Machine Learning

1. Định nghĩa

Machine learing là 1 lĩnh vực trong khoa học máy tính sử dụng kĩ thuật thống kế để cho phép hệ thống máy tính có khả năng “học” với các dữ liệu mà không cần phải được lập trình rõ ràng. Ví dụ như các máy có thể "học" cách phân loại thư điện tử xem có phải thư rác(spam) hay không và tự động xếp thư vào thư mục tương ứng. Machine learning có mối quan hệ rất gần với các khái niệm Trí tuệ nhân tạo(AI), thống kê và Big data. Machine learning giúp cho máy tính có khả năng nhận thức cơ bản của con người như nghe, nhìn, hiểu được ngôn ngữ, giải toán, lập trình, … và từ đó Machine learning sử dụng các mô hình thống kê để "ghi nhớ" lại sự phân bố của dữ liệu của big data

1. Phân loại:

* Học có giám sát (Supervised Learning)

Trong phương pháp này chúng ta sẽ “huấn luyện” máy tính dựa trên những quan sát có dãn nhãn (labeled data). Bằng việc ghi nhớ và tổng quát từ 1 số quy tắc của tập câu hỏi và trả lời có đáp án từ trước máy tính có thể trả lời được những câu hỏi khác dù chưa gặp phải nhưng có mối liên quan. Supervised learning là thuật toán dự đoán đầu ra (outcome) của một dữ liệu mới (new input) dựa trên các cặp (input, outcome) đã biết từ trước. Cặp dữ liệu này còn được gọi là (data, label), tức (dữ liệu, nhãn). Supervised learning là nhóm phổ biến nhất trong các thuật toán Machine Learning.

* Học không giám sát (Unsupervised Learning)

Trong thuật toán này, chúng ta không biết được outcome hay nhãn mà chỉ có dữ liệu đầu vào. Thuật toán unsupervised learning sẽ dựa vào cấu trúc của dữ liệu để thực hiện một công việc nào đó, ví dụ như phân nhóm (clustering) hoặc giảm số chiều của dữ liệu (dimension reduction) để thuận tiện trong việc lưu trữ và tính toán.

* Học củng cố (Reinforcement Learning)

Reinforcement learning là các bài toán giúp cho một hệ thống tự động xác định hành vi dựa trên hoàn cảnh để đạt được lợi ích cao nhất (maximizing the performance). Hiện tại, Reinforcement learning chủ yếu được áp dụng vào Lý Thuyết Trò Chơi (Game Theory), các thuật toán cần xác định nước đi tiếp theo để đạt được điểm số cao nhất.

1. Các bước của 1 bài toán học máy

Một bài toán học máy cần trải qua 3 bước chính:

* **Chọn mô hình**: Chọn một mô hình thống kê cho tập dữ liệu. Ví dụ như mô hình thống kê Bec-nu-li, mô hình phân phối chuẩn.
* **Tìm tham số**: Các mô hình thống kê có các tham số tương ứng, nhiệm vụ lúc này là tìm các tham số này sao cho phù hợp với tập dữ liệu nhất có thể.
* **Suy luận**: Sau khi có được mô hình và tham số, ta có thể dựa vào chúng để đưa ra suy luận cho một đầu vào mới nào đó.

Bất cứ một bài toán học máy nào cũng đều cần có dữ liệu để huấn luyện, ta có thể coi nó là điều kiện tiên quyết. Dữ liệu sau khi có được cần phải:

* **Chuẩn hoá**: Tất cả các dữ liệu đầu vào đều cần được chuẩn hoá để máy tính có thể xử lý được. Quá trình chuẩn hoá bao gồm số hoá dữ liệu, co giãn thông số cho phù hợp với bài toán. Việc chuẩn hoá này ảnh hưởng trực tiếp tới tốc độ huấn luyện cũng như cả hiệu quả huấn luyện. Cụ thể ra sao thì ta sẽ cùng thảo luận trong một bài viết khác.
* **Phân chia**: Việc mô hình được chọn rất khớp với tập dữ liệu đang có không có nghĩa là giả thuyết của ta là đúng mà có thể xảy ra tình huống dữ liệu thật lại không khớp. Vấn đề này trong học máy được gọi là khớp quá (*Overfitting*). Vì vậy khi huấn luyện người ta phải phân chia dữ liệu ra thành 3 loại để có thể kiểm chứng được phần nào mức độ tổng quát của mô hình. Cụ thể 3 loại đó là:
  + **Tập huấn luyện** (*Training set*): Chiếm 60%. Dùng để học khi huấn luyện.
  + **Tập kiểm chứng** (*Cross validation set*): Chiếm 20%. Dùng để kiểm chứng mô hình khi huấn luyện.
  + **Tập kiểm tra** (*Test set*): Chiếm 20%. Dùng để kiểm tra xem mô hình đã phù hợp chưa sau khi huấn luyện.

Lưu ý rằng, tập kiểm tra ta phải lọc riêng ra và không được sờ tới, sử dụng nó trong khi huấn luyện. Còn tập huấn luyện và tập kiểm chứng thì nên xáo trộn đổi cho nhau để mô hình của ta được huấn luyện với các mẫu ngẫu nhiên nhất có thể