Phân loại các thuật toán Machine Learning

1, Đặt vấn đề

Thật ra nói về phần loại thì không có tính cố định, vấn đề chúng ta cần dựa theo tiêu chí nào. Ví dụ: Muốn phân loại sinh viên trong một lớp, nếu phân loại theo giới tính ta có: nam, nữ. Nếu phân loại theo lực học ta có: Xuất sắc, giỏi, khá, trung bình. Nếu phân loại theo nhóm làm bài tập ta có: Nhóm 1,2,3,…

Trong Machine Learning, nhắc đến phân loại ta dựa trên phân loại theo phương thức học là chủ yếu. Theo cách phân loại này ta có 4 loại chính: Học giám sát (Supervise Learning), học không giám sát (Unsupervise Learning), học bán giám sát (Semi – Supervise Learning), học củng cố (Reinforcement Learning).

2, Phân loại

2.1 Supervise Learning

Supervise Learning là một thuật toán dự đoán đầu ra (outcome) của một dữ liệu mới (new input) dựa trên cặp đầu vào, đầu ra (input, outcome) đã biết trước. Cặp dữ liệu còn được gọi là (data, label) tức (dữ liệu, nhãn dán). Supervise Learning là thuật toán phổ biến nhất trong Machine Learning.

Hiểu nôm na như này: Có một đứa bé tầm hai tuổi mà mình muốn dạy nó nhận biết một con mèo thì ta sẽ làm gì ? Tất nhiên là ta sẽ phải cho đứa bé đó biết hình ảnh con mèo như nào. Vậy ở đây, data là những hình ảnh, label là "con mèo". Vậy data này lấy ở đâu ? Các bước xử lý data như nào ? Thì mình xin được đề cập trong các bài tiếp theo.

Thuật toán Supervise Learning tiếp tục chia ra làm hai loại chính: Classification, Regression.

- Classification (Phân loại)
  Một bài toán gọi là Classification nếu các label của input data được chia thành một số hữu hạn nhóm. Ví dụ: Phân loại số, mail có spam hay không, một bạn có pass training IT PTIT hay không,… Đây là bài toán rất phổ biến trong Machine learning.
- Regression (Hồi quy)
  Từ "hồi quy" ở đây là chỉ mối quan hệ giữa các biến không phụ thuộc và biến phụ thuộc (Để hiểu hơn bạn nên học môn xác suất thống kê). Label ở đây không được chia thành nhóm như Classification mà là một giá trị cụ thể. Một ví dụ kinh điển đó là dự đoán giá nhà. Bạn thu thập được số m2 của một ngôi nhà, số phòng, cách trung tâm thành phố z km. Hãy dự đoán nhà này bán với giá bao nhiêu ?

2.2 Unsupervise Learning

Không như Supervise Learning, Unsupervise Learning ta chỉ biết dữ liệu đầu vào mà không biết nhãn của nó. Do vậy học không giám sát nó sẽ tự học, tự phát hiện ra cấu trúc

dữ liệu để thực hiện một công việc nào đó, ví dụ như phân nhóm (Clustering), giảm số chiều dữ liệu (Dimension reduction) để thuận tiện trong việc lưu trữ và tính toán.

- Clustering (Phân nhóm)
  Thuật toán Unsupervise Learning dựa trên sự liên quan giữa các dữ liệu để phân nhóm. Ví dụ, một ngân hàng cần phân loại các nhóm khách hàng của họ để đưa ra những gói dịch vụ phù hợp cho mỗi nhóm khách hàng đó. Hay đơn giản hơn ta cho một đứa trẻ phân loại hình học, có thể nó chưa biết đây là hình gì, gọi là gì nhưng nhiều khả năng nó sẽ phân loại được theo hình dạng (Tam giác, hình chữ nhật,…), theo màu sắc (đỏ, vàng,…).
- Association

  Là bài toán khi ta khám phá một quy luật dựa trên tập dữ liệu cho trước. Vấn đề này các bạn đọc thêm về Recommendation System (Hệ thống gợi ý) rất phổ biến trong các trang web bán hàng, ứng dụng xem phim, nhạc, báo,…

2.3 Semi – Supervise Learning

Khi ta có một lượng lớn giữ liệu nhưng chỉ có một trong số đó được gán nhãn. Bài toán nhóm này nằm giữa hai bài toán nhóm trên. Thật ra để có một dữ liệu ngon nghẻ được dán nhãn đầy đủ để cho máy học không hề dễ. Việc dán nhãn cho ảnh rất mất thời gian và chi phí cao. Ví dụ bạn có thể có rất nhiều ảnh chụp X – Quang về phổi nhưng để biết với ảnh X – Quang như trên ta dán nhãn cho việc bị bệnh gì là điều khó, chỉ có những chuyên gia về y học biết. Việc thu thập dữ liệu chi phí thấp chủ yếu ta lấy từ Internet. Làm việc với dữ liệu tốn thời gian và công sức rất nhiều so với việc ta xây dựng thuật toán. Bạn đọc có thể google về việc dán nhãn dữ liệu ở Trung Quốc.

2.4 Reinforcement Learning

Reinforcement Learning là một bài toán giúp cho hệ thống tự động xác định hành vi dựa trên môi trường để được lợi ích cao nhất. Bài toán này thường áp dụng trong lý thuyết trò chơi. Ví dụ, bạn muốn đào tạo cho máy học cách chơi cờ vua, bạn ném cho con máy luật chơi cờ vua (mã đi như nào, tốt đi như nào,… khi nào thì thắng). Khi đó, máy sẽ tự chơi với chính nó để tìm ra phương án đánh tối ưu nhất (dành chiến thắng ván cờ).