệu huấn luyện để xây dựng mô hình ( cây )

Testing dataset : dữ liệu kiểm tra đánh giá mô hình

Feature : đặc điểm đặc trưng của thuộc tính

Feature extraction : trích xuất các tính năng, các đặc trưng của dữ liệu

Feature Selection : chọn lọc các đặc trưng

Feature Reduction : giảm độ phức tạp

Ground truth : kết quả phân loại thực tế

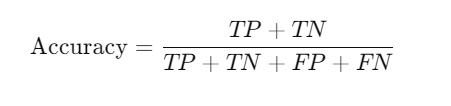
Validation : quá trình xác thực mô hình

Visualization : trực quan hóa dữ liệu

## Các phương pháp đánh giá độ chính xác của mô hình: khái niệm, giải thích các tham số, cách thực hiện, cho ví dụ minh họa. (Note: Trình bày các phương pháp, tuy nhiên bắt buộc nắm rõ và phải trình bày được phương pháp confusion matrix).

Classification Accuracy (Độ chính xác phân lớp):

* Khái niệm: Độ chính xác phân lớp là tỷ lệ các đối tượng được phân lớp chính xác trên tổng số đối tượng trong tập dữ liệu kiểm tra. Đây là một chỉ số đơn giản và trực quan, tuy nhiên có thể gây hiểu nhầm trong các tập dữ liệu không cân bằng, nơi mà lớp chiếm đa số có thể làm tăng giá trị của độ chính xác mà không phản ánh đúng hiệu suất của mô hình.
* Tham số: Không yêu cầu tham số riêng. Chỉ số này dựa trên tổng số dự đoán đúng (True Positives + True Negatives) chia cho tổng số đối tượng trong tập kiểm tra (True Positives + True Negatives + False Positives + False Negatives).
* Cách thực hiện:
* Tính tổng số dự đoán đúng (True Positives và True Negatives).
* Chia cho tổng số đối tượng trong tập kiểm tra để ra độ chính xác



2. Confusion Matrix (Ma trận nhầm lẫn):

* Khái niệm: Ma trận nhầm lẫn là một bảng hiển thị chi tiết các kết quả dự đoán của mô hình, bao gồm số lượng True Positive (TP), True Negative (TN), False Positive (FP), và False Negative (FN). Ma trận này cung cấp thông tin chi tiết về từng loại lỗi dự đoán mà mô hình gặp phải, từ đó tính toán các chỉ số khác như độ chính xác, precision, và recall.
* Tham số: Gồm các giá trị TP, TN, FP, FN cho mỗi lớp trong tập dữ liệu.
* Cách thực hiện:
* Xây dựng ma trận với các hàng và cột đại diện cho nhãn thực tế và nhãn dự đoán. Ví dụ, ma trận 2x2 cho bài toán phân lớp nhị phân sẽ có dạng như sau:

Dự đoán: Positive, Dự đoán: Negative

Thực tế: Positive, True Positive (TP), False Negative (FN)

Thực tế: Negative, False Positive (FP), True Negative (TN)

## Cho ví dụ và vẽ mô hình tổng quan cho bài toán ví dụ gồm:

5.1. Mô tả bài toán : phân biệt giới tính nam và nữ

5.2. Mô tả dữ liệu : tập dữ liệu chọn lọc các đặc trưng và gắn nhãn cho giới tính nam và nữ như :

Tóc : dài ( nữ ) , ngắn ( nam )

Chiều cao : trên m65 ( nam ) , dưới m65 ( nữ )

Khuôn mặt : mặt tròn ( nữ ) , mặt vuông ( nam )

5.3. Nhiệm vụ của bài toán phân lớp của ví dụ là làm gì?

xây dựng mô hình phân lớp để phân biệt giới tính nam và giới tính nữ

# Thuật toán K-NN

## Khái niệm/ Ý tưởng

* Thuật toán K-NN (K-Nearest Neighbors) là thuật toán phân loại thuộc nhóm phân loại có giám sát, thuật toán sẽ tính toán khoảng cách từ đối tượng đến tâm các cụm, tìm giá trị khoảng cách nhỏ nhất và gán đối tượng vào lớp tương ứng.
* Ý tưởng chính của K-NN là dự đoán nhãn cho một điểm dữ liệu mới dựa trên nhãn của K điểm gần nhất trong tập huấn luyện. Nó hoạt động theo nguyên tắc "gần mực thì đen, gần đèn thì rạng": dữ liệu mới sẽ có xu hướng thuộc về nhóm của những điểm gần nó nhất.

## Giải thích các tham số

Các tham số quan trọng trong thuật toán K-NN bao gồm:

* K: Số lượng điểm lân cận gần nhất cần xem xét. K ảnh hưởng trực tiếp đến kết quả của thuật toán. Nếu K quá nhỏ, thuật toán có thể bị ảnh hưởng bởi nhiễu (overfitting); nếu K quá lớn, nó có thể làm mất tính cục bộ của thuật toán (underfitting).
* Khoảng cách: Phương pháp đo khoảng cách giữa các điểm dữ liệu, thường là khoảng cách Euclidean, nhưng cũng có thể là khoảng cách Manhattan hoặc Cosine.
* Trọng số (Weights): Trọng số có thể được gán cho các điểm lân cận dựa trên khoảng cách của chúng đến điểm cần phân loại. Điểm gần hơn có thể được gán trọng số cao hơn.
* Tập dữ liệu huấn luyện: Tập dữ liệu chứa các điểm dữ liệu đã biết nhãn để so sánh.

## Quy trình thực hiện Phân lớp dựa trên thuật toán KNN

B1. Thu thập dữ liệu huấn luyện: Tập dữ liệu bao gồm các mẫu dữ liệu có nhãn đã biết.

B2. Xác định giá trị của K: Lựa chọn một giá trị cho tham số K.

B3. Tính toán khoảng cách: Tính khoảng cách giữa điểm cần phân loại với tất cả các điểm trong tập huấn luyện.

B4. Tìm K điểm lân cận nhất: Chọn ra K điểm có khoảng cách nhỏ nhất đến điểm cần phân loại.

B5. Bỏ phiếu nhãn: Nhãn của điểm cần phân loại được xác định bằng cách bỏ phiếu đa số từ nhãn của K điểm lân cận.

B6. Gán nhãn cho điểm cần phân loại: Gán nhãn mà có số phiếu cao nhất cho điểm dữ liệu mới.

## Viết mã giả

## Ví dụ minh họa: Có thể lấy bài tập ở Phần II để thực hiện phần này.

4.1. Mô tả bài toán:

* Bài toán yêu cầu giải quyết một vấn đề phân loại bằng phương pháp K-Nearest Neighbors (KNN). Đây là một bài toán phân loại trong học máy, nơi mà mục tiêu là xác định lớp của một mẫu chưa biết dựa trên các mẫu đã biết trong tập dữ liệu.
* Trong trường hợp này, ta có một tập dữ liệu chứa các đặc điểm (features) của các loài côn trùng và lớp (class) mà chúng thuộc về. Mỗi mẫu dữ liệu bao gồm chiều dài bụng và chiều dài râu của côn trùng, và lớp côn trùng tương ứng.
* Mục tiêu là sử dụng tập dữ liệu này để dự đoán lớp của một mẫu côn trùng mới có đặc điểm chiều dài bụng và chiều dài râu đã biết nhưng chưa rõ lớp.

4.2. Mô tả dữ liệu:

A screenshot of a graph

Description automatically generated

4.3. Thực hiện tính tay từng bước theo Quy trình được trình bày ở phần 2. Trình bày kết quả, phân tích và kết luận.

**Bước 1**: Dữ liệu cho ( data training)

|  |  |  |
| --- | --- | --- |
| **X1** | **X2** | **Class** |
| 4 | 3 | + |
| 3 | 7 | + |
| 7 | 4 | + |
| 4 | 1 | + |
| 6 | 5 | + |
| 5 | 6 | + |
| 3 | 7 | + |
| 6 | 2 | + |
| 4 | 6 | - |
| 4 | 4 | - |
| 5 | 8 | - |
| 7 | 8 | - |
| 7 | 6 | - |
| 4 | 10 | - |
| 9 | 7 | - |
| 5 | 4 | - |
| 8 | 5 | - |
| 6 | 6 | - |
| 7 | 4 | - |
| 8 | 8 | - |

Dữ liệu test (testing) : ( 7, 5 ).

Dựa vào K= {1, 3, 5, 9} láng giềng gần nhất.

**Bước 2**: Tính khoảng cách :

d = sqrt( ( x\_2 - x\_1 )^2 + ( y\_2 - y\_1 )^2 )

|  |  |  |  |
| --- | --- | --- | --- |
| **X1** | **X2** | **Class** | **d** |
| 4 | 3 | + | 3.6 |
| 3 | 7 | + | 4.8 |
| 7 | 4 | + | 1 |
| 4 | 1 | + | 5 |
| 6 | 5 | + | 1 |
| 5 | 6 | + | 2.2 |
| 3 | 7 | + | 4.5 |
| 6 | 2 | + | 3.2 |
| 4 | 6 | - | 3.2 |
| 4 | 4 | - | 3.2 |
| 5 | 8 | - | 3.6 |
| 7 | 8 | - | 3 |
| 7 | 6 | - | 1 |
| 4 | 10 | - | 5.8 |
| 9 | 7 | - | 2.8 |
| 5 | 4 | - | 2.2 |
| 8 | 5 | - | 1 |
| 6 | 6 | - | 1.4 |
| 7 | 4 | - | 1 |
| 8 | 8 | - | 3.2 |

**Bước 3**: Sắp xếp các láng giềng theo khoảng cách

Sau khi tính toán khoảng cách, sắp xếp các láng giềng từ nhỏ đến lớn theo khoảng cách.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **X1** | **X2** | **Class** | **d** | **K** |
| 7 | 4 | + | 1 | 1 |
| 6 | 5 | + | 1 | 2 |
| 7 | 6 | - | 1 | 3 |
| 8 | 5 | - | 1 | 4 |
| 7 | 4 | - | 1 | 5 |
| 6 | 6 | - | 1.4 | 6 |
| 5 | 6 | + | 2.2 | 7 |
| 5 | 4 | - | 2.2 | 8 |
| 9 | 7 | - | 2.8 | 9 |
| 7 | 8 | - | 3 | 10 |
| 6 | 2 | + | 3.2 |  |
| 4 | 6 | - | 3.2 |  |
| 4 | 4 | - | 3.2 |  |
| 8 | 8 | - | 3.2 |  |
| 4 | 3 | + | 3.6 |  |
| 5 | 8 | - | 3.6 |  |
| 3 | 7 | + | 4.5 |  |
| 3 | 7 | + | 4.8 |  |
| 4 | 1 | + | 5 |  |
| 4 | 10 | - | 5.8 |  |

**Bước 4**: Xác định nhãn dựa của ( 7, 5 ) trên số lượng láng giềng gần nhất với mỗi giá trị 𝑘:

k (1, 3, 5, và 9 láng giềng gần nhất), xác định lớp (nhãn) của các láng giềng gần nhất:

* **k=1:**
  + Láng giềng gần nhất là: ( 7, 4 )(+).
  + → Dự đoán: +
* **k=3:**
  + Ba láng giềng gần nhất là: (7, 4 )(+), ( 6, 5 )(+), ( 7, 6 )(-).
  + → Đa số là: +
  + → Dự đoán: +
* **k=5:**
  + Năm láng giềng gần nhất là: (7, 4 )(+), ( 6, 5 )(+), ( 7, 6 )(-), ( 8, 5 )(-), ( 7, 4 )(-).
  + → Đa số là: -
  + → Dự đoán: -
* **k=9:**
  + Chín láng giềng gần nhất là :

(7, 4 )(+), ( 6, 5 )(+), ( 7, 6 )(-), ( 8, 5 )(-), ( 7, 4 )(-), (6 , 6 )(-), ( 5, 6 )(+), ( 5, 4 )(-), ( 9 , 7 )(-).

* + → Đa số là: -
  + → Dự đoán: -

11. Viết chương trình (source code) cho bài toán được lấy minh họa ở phần 4.

<https://github.com/Vanhieu1205/KPDL_KNN>

A screen shot of a computer screen

Description automatically generated

12. Trình bày Ưu/Nhược điểm của thuật toán KNN

**Ưu điểm:**

1. Đơn giản và dễ hiểu: Thuật toán dễ cài đặt và không đòi hỏi các giả định phức tạp về phân phối dữ liệu.

2. Linh hoạt: K-NN có thể được sử dụng cho cả bài toán phân loại và hồi quy.

3. Hiệu quả với dữ liệu đa chiều: K-NN có thể hoạt động tốt trên các tập dữ liệu đa chiều mà không cần biến đổi nhiều.

**Nhược điểm:**

1. Tính toán phức tạp: Khi kích thước tập dữ liệu lớn, việc tính toán khoảng cách với từng điểm dữ liệu sẽ mất nhiều thời gian.

2. Độ nhạy với nhiễu: K-NN rất nhạy cảm với các điểm dữ liệu nhiễu (outliers).

3. Yêu cầu tài nguyên: Thuật toán cần lưu trữ tất cả dữ liệu huấn luyện và tìm kiếm trong toàn bộ dữ liệu này để đưa ra dự đoán, đòi hỏi nhiều bộ nhớ và tài nguyên xử lý.

# PHẦN II: THỰC HÀNH

## Ex1: Cho một tập dữ liệu

**Bước 1**: Dữ liệu cho ( data training)

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **x** | 0.5 | 3.0 | 4.5 | 4.6 | 4.9 | 5.2 | 5.3 | 5.5 | 7.0 | 9.5 |
| **y** | - | - | + | + | + | - | - | + | - | - |

Dữ liệu test (testing) : 𝑧 = 5.0.

Dựa vào K= {1, 3, 5, 9} láng giềng gần nhất.

**Bước 2**: Tính khoảng cách :

d = sqrt(( x - z )^2 ).

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **x** | 0.5 | 3.0 | 4.5 | 4.6 | 4.9 | 5.2 | 5.3 | 5.5 | 7.0 | 9.5 |
| **y** | - | - | + | + | + | - | - | + | - | - |
| **x-z** | 4.5 | 2.0 | 0.5 | 0.4 | 0.1 | 0.2 | 0.3 | 0.5 | 2.0 | 4.5 |

**Bước 3**: Sắp xếp các láng giềng theo khoảng cách

Sau khi tính toán khoảng cách, sắp xếp các láng giềng từ nhỏ đến lớn theo khoảng cách.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **x** | 4.9 | 5.2 | 5.3 | 4.6 | 4.5 | 5.5 | 9.5 | 3.0 | 7.0 | 0.5 |
| **y** | + | - | - | + | + | + | - | - | - | - |
| **x-z** | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.5 | 4.5 | 2.0 | 2.0 | 4.5 |

**Bước 4**: Xác định nhãn dựa trên số lượng láng giềng gần nhất với mỗi giá trị 𝑘:

k (1, 3, 5, và 9 láng giềng gần nhất), xác định lớp (nhãn) của các láng giềng gần nhất và phân loại điểm

* **k=1:**
  + Láng giềng gần nhất là 4.9(+).
  + → Dự đoán: +
* **k=3:**
  + Ba láng giềng gần nhất là 4.9(+),5.2(-),5.3(-).
  + → Đa số là -.
  + → Dự đoán: -
* **k=5:**
  + Năm láng giềng gần nhất là 4.9(+),5.2(-),5.3(-),4.5(+),4.6(+).
  + → Đa số là +.
  + → Dự đoán: +
* **k=9:**
  + Chín láng giềng gần nhất là :

4.9(+), 5.2(-), 5.3(-), 5.5(+), 4.6(+), 4.5(+), 3.0(−), 7.0(−), 9.5(−)

* + → Đa số là -
  + → Dự đoán: -

## Ex2: Given a collection dataset as following: Using KNN method

**a)**

**Bước 1**: Dữ liệu cho ( data training)

|  |  |  |
| --- | --- | --- |
| **X1** | **X2** | **Class** |
| 4 | 3 | + |
| 3 | 7 | + |
| 7 | 4 | + |
| 4 | 1 | + |
| 6 | 5 | + |
| 5 | 6 | + |
| 3 | 7 | + |
| 6 | 2 | + |
| 4 | 6 | - |
| 4 | 4 | - |
| 5 | 8 | - |
| 7 | 8 | - |
| 7 | 6 | - |
| 4 | 10 | - |
| 9 | 7 | - |
| 5 | 4 | - |
| 8 | 5 | - |
| 6 | 6 | - |
| 7 | 4 | - |
| 8 | 8 | - |

Dữ liệu test (testing) : ( 7, 5 ).

Dựa vào K= {1, 3, 5, 9} láng giềng gần nhất.

**Bước 2**: Tính khoảng cách :

d = sqrt( ( x\_2 - x\_1 )^2 + ( y\_2 - y\_1 )^2 )

|  |  |  |  |
| --- | --- | --- | --- |
| **X1** | **X2** | **Class** | **d** |
| 4 | 3 | + | 3.6 |
| 3 | 7 | + | 4.8 |
| 7 | 4 | + | 1 |
| 4 | 1 | + | 5 |
| 6 | 5 | + | 1 |
| 5 | 6 | + | 2.2 |
| 3 | 7 | + | 4.5 |
| 6 | 2 | + | 3.2 |
| 4 | 6 | - | 3.2 |
| 4 | 4 | - | 3.2 |
| 5 | 8 | - | 3.6 |
| 7 | 8 | - | 3 |
| 7 | 6 | - | 1 |
| 4 | 10 | - | 5.8 |
| 9 | 7 | - | 2.8 |
| 5 | 4 | - | 2.2 |
| 8 | 5 | - | 1 |
| 6 | 6 | - | 1.4 |
| 7 | 4 | - | 1 |
| 8 | 8 | - | 3.2 |

**Bước 3**: Sắp xếp các láng giềng theo khoảng cách

Sau khi tính toán khoảng cách, sắp xếp các láng giềng từ nhỏ đến lớn theo khoảng cách.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **X1** | **X2** | **Class** | **d** | **K** |
| 7 | 4 | + | 1 | 1 |
| 6 | 5 | + | 1 | 2 |
| 7 | 6 | - | 1 | 3 |
| 8 | 5 | - | 1 | 4 |
| 7 | 4 | - | 1 | 5 |
| 6 | 6 | - | 1.4 | 6 |
| 5 | 6 | + | 2.2 | 7 |
| 5 | 4 | - | 2.2 | 8 |
| 9 | 7 | - | 2.8 | 9 |
| 7 | 8 | - | 3 | 10 |
| 6 | 2 | + | 3.2 |  |
| 4 | 6 | - | 3.2 |  |
| 4 | 4 | - | 3.2 |  |
| 8 | 8 | - | 3.2 |  |
| 4 | 3 | + | 3.6 |  |
| 5 | 8 | - | 3.6 |  |
| 3 | 7 | + | 4.5 |  |
| 3 | 7 | + | 4.8 |  |
| 4 | 1 | + | 5 |  |
| 4 | 10 | - | 5.8 |  |

**Bước 4**: Xác định nhãn dựa của ( 7, 5 ) trên số lượng láng giềng gần nhất với mỗi giá trị 𝑘:

k (1, 3, 5, và 9 láng giềng gần nhất), xác định lớp (nhãn) của các láng giềng gần nhất:

* **k=1:**
  + Láng giềng gần nhất là: ( 7, 4 )(+).
  + → Dự đoán: +
* **k=3:**
  + Ba láng giềng gần nhất là: (7, 4 )(+), ( 6, 5 )(+), ( 7, 6 )(-).
  + → Đa số là: +
  + → Dự đoán: +
* **k=5:**
  + Năm láng giềng gần nhất là: (7, 4 )(+), ( 6, 5 )(+), ( 7, 6 )(-), ( 8, 5 )(-), ( 7, 4 )(-).
  + → Đa số là: -
  + → Dự đoán: -
* **k=9:**
  + Chín láng giềng gần nhất là :

(7, 4 )(+), ( 6, 5 )(+), ( 7, 6 )(-), ( 8, 5 )(-), ( 7, 4 )(-), (6 , 6 )(-), ( 5, 6 )(+), ( 5, 4 )(-), ( 9 , 7 )(-).

* + → Đa số là: -
  + → Dự đoán: -

**b)**

**Bước 1**: Dữ liệu cho ( data training)

|  |  |  |
| --- | --- | --- |
| **X1** | **X2** | **Class** |
| 4 | 3 | + |
| 3 | 7 | + |
| 7 | 4 | + |
| 4 | 1 | + |
| 6 | 5 | + |
| 5 | 6 | + |
| 3 | 7 | + |
| 6 | 2 | + |
| 4 | 6 | - |
| 4 | 4 | - |
| 5 | 8 | - |
| 7 | 8 | - |
| 7 | 6 | - |
| 4 | 10 | - |
| 9 | 7 | - |
| 5 | 4 | - |
| 8 | 5 | - |
| 6 | 6 | - |
| 7 | 4 | - |
| 8 | 8 | - |

**Giả sử chọn :**

* Với 80% dữ liệu training:

|  |  |  |
| --- | --- | --- |
| **X1** | **X2** | **Class** |
| 4 | 3 | + |
| 3 | 7 | + |
| 7 | 4 | + |
| 4 | 1 | + |
| 6 | 5 | + |
| 5 | 6 | + |
| 4 | 6 | - |
| 4 | 4 | - |
| 5 | 8 | - |
| 7 | 8 | - |
| 7 | 6 | - |
| 4 | 10 | - |
| 9 | 7 | - |
| 5 | 4 | - |
| 8 | 5 | - |
| 6 | 6 | - |

* Và 20% dữ liệu testing

|  |  |  |
| --- | --- | --- |
| 6 | 2 | + |
| 3 | 7 | + |
| 7 | 4 | - |
| 8 | 8 | - |

**Bước 2**: Tính khoảng cách và kiểm tra kết quả :

* **Điểm kiểm tra (6, 2):**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **X1** | **X2** | **Class** | **d** | **K** |
| 4 | 3 | + | 2.24 | 1 |
| 7 | 4 | + | 2.24 | 2 |
| 4 | 1 | + | 2.24 | 3 |
| 5 | 4 | - | 2.24 | 4 |
| 4 | 4 | - | 2.83 | 5 |
| 6 | 5 | + | 3.00 | 6 |
| 8 | 5 | - | 3.61 | 7 |
| 6 | 6 | - | 4.00 | 8 |
| 5 | 6 | + | 4.12 | 9 |
| 7 | 6 | - | 4.12 |  |
| 4 | 6 | - | 4.47 |  |
| 3 | 7 | + | 5.83 |  |
| 9 | 7 | - | 5.83 |  |
| 5 | 8 | - | 6.08 |  |
| 7 | 8 | - | 6.08 |  |
| 4 | 10 | - | 8.25 |  |

* **k=1:**
  + Láng giềng gần nhất là: (4, 3)(+).
  + → Dự đoán: +
* **k=3:**
  + Ba láng giềng gần nhất là: (4, 3)(+), (7, 4)(+), (4, 1)(+).
  + → Đa số là: +
  + → Dự đoán: +
* **k=5:**
  + Năm láng giềng gần nhất là: (4, 3)(+), (7, 4)(+), (4, 1)(+), (5, 4 )(-), (4, 4)(-).
  + → Đa số là: +
  + → Dự đoán: +
* **k=9:**
  + Chín láng giềng gần nhất là :

(4, 3)(+), (7, 4)(+), (4, 1)(+), (5, 4 )(-), (4, 4)(-), (6 , 5)(+), (8, 5)(-), (6, 6 )(-), (5 , 6 )(+).

* + → Đa số là: +
  + → Dự đoán: +
* **Điểm kiểm tra (3, 7):**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **X1** | **X2** | **Class** | **d** | **K** |
| 3 | 7 | + | 0.00 | 1 |
| 4 | 6 | - | 1.41 | 2 |
| 5 | 6 | + | 2.24 | 3 |
| 5 | 8 | - | 2.24 | 4 |
| 4 | 4 | - | 3.16 | 5 |
| 6 | 6 | - | 3.16 | 6 |
| 4 | 10 | - | 3.16 | 7 |
| 5 | 4 | - | 3.61 | 8 |
| 6 | 5 | + | 3.61 | 9 |
| 4 | 3 | + | 4.12 |  |
| 7 | 6 | - | 4.12 |  |
| 7 | 8 | - | 4.12 |  |
| 7 | 4 | + | 5.00 |  |
| 8 | 5 | - | 5.39 |  |
| 9 | 7 | - | 6.00 |  |
| 4 | 1 | + | 6.08 |  |

* **k=1:**
  + Láng giềng gần nhất là: (3, 7)(+).
  + → Dự đoán: +
* **k=3:**
  + Ba láng giềng gần nhất là: (3, 7)(+), (4, 6)(-), (5, 6)(+).
  + → Đa số là: +
  + → Dự đoán: +
* **k=5:**
  + Năm láng giềng gần nhất là: (3, 7)(+), (4, 6)(-), (5, 6)(+), (5, 8 )(-), (4, 4)(-).
  + → Đa số là: -
  + → Dự đoán: -
* **k=9:**
  + Chín láng giềng gần nhất là :

(3, 7)(+), (4, 6)(-), (5, 6)(+), (5, 8 )(-), (4, 4)(-), (6 , 6)(-), (4, 10)(-), (5, 4)(-), (6 , 5 )(+).

* + → Đa số là: -
  + → Dự đoán: -
* **Điểm kiểm tra (7, 4):**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **X1** | **X2** | **Class** | **d** | **K** |
| 7 | 4 | + | 0.00 | 1 |
| 6 | 5 | + | 1.41 | 2 |
| 8 | 5 | - | 1.41 | 3 |
| 5 | 4 | - | 2.00 | 4 |
| 7 | 6 | - | 2.00 | 5 |
| 6 | 6 | - | 2.24 | 6 |
| 5 | 6 | + | 2.83 | 7 |
| 4 | 4 | - | 3.00 | 8 |
| 4 | 3 | + | 3.16 | 9 |
| 4 | 6 | - | 3.61 |  |
| 9 | 7 | - | 3.61 |  |
| 7 | 8 | - | 4.00 |  |
| 4 | 1 | + | 4.24 |  |
| 5 | 8 | - | 4.47 |  |
| 3 | 7 | + | 5.00 |  |
| 4 | 10 | - | 6.71 |  |

* **k=1:**
  + Láng giềng gần nhất là: (7, 4)(+).
  + → Dự đoán: +
* **k=3:**
  + Ba láng giềng gần nhất là: (7, 4)(+), (6, 5)(+), (8, 5)(-).
  + → Đa số là: +
  + → Dự đoán: +
* **k=5:**
  + Năm láng giềng gần nhất là: (7, 4)(+), (6, 5)(+), (8, 5)(-), (5, 4)(-), (7, 6)(-).
  + → Đa số là: -
  + → Dự đoán: -
* **k=9:**
  + Chín láng giềng gần nhất là :

(7, 4)(+), (6, 5)(+), (8, 5)(-), (5, 4)(-), (7, 6)(-), (6 , 6)(-), (5, 6)(+), (4, 4)(-), (4 , 3)(+).

* + → Đa số là: -
  + → Dự đoán: -
* **Điểm kiểm tra (8, 8):**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **X1** | **X2** | **Class** | **d** | **K** |
| 7 | 8 | - | 1.00 | 1 |
| 9 | 7 | - | 1.41 | 2 |
| 7 | 6 | - | 2.24 | 3 |
| 6 | 6 | - | 2.83 | 4 |
| 8 | 5 | - | 3.00 | 5 |
| 5 | 8 | - | 3.00 | 6 |
| 6 | 5 | + | 3.61 | 7 |
| 5 | 6 | + | 3.61 | 8 |
| 7 | 4 | + | 4.12 | 9 |
| 4 | 6 | - | 4.47 |  |
| 4 | 10 | - | 4.47 |  |
| 5 | 4 | - | 5.00 |  |
| 3 | 7 | + | 5.10 |  |
| 4 | 4 | - | 5.66 |  |
| 4 | 3 | + | 6.40 |  |
| 4 | 1 | + | 8.06 |  |

* **k=1:**
  + Láng giềng gần nhất là: (7, 8)(-).
  + → Dự đoán: -
* **k=3:**
  + Ba láng giềng gần nhất là: (7, 8)(-), (9, 7)(-), (7, 6)(-).
  + → Đa số là: -
  + → Dự đoán: -
* **k=5:**
  + Năm láng giềng gần nhất là: (7, 8)(-), (9, 7)(-), (7, 6)(-), (6, 6)(-), (8, 5)(-).
  + → Đa số là: -
  + → Dự đoán: -
* **k=9:**
  + Chín láng giềng gần nhất là :

(7, 8)(-), (9, 7)(-), (7, 6)(-), (6, 6)(-), (8, 5)(-), (5 , 8)(-), (6, 5)(+), (5, 6)(+), (7 , 4)(+).

* + → Đa số là: -
  + → Dự đoán: -

**Bước 3**: đánh giá hiệu xuất mô hình :

**Dữ liệu thực tế và dự đoán:**

* **Điểm kiểm tra 1: (6, 2).**
  + Nhãn thực tế: +
  + Dự đoán với k=1: +
  + Dự đoán với k=3: +
  + Dự đoán với k=5: +
  + Dự đoán với k=9: +
* **Điểm kiểm tra 2: (3,7).**
  + Nhãn thực tế: +
  + Dự đoán với k=1: +
  + Dự đoán với k=3: +
  + Dự đoán với k=5: -
  + Dự đoán với k=9: -
* **Điểm kiểm tra 3: (7,4).**
  + Nhãn thực tế: -
  + Dự đoán với k=1: +
  + Dự đoán với k=3: +
  + Dự đoán với k=5: -
  + Dự đoán với k=9: -
* **Điểm kiểm tra 4: (8,8).**
  + Nhãn thực tế: -
  + Dự đoán với k=1: -
  + Dự đoán với k=3: -
  + Dự đoán với k=5: -
  + Dự đoán với k=9: -

**Ma trận và đánh giá:**

* + **Độ chính xác(Accuracy) = (TN+TP) / (TP+FP+TN+FN).**
  + **Độ Nhạy (Recall): TP/(TP+FN)**
  + **Độ Đặc Hiệu (Specificity): TN/(TN+FP)**

* **K = 1:**
  + **Điểm kiểm tra 1**: Nhãn thực tế là +, dự đoán là + (đúng).
  + **Điểm kiểm tra 2**: Nhãn thực tế là +, dự đoán là + (đúng).
  + **Điểm kiểm tra 3**: Nhãn thực tế là -, dự đoán là + (sai).
  + **Điểm kiểm tra 4**: Nhãn thực tế là -, dự đoán là - (đúng).

**Ma trận K = 1:**

|  |  |  |
| --- | --- | --- |
|  | Dự đoán + | Dự đoán - |
| Thực tế + | 2 | 0 |
| Thực tế - | 1 | 1 |

**Độ chính xác K = 1:**

**Accuracy = 3/4 ​= 0.75 (75%)**

* **K = 3:**
  + **Điểm kiểm tra 1**: Nhãn thực tế là +, dự đoán là + (đúng).
  + **Điểm kiểm tra 2**: Nhãn thực tế là +, dự đoán là + (đúng).
  + **Điểm kiểm tra 3**: Nhãn thực tế là -, dự đoán là + (sai).
  + **Điểm kiểm tra 4**: Nhãn thực tế là -, dự đoán là - (đúng).

**Ma trận K = 3:**

|  |  |  |
| --- | --- | --- |
|  | Dự đoán + | Dự đoán - |
| Thực tế + | 2 | 0 |
| Thực tế - | 1 | 1 |

**Độ chính xác K = 3:**

**Accuracy = 3/4 ​= 0.75 ( 75% )**

* **K = 5:**
  + **Điểm kiểm tra 1**: Nhãn thực tế là +, dự đoán là + (đúng).
  + **Điểm kiểm tra 2**: Nhãn thực tế là +, dự đoán là - (sai).
  + **Điểm kiểm tra 3**: Nhãn thực tế là -, dự đoán là - (đúng).
  + **Điểm kiểm tra 4**: Nhãn thực tế là -, dự đoán là - (đúng).

**Ma trận K = 5:**

|  |  |  |
| --- | --- | --- |
|  | Dự đoán + | Dự đoán - |
| Thực tế + | 1 | 1 |
| Thực tế - | 0 | 2 |

**Độ chính xác K = 5:**

**Accuracy = 3/4 ​= 0.75 (75%)**

* **K = 9:**
  + **Điểm kiểm tra 1**: Nhãn thực tế là +, dự đoán là + (đúng).
  + **Điểm kiểm tra 2**: Nhãn thực tế là +, dự đoán là + (đúng).
  + **Điểm kiểm tra 3**: Nhãn thực tế là -, dự đoán là + (sai).
  + **Điểm kiểm tra 4**: Nhãn thực tế là -, dự đoán là - (đúng).

**Ma trận K = 9:**

|  |  |  |
| --- | --- | --- |
|  | Dự đoán + | Dự đoán - |
| Thực tế + | 1 | 1 |
| Thực tế - | 0 | 2 |

**Độ chính xác K = 9:**

**Accuracy = 3/4 ​= 0.75 (75%)**