Matplotlib: thư viện hiển thị dữ liệu dưới dạng đồ thị hay là biểu đồ

Graphical user interface, text, application

Description automatically generated

**-*Covariance*:** Cả hiệp phương sai và tương quan là hai khái niệm trong lĩnh vực xác suất thống kê, đều nói về mối quan hệ giữa các biến.

+ Cả hiệp phương sai và tương quan đều là những công cụ quan trọng được sử dụng trong việc thăm dò dữ liệu để lựa chọn đối tượng địa lý và phân tích đa biến.

Ví dụ, một nhà đầu tư đang tìm cách phân tán rủi ro của danh mục đầu tư có thể tìm kiếm các cổ phiếu có hiệp phương sai cao, vì điều đó cho thấy giá của chúng tăng cùng một lúc. Tuy nhiên, chỉ riêng một chuyển động tương tự là chưa đủ. Sau đó, nhà đầu tư sẽ sử dụng số liệu tương quan để xác định mức độ liên kết chặt chẽ giữa các giá cổ phiếu đó với nhau.

***-Hồi quy tuyến tính***: là 1 thuật toán học có giám sát, trong đó đầu ra dự đoán là liên tục và có độ dốc không đổi

Nó được sử dụng để dự đoán các giá trị trong 1 phạm vi liên tục, (ví dụ: doanh số, giá cả) thay vì cố gắng phân loại chúng thành các danh mục hay nhóm (ví dụ mèo, chó)

Nó có thể được sử dụng cho các trường hợp chúng ta muốn dự đoán một số lượng liên tục. Ví dụ, dự đoán giao thông ở một cửa hàng bán lẻ, dự đoán thời gian người dùng dừng lại một trang nào đó hoặc số trang đã truy cập vào một website nào đó v.v...

\*\* 2 loại \*\*

+Simple regression

+Multivariable regression

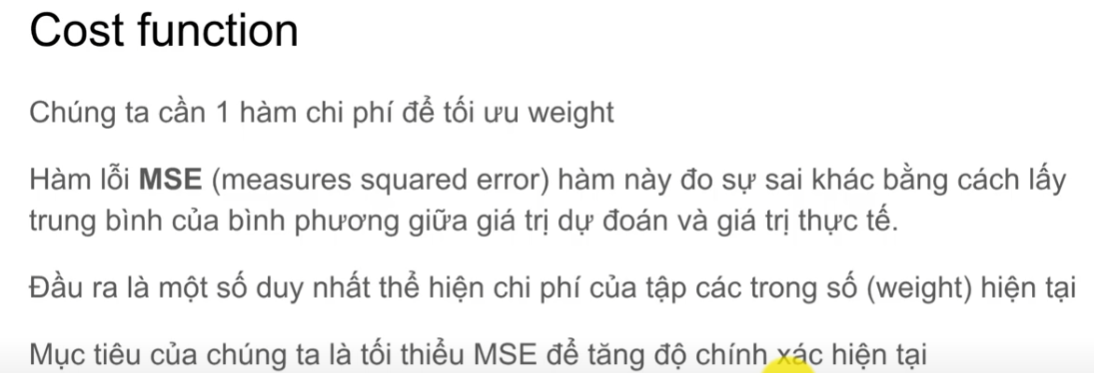
\*\*\*\* Học có giám sát thì bộ dataset phải gán nhãn

Radio: biến độc lập. Trong machine learning gọi là features

Weight hệ số của biến độc lập. ……………………………. Weight

Bias: gía trị lệch để bù đắp sai số

***Sales= weight \* radio+bias***



Timeline

Description automatically generated

***Hồi quy logistic:*** Graphical user interface, text, application

Description automatically generated

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Chart

Description automatically generated with medium confidence

Graphical user interface, text, application

Description automatically generated

**Neural Network**

Diagram

Description automatically generated with medium confidence

<h4>Học có giám sát và học không giám sát</h4>

+ Học có giám sát là bài toán hồi quy

<h4>Học có giám sát là bài toán phân loại</h4>

Bao gồm: Multiclass (Phân loại nhiều lớp)

+ có rất nhiều nhãn nhưng 1 y chỉ ứng với 1 x

+ check Spam filtering, y in {spam, normal}

+ Discovery of network attacks

+ Financial risk estimation: y in{high, normal, no}

Multilabel (Phân loại đa nhãn)

+ ex: Birds nest tree

+ Output y is subset of labels (Mỗi output là 1 tập nhỏ các lớp, mỗi quan sát x có thể có nhiều nhãn)

+Image tagging y = {bird, nest, tree}

+sentiment analysis <br>

+ Prediction of stock indices <br>

<h4>Học không giám sát</h4>

+ Clustering data into cluster (vân tay , ...) <br>

+ Comunity detection <br>

+ Trend detection <br>

Overfitting: làm rất tốt với tập dữ liệu train nhưng với tập thực tế thì lại rất tệ (quá khít, quá khớp) <br>

Underfitting: Làm tệ với cả 2 tập dữ liệu <br>

+ Regularization là 1 mô hình giúp chúng ta huấn luyện được mô hình tốt và tránh được hiện tượng overfitting <br>

<h4>Tiền xử lý dữ liệu</h4>

Thời gian dành cho phân tích dữ liệu <br>

+ Thu thập dữ liệu: 19% <br>

+ Thu xếp và làm sạch dữ liệu: 60% <br>

+ Tạo tập dữ liệu huấn luyện: 3% <br>

+ Khai phá: 9% <br>

+ Cải thiện thuật toán: 4% <br>

+ Khác: 5% <br>

Tiền xử lý dữ liêu để làm gì? Thuận tiện trong việc lưu trữ, truy vấn <br>

Các mô hình học máy thường làm việc với các mô hình dữ liệu có cấu trúc ma trận vector, chuỗi,.... <br>

<h4>Source riêng https://users.soict.hust.edu.vn/khoattq/ml-dm-course/131404-IT3190/ (web xấu vãi ò)</h4>

<h4> Sách tham khảo </h4>

<p>

[1] Tom Mitchell. Machine Learning. McGraw-Hill, 1997.

[2] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. Deep Learning. MIT press, 2016.

[3] Jiawei Han, Micheline Kamber, Jian Pei. Data Mining: Concepts and Techniques (3rd Edition). Morgan Kaufmann, 2011.

[4] Trevor Hastie,‎ Robert Tibshirani,‎ Jerome Friedman. The Elements of Statistical Learning: Data Mining, Inference, and Prediction (12th Edition). Springer, 201</p>