Chương 5: Các thuật toán học tập có giám sát (supervised learning)

4.1 Giới thiệu học tập có giám sát

Quy trình trong hc giám sát:

1. Tập dữ liệu huấn luyện: chia thnahf hai phần chính-dữ liệu huấn luyện và dữ liệu kiểm thử. Dữ liệu huấn luyện chứa các mẫu dữ liệu đã được gán nhãn
2. Chọn mô hình: Chọn một mô hình học máy phù hợp với bài toán cụ thể. Các mô hình thông thường bao gồm các thuật toán như: Suppoet Vector machines, decision trees, random forest, neural networks và nhiều thuật toán khác.
3. Huấn luyện mô hình: Áp dụng dữ liệu huấn luyện vào mô hình để nó học cách ánh xạ từ dữ liệu đầu vào đến đầu ra . Trong quá trình này, mô hình điều chỉnh các tham số của nó để giảm sai số giữa đầu ra dự đoán và đầu ra thực tế
4. Kiểm thử mô hình: sử dụng dữ liệu kiểm thử để đánh giá hiệu suất của mô hình trên các dữ liệu mới mà nó chưa từng thấy trước đó
5. Tinh chỉnh mô hình: Nếu mô hình không hoạt động tốt trên dữ liệu kiểm thử, có thể cần phải điều chỉnh các tham số hoặc chọn một kiến trúc mô hình khác để cải thiện hiệu suất.

ứng dụng của học máy có giám sát:

* Nhận dạng hình ảnh và video( computer vision) : sử dụng AI để nhận dạng và phân tích nội dung từ hình ảnh và video, từ việc nhận biết khuôn mặt cho đến phân loại đối tượng
* Dịch ngôn ngữ: dùng để tự động dịch các ngôn ngữ khác nhau dựa trên dữ liệu đã được gán nhãn với các cặp câu tương ứng
* Y tế chăm sóc sức khỏe: sử dụng AI trong chẩn đoán bệnh, phát hiện tế bào ác tính và tư vấn về điều trị dựa trên dữ liệu y tế
* Xe tự lái: sử dụng AI để điều khiển các phương tiện tự động, giúp giảm tai nạn giao thông,...
* Dự đoán và phân tích dữ liệu: sử dụng AI để dự đoán xu hướng, biểu đồ và phân tích dữ liệu lớn.
* Tư vấn sản phẩm và tiếp thị
* Tiếp thị dựa trên hành vi
* Nhận dạng giọng nói

4.2 Giới thiệu bài toán phân loại và bài toán hồi quy

1. Bài toán phân loại

* Bài toán phân loại (classification) là một trong những bài toán quan trọng trong lĩnh vực học có giám sát ( superviesed learning)
* Trong bài toán này, mục tiêu là xây dựng một mô hình phù hợp
* Có nhiều ứng dụng: nhận diện ảnh, dịch ngôn ngữ tự nhiên, y học, spam filtering

2 Bài toán hồi quy

* Bài toán hồi quy (regression) là một dạng bài toán quan trọng trong lĩnh vực học có giám sát (supervised learning)
* Mục tiêu là dự đoán giá trị liên tục cho mỗi mẫu dữ liệu
* Hồi quy là một phương pháp thống kê và máy học được sử dụng để mô hình hóa mỗi quan hệ giữa một biến phụ thuộc và một hoặc nhiều biến độc lập
* Mục tiêu của hồi quy là tim ra một mô hình có thể mô tả và dự đoán mỗi quan hệ giữa các biến
* Pp này đặc biệt hữu ích khi chúng ta muốn hiểu và dự đoán giá trị của một biến phụ thuộc dựa trên các giá trị của một biến hoặc nhiều biến độc lập
* Ứng dụng bài toán hồi quy trong thực tế:
  + Dự đoán giá cả và tài sản
  + Quản lí doanh nghiệp và tài chính
  + Y tế
  + Nông nghiệp và môi trường
  + Khoa học và xã hội
  + Công nghiệp và kĩ thuật
  + Tài chính cá nhân

4.3 Hồi quy tuyến tính

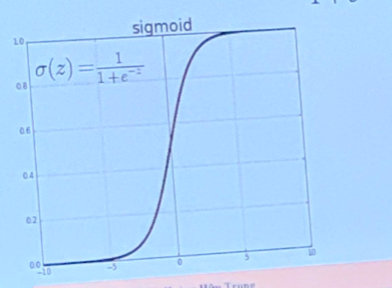
* Hồi quy tuyến tính là một phương pháp trong thống kê và máy học, được sử dụng để mô hình hóa mối quan hệ tuyến tính giữa một hoặc nhiều biến độc lập và một biến phụ thuộc
* Trong hồi quy tuyến tính, giả sử rằng mỗi quan hệ giữa các biến có thể được biểu diễn bằng một đường thẳng.

Trong đó:

* + - là biến phụ thuộc
    - B0 là hệ số chặn
    - Bi là hệ số góc
    - i là số lượng biến độc lập

**HỒI QUY LOGISTIC**

* Hồi quy logistic là một phương pháp trong thống kê và học máy được sử dụng để mô hình hóa và dự đoán xác suất của một biến phụ thuộc nhị phân (binary outcome), có nghãi là biến phụ thuộc chỉ nhân một trong hai giá trị có thể: 0 hoặc 1, “có” hoặc “không”, “thành công” hoặc “ thất bại”.
* Hồi quy logistic sử dụng hàm logistic ( hoặc hàm sigmoid) để chuyển đổi đầu ra của một hàm tuyến tính thành một giá trị nằm trong khoảng (0,1). Công thức hàm sigmoid như sau



* + - Công thức tổng quát của hồi quy logistic
* P(Y=1): xác xuất của sự kiện Y xảy ra

Ưu điểm của hồi quy Logistic

+ Dễ hiểu và dễ triển khai: có thể áp dụng một cách nhanh chóng mà không đòi hỏi nhiều công sức trong việc triển khai

+ TÍnh linh hoạt: Linear Regression có thể được áp dụng cho nhiều dạng dữ liệu và có thể mở rộng để xử lí các vấn đề phức tạp hơn thông qua các biến độc lập tuyến tính hoặc phi tuyến tính

+ Dễ hiểu và dễ giải thích: Kết quả của mô hình Linear Regression thường dễ giải thích cho người không chuyên môn

+ Hiệu suất tốt khi dữ liệu đơn giản: Khi mối quan hệ giữa biến độc lập và biến phụ thuộc là tuyến tính, Logistic Regression thường cho kết quả tốt và ổn định.

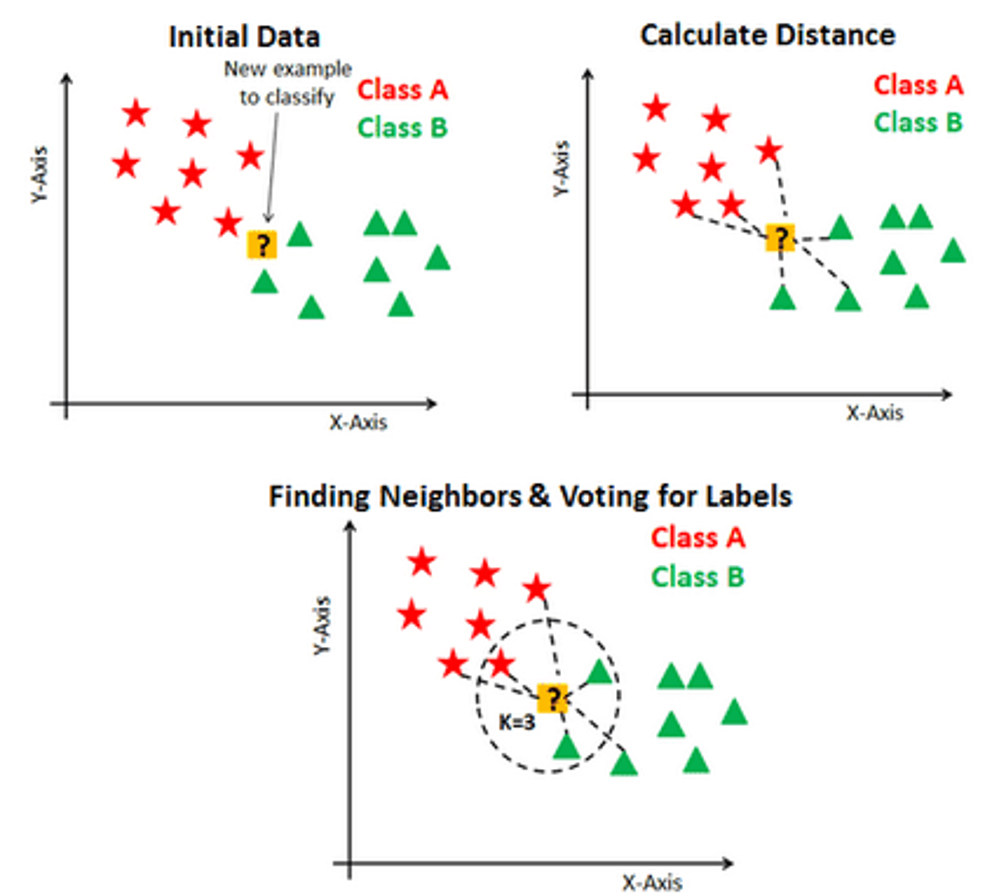
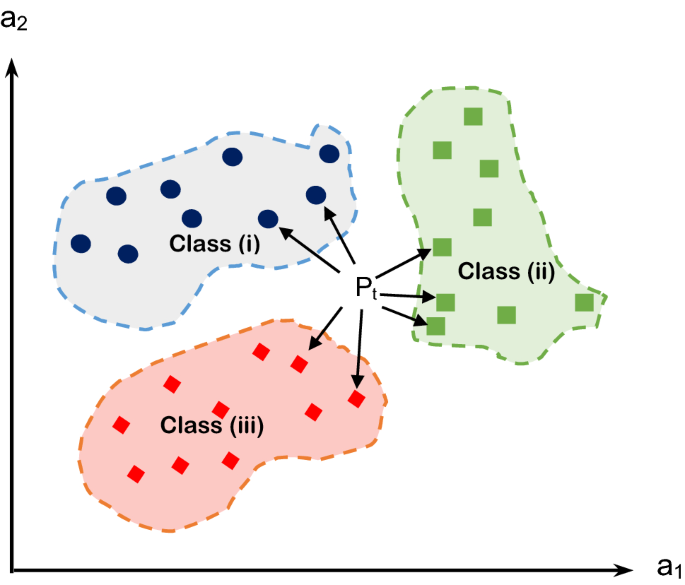
+ Nhạy cảm với dữ liệu nhiễu: Logistic có thể nhạy cảm với dữ liệu nhiễu, điều này có thể làm ảnh hưởng đến hiệu suất của mô hình

+ Mối quan hệ phi tuyến tính: nếu mối quan hệ giữa biến độc lập và biến phụ thuộc là phi tuyến tính, Logistic Regression chủ yếu được sử dụng để mô hình hóa mối quan hệ giữa một biến độc lập và một biến phụ thuộc. Đối với mối quan hệ phức tạp hơn, nó có thể không đủ linh hoạt

+ Assumpiton về độc lập của các biến : Logistic Regression đặt ra giả định về độc lập giữa các biến độc lập, điều này có thể không đáp ứng trong thực tế.

+ Không xử lí tốt với biến độc lập có phân phối không đồng đềuL Khi các biến độc lập không được phân phối đồng đều trong không gian giá trị. Logistic Regression có thể cho kết quả không chính xác.

GIỚI THIỆU K-NEAREST NEIGHBOOR

* Thuật toán K-nearest neighboor( KNN) là một kĩ thuật học máy phân loại học có giám sát phổ biến được sử dụng cho các nhiệm vụ phân loại và hồi quy.
* KNN dựa trên ý tưởng rằng các điểm dữ liệu tương tự có xu hướng có nhãn hoặc giá trị tương tự
* Trong giai đoạn huấn luyện, thuật toán KNN lưu trữ toán bộ tập dữ liệu huấn luyện làm tài liệu tham khảo
* Khi đưa ra dự đoán, nó sẽ tính toán khoẳng cách giữa điểm dữ liệu đầu vào và tất cả các mẫu huấn luyện bằng cách sử dụng thước đo khoảng cách đã chọn, chẳng hạn như khoảng cách Ecuclide.
* KNN xác định K hàng xóm gần nhất với điểm dữ liệu đầu vào dựa trên khoảng cách của chúng.
* Đối với hồi quy, nó tính toán giá trị trung bình hoặc trung bình có trọng số của các giá trị mục tiêu của K lân cận để dự đoán giá trị cho điểm dữ liệu đầu vào.
* Với KNN, trong bài toán Classification, label của một điểm dữ liệu mới (hay kết qủa của câu hỏi trong bài thi ) được suy ra trực tiếp từ K điểm dữ liệu gần nhất trong trainning dataset.
* Nhãn lớp của 1 test data có thể được quyết định bằng bỏ phiếu hoặc đa số (major voting) giữa các điểm gần nhất hoặc nó có thể được suy ra bằng cách đánh trọng số khác nhau cho mỗi trong các điểm gần nhất đó rồi suy ra nhãn lớp.
* 
* Sự khác biệt giữa các thuật ngữ này là về mặt kỹ thuật kỹ thuật, “bỏ phiếu đa số” yêu cầu đa số lớn hơn 50%, điều này chủ yếu có hiệu quả khi chỉ có hai loại.
* Khi có nhiều lớp, VD: ba loại không nhất thiết cần 50% phiếu bầu để đưa ra kết luận về một lớp, có thể chỉ định nhãn lớp với số phiếu bầu lớn nhất và lớn hơn 33,31%.
* 

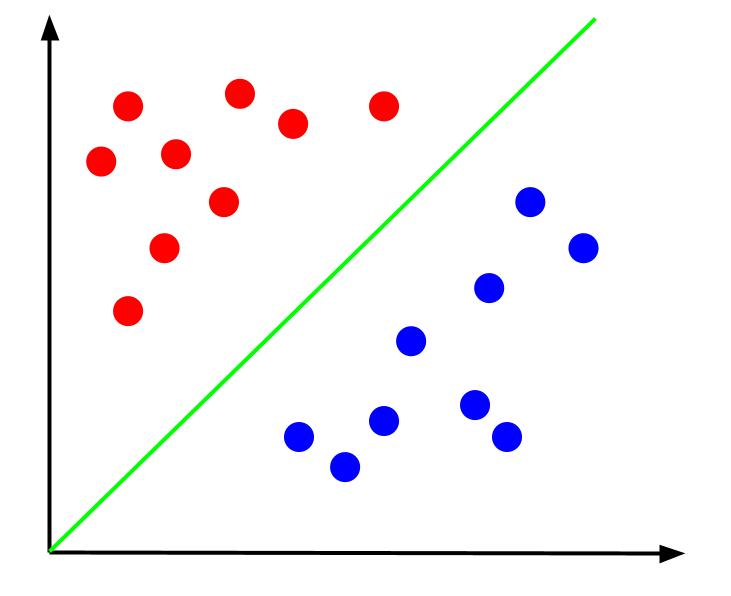
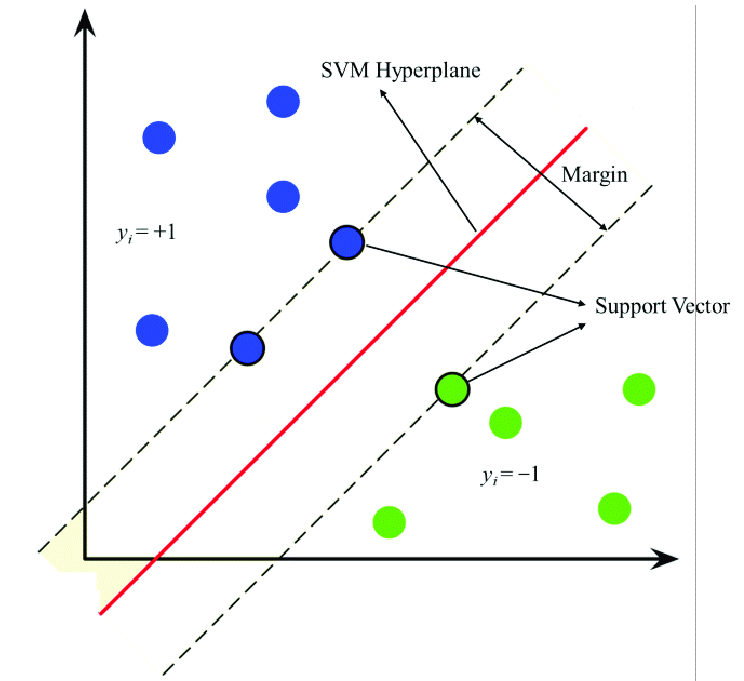
Ưu điểm của KNN

* Dễ triển khai và dễ hiểu: KNN là một thuật toán đơn giản và dễ hiểu. Điều này làm cho nó phù hợp cho cả người mới học và người làm việc trong lĩnh vực học máy
* Không yêu cầu giả định về phân phối: KNN không đòi hỏi giả định về phân phối dữ liệu. Nó có thể hoạt động tốt với dữ liệu phi tuyến tính và không chuẩn hóa.
* Hiệu quả đối với dữ liệu lớn: KNN có thể được sử dụng với dữ liệu lớn mà không cần mô hình. Nó thực hiện phân loại dựa trên K điểm gần nhất mà không cần tạo biểu đồ quyết định phức tạp.
* Có khả năng cập nhật dữ liệu: KNN cho phép cập nhật dữ liệu mới một cách dễ dàng. Khi có dữ liệu mới, bạn có thể tính toán lại K điểm gần nhất và cập nhật mô hình
* Khả năng hoạt động với đa dạng các loại đặc trưng: KNN có thể hoạt động với các loại đặc trưng nhưu đặc trưng hình ảnh, văn bản và dữ liệu số

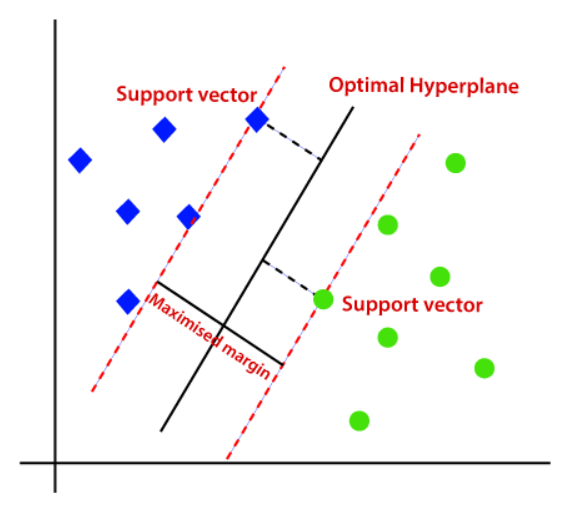
Nhược điểm của KNN

* Nhạy cảm với nhiễu: KNN nhạy cảm với nhiễu và dữ liệu bất thường. Dữ liệu nhiễu hoặc ngoại lai có thể ảnh hưởng đến hiệu suất của mô hình.
* Yêu cầu lựa chọn giá trị K phù hợp: Lựa chọn giá trị K thích hợp là một thách thức. Giá trị K quá nhỏ sẽ dẫn đế overfiting, trong khi giá trị K quá lớn có thể dẫn đến underfiting
* Tính toán độ phức tạp cao: KNN phải tính toán khoảng cách giữa điểm dữ liệu và tất cả các điểm dữ liệu khác trong tập huấn luyện. Điều này có thể làm cho thuật toán trở nên tốn thời gian đối với tập dữ liệu lớn
* Không có khả năng học được cấu trúc dữ liệu ẩn : KNN không có khả năng học được cấu trúc dữ liệu ẩn trong dữ liệu, Nó chỉ dựa vào thông tin từ các điểm gần nhất.
* Khó xử lí với dữ liệu không cân bằng: KNN không xử lí tốt với dữ liệu không cân bằng, nới một lớp có số lượng lớn hơn hẳn so với các lớp khác.

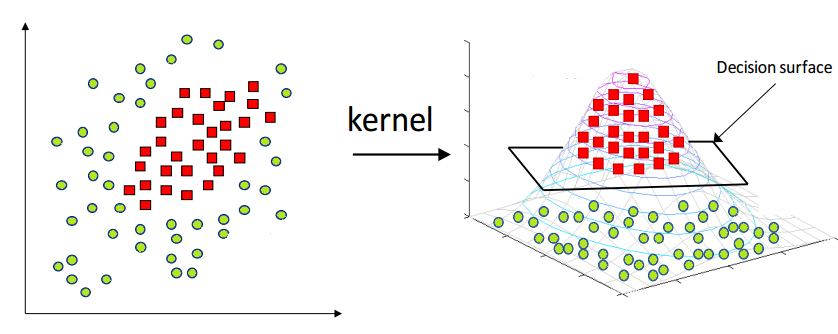
SUPPORT VECTOR MACHINE (SVM)

* Máy vector hỗ trợ ( support vector machine-SVM) là một thuật toán máy học thuộc lớp các thuật toán học có giám sát, thường được sử dụng chô cả bài toán phân loại và hồi quy.
* Dây là một thuật toán khá hiệu quả trong lớp các bài toán phân loại nhị phân và dự báo của học có giám sát.
* 
* SVM được thiết kế để tìm ra ranh giới quyết định tối ưu giữa các lớp dữ liệu
* Thuật toán này có ưu điểm là hoạt động tốt đối với những mẫu dữ liệu có kích thước lớn và thường mang lại kết quả vượt trội so với lớp các thuật toán khác trong học có giám sát.
* SVM tìm kiếm ranh giới phân chia giữa các điểm dữ liệu của các lớp sao cho khoảng cách từ các điểm đến ranh giới đó là lớn nhất.
* Khoảng cách từ các điểm dữ liệu đến ranh giới được gọi là “lề”( margin) và SVM cố gắng tối đa hóa lề này.
* 

Các vec to hỗ trợ :

* Một điểm trong không gian vector có thể được coi kà một vector từ gốc tọa độ tới điểm đó
* Các điểm dữ liệu nằm trên hoặc gần nhất với siêu phẳng được gọi là vector hỗ trợ, chúng ảnh hưởng đến vị trí và hướng của siêu phẳng.
* Các vecto này được sử dụng để tối ưu hóa lề và nếu xóa các điểm này, vị trí của siêu phẳng sẽ thay đổi
* Một điểm lưu ý nữa đó là các vecto hỗ trợ phải cách đều siêu phẳng
* 

Kernal SVM

* Ý tưởng cơ bản của Kernel SVM tìm một phép biến đổi sao cho dữ liệu ban đầu là không phan biệt tuyến tính được biến sang không gian mới
* ở không gian mới này dữ liệu trở nên phân biệt tuyến tính
* 

Ưu điểm của SVM

* Hiệu suất cao trong không gian chiều cao: Đặc biệt là trong các không gian có nhiều đặc trưng
* Hiệu suất tôs trong các tập dữ liệu lớn
* Khả năng xử lí dữ liệu không cân bằng: Đối mặt với vấn đề không cân bằng bằng cách chọn kernel thích hợp
* Khả năng tùy chỉnh kernel: SVM cho phép sử dụng nhiều loại kernel khác nhau tùy thuộc vào tbaif toán cụ thể , vd: tuyến tính, đa thucs, radial basis function(RBF),…
* Khả năng xử lí nhiễu: SVM chống lại nhiễu tốt , ví nó tìm ra ranh giới quyết định có biên lớn nhất, giảm thiểu ảnh hưởng của nhiễu.

Nhược điểm SVm

* Độ phức tạp tính toán: Đối với các bộ dữ liệu lớn, đặc biệt là khi số chiều tăng lên, SVM có thể yêu cầu nhiều tài nguyên.
* Nhạy cảm với lựa chọn tham số: SVM yêu cầu lựa chọn các tham số như kernel, C(cost), và gamma.
* Khó khăn khi xử lí dữ liệu lớn: Vì yêu cầu lưu toàn bộ tập huấn luyện trong bộ nhớ
* Kông phù hợp cho các bài toán nhiều lớp: Đối với bài toán nhiều lớp, phải sử dụng phương pháp one-vs-one hoặc one-vs-all, điều này có thể làm tăng độ phức tạp của mô hình
* Không thích hợp cho dữ liệu không tuyến tính và lớn: Sử dụng kernel function có thể gây tăng độ phức tạp tính toán.