Machine learning

ML được chia làm 3 loại: học có giám sát, học không giám sát và học tăng cường.

1. Các thuật toán học có giám sát

Học có giám sát sẽ sử dụng dữ liệu được gán nhãn để học.

Có 2 loại bài toán cơ bản của học có giám sát đó là: phân loại và hồi quy (classification và regression).

Các thuật toán sẽ đề cập tới trong loại học máy này bao gồm: linear regression, logistic regression, CART, mạng bayes và KNN.

Ngoài ra còn có bagging with random forest và boost with xgboost thuộc loại kết hợp.

1. Học không giám sát.

Học không giám sát sẽ học trên dữ liệu không nhãn.

Các kiểu bài toán ứng dụng với loại học không giám sát là: tìm mối liên kết, phân cụm dữ liệu giống nhau và giám kích thước dữ liệu. (association, clustering và dimentionality reduction)

Các thuật toán của loại học không giám sát chúng ta sẽ đề cập đến là: apriori, k-mean, PCA

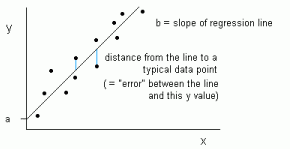
1. Học tăng cường.

Học tăng cường là loại học máy mà model của chúng ta sẽ thử - sai – làm lại để có thể đạt được kết quả tốt nhất. Tóm lại với học giám sát, chúng ta sẽ rút kinh nghiệm từ những lần làm sai.

10 thuật toán ML cho người mới bắt đầu

1. Linear regression

Thuật toán linear regression hay còn gọi là hồi quy tuyến tính dựa trên các điểm dữ liệu có được (x,y) – x là dữ liệu, y là nhãn – thuật toán sẽ xây dựng nên một hàm tuyến tính y = ax+b sao cho nó gần nhất với tối đa các điểm dữ liệu



Như trong hình ta có thể thấy đường thẳng đi qua giữa các điểm dữ liệu thành một đường chéo.

Kết quả của thuật toán là tìm ra được a và b thông qua tập dữ liệu training.

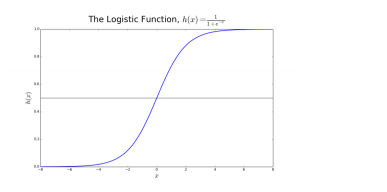
Ví dụ:

1. Logistic regression

Thuật toán này tương tự linear regression nhưng thay vì tìm ra một hàm tuyến tính (y=ax+b) thì sẽ tìm ra một hàm thông qua hàm logistic (h(x)= 1/ (1 + e^ -x) ).

Kết quả đầu ra của model xây với thuật toán logistic regression là một tỷ lệ (0->1).

P(x) = e ^ (b0 +b1x) / (1 + e(b0 + b1x))    
Mục đích của thuật toán là tìm ra b0 và b1 để xây dựng hàm.



Nếu P(x) > 0.5 ta sẽ kết luận x có nhãn 1 và ngược lại (với giả sử chỉ có 2 nhãn 0 và 1).

Ví dụ:

1. CART

Cây phân loại và cây hồi quy cùng thuộc họ thuật toán cây quyết định.

Ví dụ:

1. Naive bayes (phân loại bayes).

Thuật toán này dựa trên công thức bayes để dự đoán kết quả của input.



Ta sẽ tính xác suất nhận nhãn + và xác suất nhận nhãn – rồi xem kết quả nào lớn hơn thì kết luận.

Ví dụ:

1. KNN

KNN hay còn gọi là k láng giềng gần nhất.

Đại loại thuật toán này sẽ tìm k điểm gần nhất với dữ liệu test để kết luận nhãn của dữ liệu đó.

Ví dụ:

Tiếp theo là nhóm thuật toán học không giám sát

1. Apriori
2. K-means
3. PCA

Ensemble learning techniques:

1. Bagging with random forests
2. Boosting with adaboost