**Lesson 4: Machine Learning Validation, Evaluation and Probability**

1. **Review Regression Techniques :**

Như vậy , sau 3 buổi học, chúng ta đã đi qua được thuật toán Regression , với 3 bài toán chính :

**Linear Regression**

Y = ax + b

**Non Linear Regression**

**Y = ax^n + b**

**Logistic Regression**

y = σ(w0 +w1 ∗ + w2 ∗ )

**Warm up** : Dựa vào kiến thức đã học, hãy trả lời các câu hỏi.

**1. Nhận xét về các mối tương quan ( correlation ) giữa x và y trong các biểu đồ sau. Với biểu đồ nào thì có thể dùng Linear regression để dự đoán. Và dự đoán bằng cách nào ? ( Có thể trả lời theo thuật toán , hoặc kiến thức phương trình đường thẳng )**

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

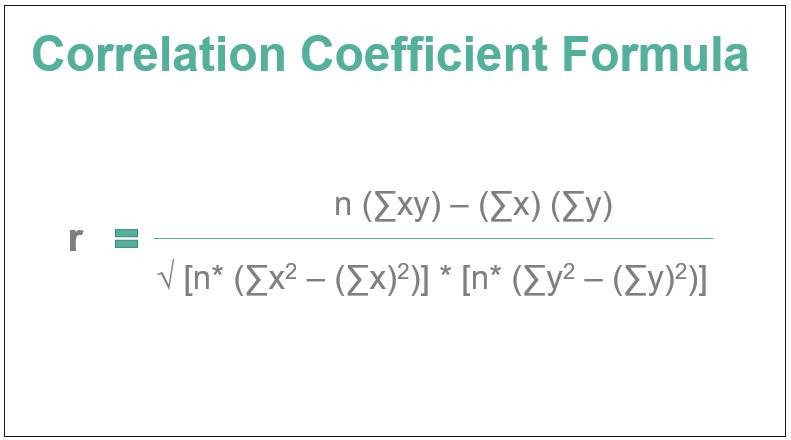
Table, calendar

Description automatically generated

Với dữ liệu ở trên thì có thể xây dựng được một linear regression để dự đoán không ? , và độ chính xác sẽ là bao nhiêu ( gợi ý : Correlation Coefficient )

correlation coefficient is calculated by **first determining the covariance of the variables and then dividing that quantity by the product of** those variables' standard deviations.

**Công thức**: correlation coefficient = covariance(X, Y) / (stdv(X) \* stdv(Y))



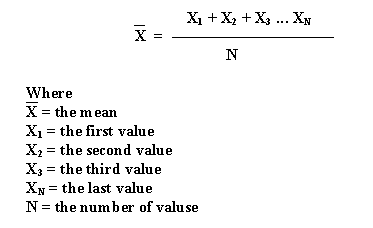
Hoặc mở python lên

from scipy.stats import pearsonr

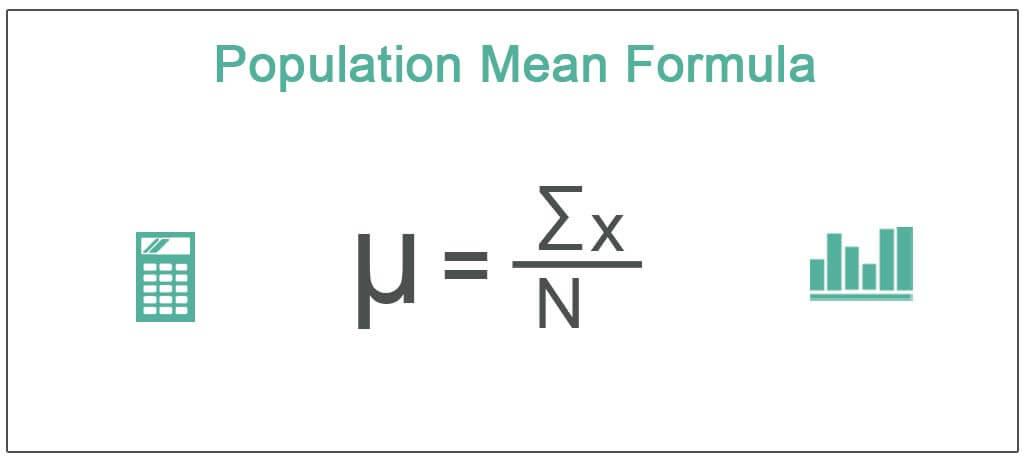
r = pearsonr(data1, data2)

**Ôn lại các chỉ số quan trọng trong statistics và ứng dụng của nó trong Linear Regression :**

**Mean : Tức giá trị trung bình**



**Công thức rút gọn:**

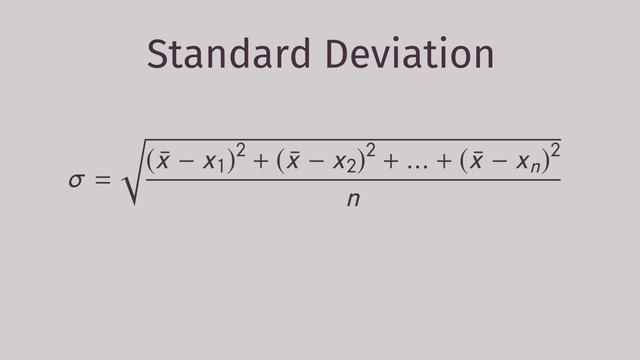


**Phương sai (Variance) :** là sự phân tán của các biến trong bộ dữ liệu , Phương sai càng gần không, các điểm dữ liệu càng nhóm lại gần nhau ( dữ liệu bị phân tán quá mức sẽ ảnh hưởng đến kết quả dự đoán, do bộ dữ liệu bị nhiễu quá lớn)

Table, Excel

Description automatically generated

**Standard deviation :** Độ lệch chuẩn , bằng căn bậc 2 của phương sai , chỉ mức độ lệch khỏi giá trị trung bình.



**Standard error** : hoặc standard error of the mean , tức là độ lệch chuẩn của giá trị trung bình của mẫu so với giá trị trung bình của tập hợp.

**Khái niệm sample , population** .

Sample là một mẫu nhỏ

Population là một mẫu lớn, một population có thể có nhiều sample.

* Standard error of the Mean là độ lệch của mean của các sample so với Mean của mẫu. Đối với các bài toán nhỏ thì Standard Deviation cũng là Standard Error.

Ví dụ : Khi các bạn chạy model , các bạn hay chia tập dữ liệu ra làm 2 sample là train / test.

Vậy sample train và sample test sẽ có giá trị trung bình khác nhau, và cũng khác so với tập hợp ban đầu. => Khác như thế nào ? => Tính standard error of the mean.

**Vì sao phải tính mean , standard deviation ?**

* Vì khi các bạn dự đoán 1 kết quả, kết quả được dự đoán sẽ lệch với giá trị thực tế trong khoảng +/- standard deviation.

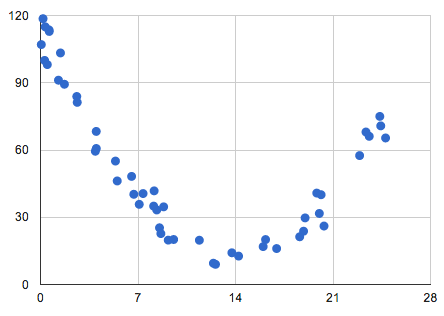
Ví dụ : Dự đoán 1 chiếc xe 2 năm tuổi sẽ có giá 40k usd +/- 5k usd. Giá thực tế là 43k usd => Predicted value dao động trong khoảng cho phép và model vẫn dự đoán tốt.

Nếu là thuần thống kê thì sẽ có kết luận là 95% confident that chiếc xe 2 năm tuổi sẽ nằm trong khoảng giá từ 35k usd – 45k usd.

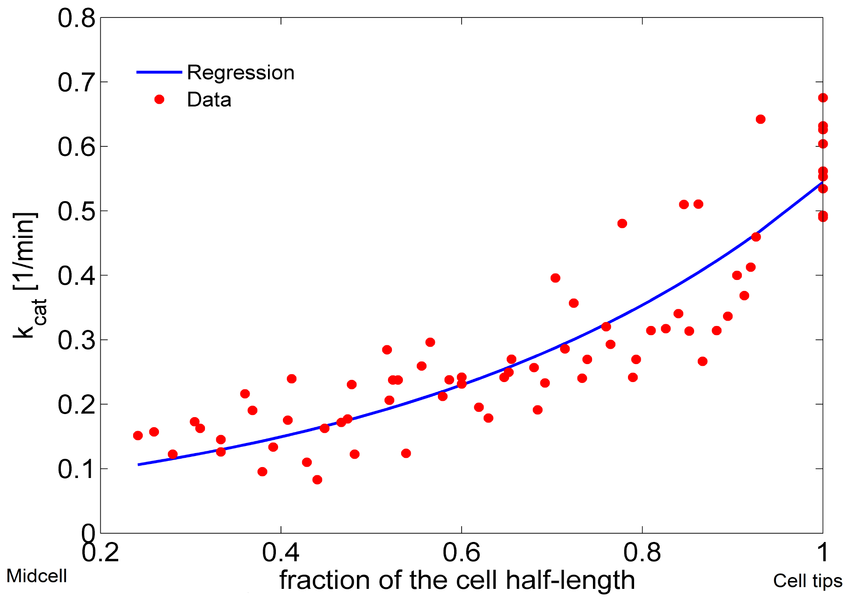
Như vậy, Standard Deviation , hay standard error cũng là một yếu tố rất quan trọng để biết khoảng chính xác của mô hình dự đoán.

**Non – Linear Regression :**

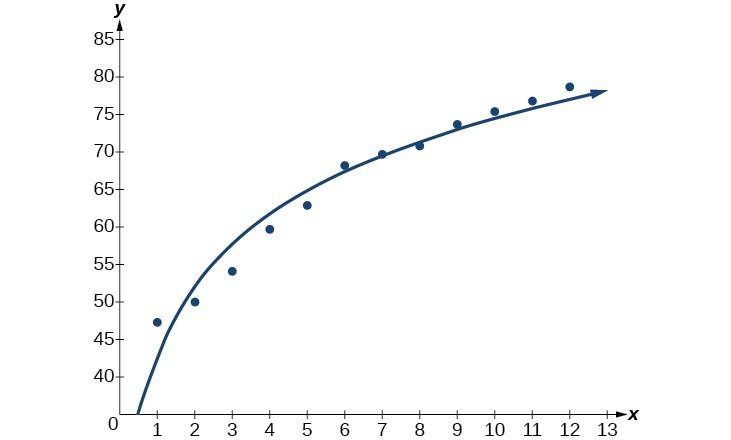
**Quadratic regression : Y = ax^2 + bx +c**



**Exponential regression : y = a\*b^x**



**Logarithmic Regression : y = a logb (x)**



**Các bạn có nhận xét gì về Sự khác nhau của Exponential Regression & Logarithmic Regression không ?**

**Như vậy , đối với phần data validation của thuật toán Regression , chúng ta có những cách :**

* **Cơ bản :** Check Null values , check distribution , check logic
* **Nâng cao :** Check correlation , check causation ( để xem biến độc lập và biến phụ thuộc có tương quan không, có mối quan hệ nguyên nhân kết quả không ).

**Phần Data Evaluation , chúng ta có những cách :**

* Mean , Min , Max , Median ( IQR )
* Standard Deviation , Covariance , Variance ( để kiểm tra độ hội tụ, độ lệch chuẩn của tập dữ liệu)

**Đấy là những phần cơ bản.**

**Còn model Evaluation thì các bạn đã học ở bài trước rồi, Mean Absolute Error và Mean Square Error.**

Vậy , đối với bài toán về Regression cho những biến số ( numerical data ) thì chúng ta sử dụng những phương pháp trên, còn với bài toán Regression cho những biến không phải là số ( categorical , nominal data ) thì chúng ta phải làm như thế nào ?

Nhắc lại một chút về các loại data.

* **Numerical data** : là số thực , từ -∞ đến +∞. Có mối quan hệ hơn thua. Hay được gọi chung là quantitative data. ( dữ liệu định lượng)

Ví dụ : Tiền , tuổi , khoảng cách.

10 tuổi < 20 tuổi < 30 tuổi

$100,000 < $200,000 < $300,000

* **Categorical data / nominal data :** là số mang ý nghĩa đại diện, không có mối quan hệ hơn thua. Hay được gọi chung là qualitative data. ( dữ liệu định tính)

Ví dụ :

Group : 1 , 2 , 3 , 4

Thì group 1 ,2 , 3 , 4 đều mang ý nghĩa đại diện, không có mối quan hệ group 1 < group 2 < group 3 < group 4

Giới tính : Nam hoặc nữ , thường sẽ được label trong tập dữ liệu là 0 , 1.

Thì 0 có nghĩa là Nam , 1 có nghĩa là nữ , hoàn toàn không có ý nghĩa 0 < 1.

**Vì sao phải cần có categorical data ( vì khi xây mô hình , chúng ta không thể dùng chữ , phải đổi chữ về số thì mới xây dựng được model ). Gọi chung là phải mã hóa dữ liệu.**

**Các phương pháp mã hóa ( Encoding ) : ( Phần này practice lab )**

* One hot encoding
* Label encoding
* IQR ( Inter Quartile Range )

**Các phương pháp để validate categorical data :**

* **Check bias :** Đếm frequency , xem frequency có cân bằng không, có dữ liệu nào lặp lại quá nhiều / quá ít không.

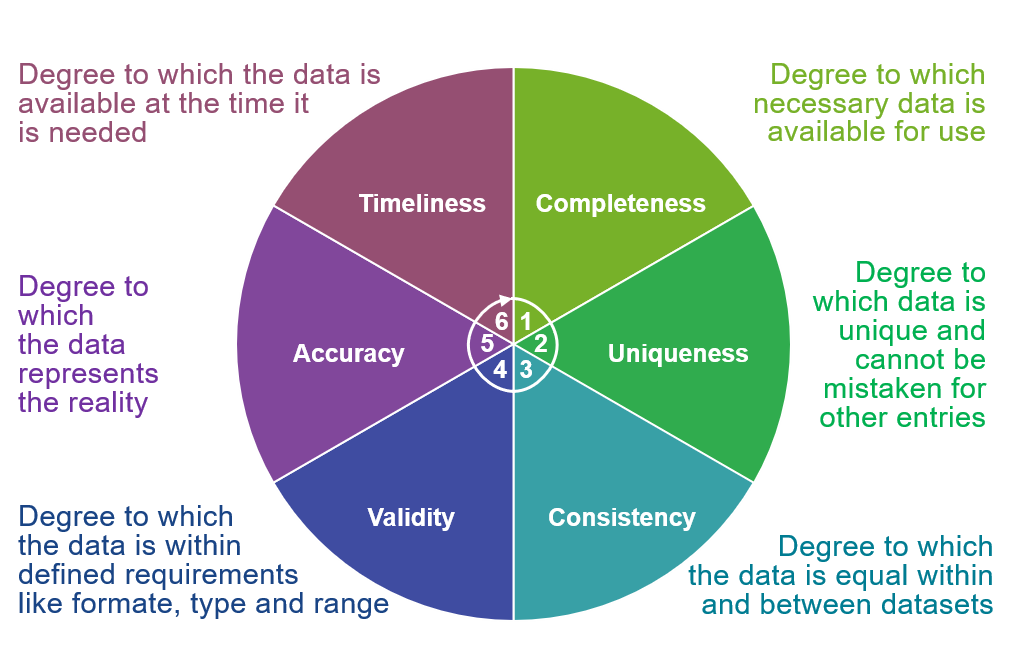
Table

Description automatically generated

* **Tập dữ liệu cân bằng**
* **Dữ liệu có bị nhiễu / lỗi không ? Nếu có thì phải xử lý**

Table

Description automatically generated

* **Tập dữ liệu ko cân bằng, ngoài ra còn có các biến vô nghĩa lọt vào ( 7, 8 , 9 )**
* **Dữ liệu có unique không ? Có bị trùng lặp không ?**
* 

**Về phần data evaluation cho categorical data , chúng ta có các phương pháp**

* **PCA ( giảm chiều dữ liệu )**

A picture containing dark

Description automatically generated

* **LDA**
* **SVD**

**Ví dụ : Dự đoán xem khả năng được cho vay / không được cho vay : những yếu tố ảnh hưởng mạnh là :**

**Thu nhập, thâm niên công việc, lịch sử tín dụng**

**Những yếu tố có thể ảnh hưởng là : Tình trạng hôn nhân , Sức khỏe**

**Những yếu tố khó ảnh hưởng là : Sở thích, địa chỉ , độ tuổi**

**Lưu ý : Trong Machine Learning , data quality rất quan trọng. !!!**

**Keep in mind : Trash in = trash out**

**Part 2 . Probability**

**Xác suất là gì ? Vì sao phải học xác suất ?**

Bayesian Machine Learning là một nhánh của Machine Learning , trong đó Machine Learning và lý thuyết xác suất có một sự liên hệ rất khăng khít. Có những bài toán classification mà máy tính cần phải áp dụng xác suất để giải quyết : Ví dụ , xác suất khách hàng sẽ chọn 1 item tiếp theo sau khi đã chọn 1 item ( Recommendation System , Related Search) . Hoặc chuẩn đoán khả năng bị bệnh của bệnh nhân. Ví dụ : Xác suất của bệnh nhân có thể tử vong do COVID 19 là cao nếu họ bị : Tiểu Đường , Ung Thư , Béo Phì ....

Gồm những nội dung : biến ngẫu nhiên, xác suất đồng thời, xác suất biên, xác suất có điều kiện, và quy tắc Bayes

1. **Biến ngẫu nhiên**

* **Biến ngẫu nhiên :** (random variable) x là một đại lượng dùng để đo những đại lượng không xác định, là outcome của một thí nghiệm.

**Ví dụ : Tung 1 đồng xu ,ta sẽ được ngẫu nhiên 1 biến x trong 2 kết quả là mặt sấp, hoặc mặt ngửa.**

+ Có 2 loại biến ngẫu nhiên : biến ngẫu nhiên rời rạc ( discrete ) hoặc liên tục ( continous)

* Biến ngẫu nhiên rời rạc sẽ lấy giá trị trong một tập hợp cho trước, tập giá trị nó là rời rạc và đếm được ( ví dụ biến x khi tung đồng xu sẽ có giá trị ngẫu nhiên trong tập hợp đã biết trước là (sấp , ngửa ), 1 trong 6 mặt của xúc xắc ,...
* Xác suất của các giá trị này có tổng = 1 ( Xác suất của mặt sấp + xác suất mặt ngửa = 100% )

1. **Biến ngẫu nhiên liên tục**

Biến ngẫu nhiên liên tục lấy các giá trị là tập con của các số thực, ví dụ : giá nhà đất ở Sài Gòn là 1 tập rất nhiều những con số. Những giá trị này có thể là hữu hạn, ví dụ thời gian làm bài của mỗi thí sinh trong một bài thi 180 phút, hoặc vô hạn, ví dụ thời gian để chiếc xe bus tiếp theo tới

* Xác suất để đầu ra bằng chính xác một giá trị nào đó là gần như bằng 0 ( ví dụ xác suất để xe bus tiếp theo 10 phút nữa sẽ tới , chỉ có thể ước lượng độ tập trung xác suất tại lân cận điểm nào đó là cao hay thấp ( ví dụ xác suất để xe bus tiếp theo tới trong vòng 15 -20 phút nữa ).

**Với biến ngẫu nhiên rời rạc,**p(x)**được hiểu là *mật độ xác suất* tại**x

*hàm mật đô xác suất* (probability density function - pdf)

A picture containing graphical user interface

Description automatically generated

### Xác suất hợp ( Joint Probability )

là xác suất của hai biến cố cùng xảy ra. Xác suất hợp của A và B được ký hiệu P(A,B)

### Ví dụ :

### X : là xác suất thí sinh được điểm x môn Toán trong kỳ thi THPT QG.

### Y : là xác suất thí sinh được điểm y môn Lý trong kỳ thi THPT QG.

(xác suất ở đây thực ra là tỉ lệ giữa số học sinh đạt x điểm môn toán và toàn bộ số học sinh)

P(X,Y) là xác suất thí sinh được x điểm môn Toán và y điểm môn Lý

Xét từng biến rời rạc ( vì điểm chỉ có từ 0 – 10)

P(x=0) , tức 0 điểm là rất thấp

P(x=10) , tức 10 điểm cũng là rất thấp

Tuy nhiên xác suất P(x=4, 5 , 6 ,7 ) , tức xác suất để được 4 , 5 ,6 ,7 điểm Toán là cao.

Tương tự,

Xác suất để P(y=0) , tức 0 điểm cũng là rất thấp

Xác suất để P(y=10) , tức 10 điểm cũng là rất thấp.

Tuy nhiên , xác suất P(y = 4, 5 , 6 ,7 ) , tức xác suất để được 4 , 5 , 6 ,7 điểm Lý là cao

Vì vậy P(X,Y) cũng có mật độ xác suất rất khác nhau.

Ví dụ P(0,10) hoặc P(10,0) là rất thấp ( Xác suất để thí sinh 10 điểm môn toán được 0 điểm môn lý hoặc ngược lại )

Và xác suất để P(5,5) hoặc P(5,6) hoặc P(6,5) là cao nhất ( Xác suất để thí sinh 5 điểm môn toán cũng được 5 điểm môn Lý...)

1. **Xác suất biên**

 là xác suất của một biến cố mà không quan tâm đến các biến cố khác

1. **Xác suất có điều kiện**

**Ví dụ :** biết rằng bây giờ đang là tháng 7, tính xác suất để nhiệt độ hôm nay cao hơn 30 độ C

Biết rằng bạn học trung bình 10 tiếng mỗi tuần , tính xác suất để thí sinh thi hơn 8 điểm

Xác suất có điều kiện (conditional probability) của một biến ngẫu nhiên x biết rằng biến ngẫu nhiên y được ký hiệu là p(x|y) , y\* là giá trị được biết  
Text, letter

Description automatically generated

**P(B|A) = P(A và B) / P(A)**

*Hay*

**P(B|A) = P(A∩B) / P(A)**

1. **Quy tắc Bayes (Quy tắc này rất quan trọng trong Machine Learning!)**

**Ta có công thức** xác suất có điều kiện p(x/y) hoặc p(y/x)

Text, letter

Description automatically generated Text, letter

Description automatically generated

Dựa vào mối quan hệ trên , vì p(x,y) = p(y,x) ta có



Biến đổi 1 chút , ta thu được:

Text, letter

Description automatically generatedText

Description automatically generatedText

Description automatically generated

Đây là 3 công thức Bayes, vậy ý nghĩa nó là gì trong Machine Learning ?

Để tính xác suất y biết x , p(y/x) , ta cần phải biết xác suất x biết y, p(x/y) và p(x) , hoặc để tính xác suất x biết y , p(x/y), ta cần phải biết xác suất y biết x , p(y/x) và p(y), và việc tính trực tiếp điều này rất phức tạp.

* Vì vậy , chúng ta hoàn toàn có thể tính được xác suất một biến, nếu đã biết xác suất của biến kia và ngược lại.

Dựa trên định luật Bayes , chúng ta có 2 mô hình phân lớp Bayes nổi tiếng :

## **Naive Bayes Classifier**

Naive Bayes là một thuật toán phân loại cho các vấn đề phân loại nhị phân (hai lớp) và đa lớp. Thuật toán Naive Bayes tính xác suất cho các yếu tố, sau đó chọn kết quả với xác suất cao nhất.

Sở dĩ gọi là Naive ( ngây thơ) vì thuật toán Naive Bayes assume là các yếu tố (features ) đều độc lập về mặt thống kê, tức là mỗi các feature đều độc lập và công bằng.

## **Gaussian Naive Bayes**

Gaussian Naive Bayes là thuật toán Naive Bayes nhưng được kếp hợp với mô hình phân bổ ( normal distribution ). Mô hình này được sử dụng cho các biến liên tục ( continuous variables) vì các biến liên tục sẽ có sự phân bổ theo từng điểm ( mật độ tập trung xác suất ).

**Practice** : [Link](https://colab.research.google.com/drive/1zSX5hICpXXNgbRtPwF7gMW5S0R8DTt6U?usp=sharing)