IBM AI Engineering Professional Certificate

# CHƯƠNG 1. Machine Learning with Python

## Giới thiệu về machine learning

Là lĩnh vực con của khoa học máy tính, tập hợp các thuật toán để giải quyết bài toán về dự đoán, phân loại.

Có các phương pháp học tập trong ML là: supervised, unsupervised, semi-supervised, self-supervised, … Trong đó có 2 phương pháp học chính và phổ thông nhất là supervised, unsupervied.

Supervised là phương pháp học có giám sát, nghĩa là khi 1 dữ liệu được đưa vào mô hình thì nó đã được gắn nhãn, hoặc có kết quả rõ ràng, ví dụ: khi đưa dữ liệu vào để máy có thể nhận diện khuôn mặt thì dữ liệu đó đã dược gắn nhãn tương ứng với tên, tuổi, …

Unsupervised là phương pháp học không giám sát, nghĩa là khi 1 dữ liệu được đưa vào mô hình thì mô hình phải cố gắng tìm ra được cấu trúc ẩn bên trong dữ liệu đó, ví dụ: đưa dữ liệu là các loài động vật thì mô hình phải cố gắng tự phân loại các loài vật.

## Bài toán của ML

### Hồi quy

Là dựa vào các thông tin có sẵn để có thể dự đoán một giá trị tiếp theo. Thuật toán phổ biến trong hồi quy là hồi quy tuyến tính.

Hồi quy tuyến tính là tìm ra mối quan hệ giữa các thông tin đầu vào với thông tin đầu ra dựa trên phương trình toán học. Có 2 phương pháp hồi quy tuyến tính là hồi quy tuyến tính đơn giản (chỉ chứa 1 tham số độc lập) và hồi quy tuyến tính bội (chứa nhiều hơn 1 tham số độc lập).

Để đánh giá một mô hình hồi quy thì thường sử dụng các số liệu sau: MAE (Mean absolute error), MSE (Mean Square Error), RSMA (Root Mean Square Error), RAE (Realative Absolute Error), RSA (Relative Square Error).

Tham số đánh giá được sử dụng nhiều nhất là MSE= 1/n \* 2 là khoảng cách bình phương trung bình giữa giá trị thực tế và giá trị mô hình dự đoán.

### Phân loại

Là nhóm các dữ liệu có điểm chung với nhau thành các nhóm. Một số thuật toán phổ biến là hồi quy logistic, k-nearest neighbor, cây quyết định, support vector machine, …

* Hồi quy logistic: là dựa vào các thông tin đã có để đưa ra một dự đoán hữu hạn nhị phân, cụ thể như có hoặc không, đúng hoặc sai, thành công hoặc thất bại, cao hoặc thấp, …

Ví dụ như ta có mối quan hệ như sau: ծ (ƟTX) 🡪 P(y=1|x)

Với mỗi giá trị theta và x, mô hình sẽ cho ra một dự đoán ŷ, và từ đó tính ra xác suất y=1 với mỗi x. Hồi quy logistic sử dụng hàm logistic (sigmoid)

Bước 1: Tạo ngẫu nhiên giá trị theta (Ɵ)

Bước 2: Với mỗi giá trị x từ dữ liệu tính toán ra ŷ (ŷ= ծ (ƟTX))

Bước 3: Tính toán mất mát giữa giá trị được tính bởi mô hình và thực tế (y - ŷ)

Bước 4: Tính toán tổng mất mát giữa mọi giá trị từ dữ liệu cho trước (J())

Bước 5: Thay đổi theta bằng giá trị khác và thực hiện lại Bước 2 đến Bước 4

Thuật toán hồi quy logistic cố gắng tìm ra các giá trị theta sao cho tổng mất mát giữa giá trị mô hình và thực tế thấp nhất (J() min). Khi đạt được đến độ chính xác yêu cầu thì mô hình ngừng lại. Để tối ưu hóa bước 5 thì có thể sử dụng gradient descent để có thể đạt hiệu suất tốt nhất.

Có thể sử dụng các phương pháp thay đổi tham số khác ngoài phương pháp gradient, mục đích của việc thay đổi tham số để tìm ra được theta phù hợp nhất, hàm mất mát giữa giá trị dự đoán và giá trị thực tế là thấp nhất. Đây cũng là bước khó và quan trọng nhất trong hồi quy logistic.

* Support Vector Machine

Đưa dữ liệu đầu vào vào các không gian đa chiều, thuật toán SVM sẽ tạo ra một mặt siêu phẳng bởi phương trình ngăn cách các dữ liệu với nhau. Với không gian 2 chiều gồm x, y thì siêu phẳng là một đường thẳng, với không gian 3 chiều thì siêu phẳng là một mặt phẳng, …

Để đưa dữ liệu vào các chiều không gian cao hơn thì sử dụng các hàm toán học khác nhau như: sigmoid, RBF, đa thức, …

Hàm SVM đạt hiệu quả tốt với các dữ liệu nhiều chiều và tối ưu hóa bộ nhớ, tuy nhiên lại không đạt được kết quả tốt với những dữ liệu lớn từ 1000 dòng và không cung cấp trực tiếp ước tính xác suất. SVM sử dụng tốt cho việc phát hiện chữ viết tay, phát hiện thư rác, phân tích dữ liệu GEN.

* K -Mean Clustering

Thuật toán này nhóm các dữ liệu dựa trên khoảng cách thành các cụm, việc xác định khoảng cách giữa các đối tượng (dữ liệu) bằng các hàm toán học khác nhau, trong đó hàm khoảng cách thông dụng nhất là euclide.

Phương pháp K -Mean sẽ đặt các vị trí k-centroid ngẫu nhiên, từ đó gom các đối tượng vào các k-centroid thành các cụm. Sau đó tính toán lại các vị trí k-centroid bằng cách tính trung bình khoảng cách của các đối tượng trong cụm. Rồi gom các đối tượng lại thành các cụm mới và thực hiện lại các bước tính toán vị trí k -centroid cho đến khi vị trí này không đổi nữa.

Nhược điểm của thuật toán này là phải xác định được trước số cụm của dữ liệu để có thể khởi tạo được bao nhiêu vị trí k -centroid, điều này là không dễ dàng.

## Các thông số để đánh giá mô hình thuật toán phân loại.

Để đánh giá một mô hình ML phân loại cần xác định các thông số sau:

Accuracy, precision, recall, f1 score, ma trận nhầm lẫn, trong đó 2 thông số thường xuyên được sử dụng là accuracy (độ chính xác), precision (độ chính xác khi kết quả dự đoán đúng).

Accuracy = (TP +TN)/ (tổng số số lượng dự đoán/phân loại)

Precision = TP/(TP+FP)

# Chương 2. Introduction to Deep Learning & Neural Networks with Keras

## 2.1 Introduction to Deep Learning