**Bài 2: Machine Learning - Lập trình AI bằng Python**

Học tập có nghĩa là thu nhận kiến thức hoặc kỹ năng thông qua học tập hoặc kinh nghiệm. Dựa trên điều này, chúng ta có thể định nghĩa học máy (ML) như sau:

Nó có thể được định nghĩa là lĩnh vực khoa học máy tính, cụ thể hơn là một ứng dụng của trí tuệ nhân tạo, cung cấp cho các hệ thống máy tính khả năng học hỏi với dữ liệu và cải thiện từ kinh nghiệm mà không cần được lập trình rõ ràng.

Về cơ bản, trọng tâm chính của học máy là cho phép máy tính học tự động mà không cần sự can thiệp của con người. Bây giờ câu hỏi đặt ra là việc học như vậy có thể được bắt đầu và thực hiện như thế nào? Nó có thể được bắt đầu bằng việc quan sát dữ liệu. Dữ liệu có thể là một số ví dụ, hướng dẫn hoặc một số kinh nghiệm trực tiếp. Sau đó, trên cơ sở đầu vào này, máy đưa ra quyết định tốt hơn bằng cách tìm kiếm một số mẫu trong dữ liệu.

**1. Các mô hình ML cơ bản :**

Thuật toán học máy giúp hệ thống máy tính học mà không cần được lập trình rõ ràng. Các thuật toán này được phân loại thành có giám sát hoặc không được giám sát. Bây giờ chúng ta hãy xem một vài thuật toán :

**a. Thuật toán học giám sát:**

Đây là thuật toán học máy được sử dụng phổ biến nhất. Nó được gọi là có giám sát vì quá trình học thuật toán từ tập dữ liệu đào tạo có thể được coi như một giáo viên giám sát quá trình học tập. Trong loại thuật toán ML này, các kết quả có thể đã được biết trước và dữ liệu huấn luyện cũng được gắn nhãn với các câu trả lời đúng. Nó có thể được hiểu như sau:

Giả sử chúng ta có các biến đầu vào x và một biến đầu ra y và chúng tôi đã áp dụng một thuật toán để học hàm ánh xạ từ đầu vào đến đầu ra chẳng hạn như:

Y = f(x)

Bây giờ, mục tiêu chính là xấp xỉ hàm ánh xạ tốt đến mức khi chúng ta có dữ liệu đầu vào mới (x), chúng ta có thể dự đoán biến đầu ra (Y) cho dữ liệu đó.

Các vấn đề về tinh gọn được giám sát chủ yếu có thể được chia thành hai loại vấn đề sau :

* **Classification :**bài toán phân loại
* **Regression :**bài toán hồi quy

Decision tree, random forest, knn, logistic regression là ví dụ về học giám sát

**b. Thuật toán học không giám sát :**

Như tên cho thấy, các loại thuật toán học máy này không có bất kỳ người giám sát nào để cung cấp bất kỳ loại hướng dẫn nào. Đó là lý do tại sao các thuật toán học máy không được giám sát phù hợp chặt chẽ với cái mà một số người gọi là trí tuệ nhân tạo thực sự. Nó có thể được hiểu như sau:

Giả sử chúng ta có biến đầu vào x, thì sẽ không có biến đầu ra tương ứng như trong thuật toán học có giám sát.

Nói một cách dễ hiểu, chúng ta có thể nói rằng trong học tập không có giám sát sẽ không có câu trả lời chính xác và không có giáo viên hướng dẫn. Các thuật toán giúp khám phá các mẫu thú vị trong dữ liệu.

Các vấn đề học tập giám sát có thể được chia thành hai loại vấn đề sau:

* **Clustering**
* **Association**

**c. Thuật toán Refoincement :**

Thuật toán này huấn luyện hệ thống đưa ra các quyết định cụ thể. Về cơ bản, máy được tiếp xúc với môi trường nơi nó tự đào tạo liên tục bằng phương pháp thử và sai. Các thuật toán này học hỏi từ kinh nghiệm trong quá khứ và cố gắng nắm bắt kiến thức tốt nhất có thể để đưa ra quyết định chính xác. Quy trình quyết định Markov là một ví dụ về các thuật toán học máy tăng cường.

**2. Các thuật toán Machine learning thông dụng :**

Trong phần này, chúng ta sẽ tìm hiểu về các thuật toán học máy phổ biến nhất. Các thuật toán được mô tả sau :

**Linear Regression :**

Nó là một trong những thuật toán nổi tiếng nhất trong thống kê và học máy.

*Khái niệm cơ bản* - Hồi quy tuyến tính chủ yếu là một mô hình tuyến tính giả định mối quan hệ tuyến tính giữa các biến đầu vào là x và biến đầu ra duy nhất là y. Nói cách khác, chúng ta có thể nói rằng y có thể được tính toán từ sự kết hợp tuyến tính của các biến đầu vào x. Mối quan hệ giữa các biến có thể được thiết lập bằng cách lắp một dòng tốt nhất.

Linear regression được phân thành 2 loại như sau :

* **Simple linear regression :**chỉ có 1 biến độc lập
* **Multiple linear regression :**có nhiều hơn 1 biến

Hồi quy tuyến tính chủ yếu được sử dụng để ước tính các giá trị thực dựa trên (các) biến liên tục. Ví dụ: tổng doanh thu của một cửa hàng trong một ngày, dựa trên giá trị thực, có thể được ước tính bằng hồi quy tuyến tính.

**Logistic Regression**

Nó là một thuật toán phân loại và còn được gọi là hồi quy logit.

Chủ yếu hồi quy logistic là một thuật toán phân loại được sử dụng để ước tính các giá trị rời rạc như 0 hoặc 1, đúng hoặc sai, có hoặc không dựa trên một tập biến độc lập nhất định. Về cơ bản, nó dự đoán xác suất do đó đầu ra của nó nằm trong khoảng từ 0 đến 1.

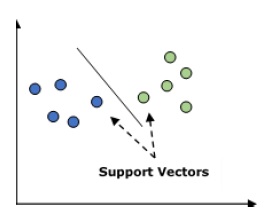
**Decision Tree**

Cây quyết định là một thuật toán học có giám sát được sử dụng chủ yếu cho các bài toán phân loại.

Về cơ bản nó là một bộ phân loại được biểu diễn dưới dạng phân vùng đệ quy dựa trên các biến độc lập. Cây quyết định có các nút tạo thành cây gốc. Cây có gốc là cây có hướng với một nút được gọi là "gốc". Gốc không có bất kỳ cạnh nào và tất cả các nút khác đều có một cạnh đến. Các nút này được gọi là lá hoặc nút quyết định. Ví dụ, hãy xem xét cây quyết định sau đây để xem liệu một người có phù hợp hay không.

**Support Vector Machine (SVM)**

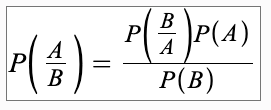
Nó được sử dụng cho cả vấn đề phân loại và hồi quy. Nhưng chủ yếu nó được sử dụng cho các bài toán phân loại. Khái niệm chính của SVM là vẽ mỗi mục dữ liệu dưới dạng một điểm trong không gian n chiều với giá trị của mỗi đối tượng là giá trị của một tọa độ cụ thể. Đây n sẽ là các tính năng chúng tôi sẽ có. Sau đây là một biểu diễn đồ họa đơn giản để hiểu khái niệm về SVM -



Trong sơ đồ trên, chúng ta có hai đặc điểm do đó trước tiên chúng ta cần vẽ hai biến này trong không gian hai chiều nơi mỗi điểm có hai tọa độ, được gọi là vectơ hỗ trợ. Dòng chia dữ liệu thành hai nhóm được phân loại khác nhau. Dòng này sẽ là bộ phân loại.

**Naïve Bayes**

Nó cũng là một kỹ thuật phân loại. Logic đằng sau kỹ thuật phân loại này là sử dụng định lý Bayes để xây dựng các bộ phân loại. Giả định là các yếu tố dự đoán là độc lập. Nói một cách đơn giản, nó giả định rằng sự hiện diện của một đối tượng cụ thể trong một lớp không liên quan đến sự hiện diện của bất kỳ đối tượng địa lý nào khác. Dưới đây là phương trình cho định lý Bayes :



Mô hình Naïve Bayes dễ xây dựng và đặc biệt hữu ích cho các tập dữ liệu lớn

**K-Nearest Neighbors (KNN)**

Nó được sử dụng để phân loại và hồi quy các vấn đề. Nó được sử dụng rộng rãi để giải quyết các vấn đề phân loại. Khái niệm chính của thuật toán này là nó được sử dụng để lưu trữ tất cả các trường hợp có sẵn và phân loại các trường hợp mới theo đa số phiếu bầu của k hàng xóm của nó. Sau đó, trường hợp được gán cho lớp phổ biến nhất trong số các hàng xóm K gần nhất của nó, được đo bằng hàm khoảng cách. Hàm khoảng cách có thể là khoảng cách Euclidean, Minkowski và Hamming. Hãy xem xét những điều sau để sử dụng KNN :

* Về mặt tính toán KNN đắt hơn các thuật toán khác được sử dụng cho các bài toán phân loại.
* Việc chuẩn hóa các biến cần thiết, nếu không, các biến phạm vi cao hơn có thể làm sai lệch nó.
* Trong KNN, chúng ta cần làm việc ở giai đoạn tiền xử lý như loại bỏ nhiễu.

**K-Means Clustering**

Như tên cho thấy, nó được sử dụng để giải quyết các vấn đề phân cụm. Về cơ bản nó là một kiểu học không giám sát. Logic chính của thuật toán K-Means là phân loại tập dữ liệu thông qua một số cụm. Làm theo các bước sau để tạo cụm bằng K-means -

* K-mean chọn k số điểm cho mỗi cụm được gọi là trung tâm.
* Bây giờ mỗi điểm dữ liệu tạo thành một cụm với các trung tâm gần nhất, tức là k cụm.
* Tìm ra các trọng tâm của mỗi cụm dựa trên các thành viên cụm hiện có
* Chúng ta cần lặp lại các bước này cho đến khi sự hội tụ xảy ra.

**Random Forest :**

Nó là một thuật toán phân loại có giám sát. Ưu điểm của thuật toán random forest là nó có thể được sử dụng cho cả loại bài toán phân loại và hồi quy. Về cơ bản, nó là tập hợp các cây quyết định (tức là rừng) hoặc bạn có thể nói là tập hợp các cây quyết định. Khái niệm cơ bản của rừng ngẫu nhiên là mỗi cây đưa ra một phân loại và khu rừng chọn những phân loại tốt nhất từ chúng. Tiếp theo là những ưu điểm của thuật toán Rừng ngẫu nhiên -

* Bộ phân loại random forest có thể được sử dụng cho cả bài toán phân loại và hồi quy.
* Ta có thể xử lý các giá trị bị thiếu
* Nó sẽ không phù hợp với mô hình ngay cả khi chúng ta có nhiều cây hơn trong rừng.