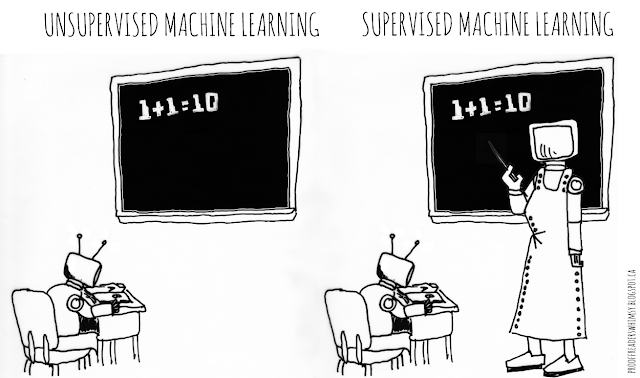
**Class 10 – Unsupervised Learning : K-means Clustering**

1. **Lý thuyết**

Như vậy , chúng ta đã hoàn thành xong phần Supervised Learning. Từ nay về sau chúng ta sẽ đi đến phần Unsupervised Learning.

Nhắc lại một chút , Unsupervised Learning là gì ?

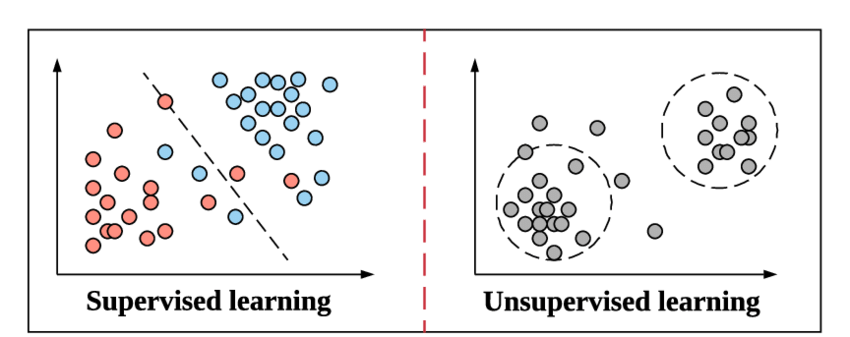
Unsupervised Learning là học không giám sát, không giám sát là vì chúng ta không biết được đầu ra ( output) trông như thế nào. Thuật toán unsupervised learning sẽ dựa vào dữ liệu mà tìm ra một mô hình hay cấu trúc nào đó ẩn trong dữ liệu cho bạn để biến nó thành những thứ có nghĩa.



Ở supervised learning , chúng ta có hàm y = f(x) , với y là dữ liệu đầu ra , f là model cần tìm và x là dữ liệu đầu vào. Y và X chúng ta đã có.

Thì ở unsupervised learning , lúc này chúng ta chỉ có ? = f(x) , vì chúng ta không biết đầu ra như thế nào , f là model và x là dữ liệu đầu vào.

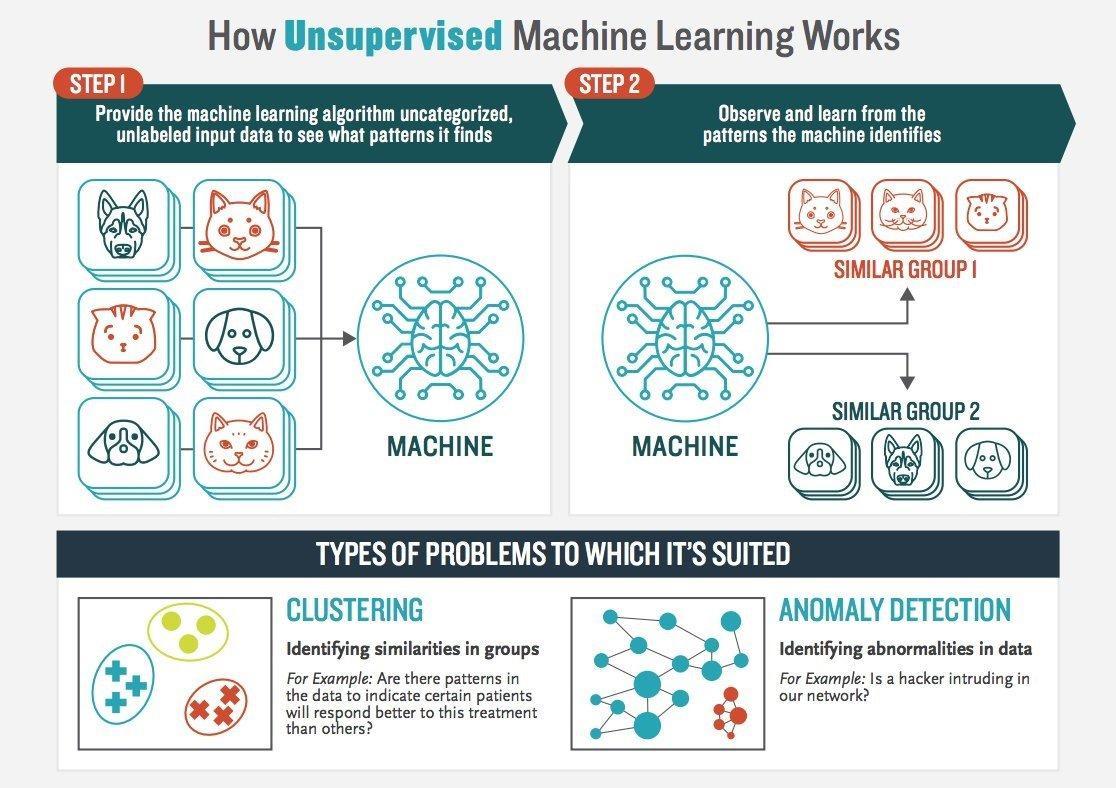
Vì vậy, đối với thuật toán unsupervised learning, không có chuyện đúng sai, chỉ có kết quả đầu ra có ý nghĩa hay không có ý nghĩa, nếu không có ý nghĩa thì chúng ta cần phải kiểm tra lại model.



Unsupervised Learning được áp dụng với những bài toán :

* Clustering : phân nhóm dựa trên những đặc trưng riêng của dữ liệu
* Association : khám phá ra quy luật dựa trên dữ liệu cho trước

Ví dụ : Phân nhóm khách hàng, Recommendation sản phẩm



Có 3 nhóm thuật toán chính :

* K Means Clustering : ( ngoài ra còn có các biến thể K modes , Kmedians ,...)
* Hierarchical Clustering :
* Probabilistic Clustering :

Buổi học ngày hôm nay, chúng ta sẽ đi vào thuật toán Clustering , và cụ thể là Kmeans clustering.

1. **Thuật toán**
2. **K Means Clustering**

Phân cụm là phương pháp tập hợp các điểm ở gần nhau trong một không gian nào đó ( không gian có thể là 2D , 3D thậm chí ND)

Ví dụ : 3 cụm dữ liệu với dữ liệu 2DChart, scatter chart

Description automatically generated

Và trong thuật toán Clustering , số K ( tức là tâm cụm ) ban đầu chúng ta khai báo rất quan trọng đối với kết quả đầu ra, vì nếu K quá nhiều hay quá ít thì cũng có thể làm cho kết quả đầu ra không có ý nghĩa ( chúng ta hay hiểu là overfit và underfit)

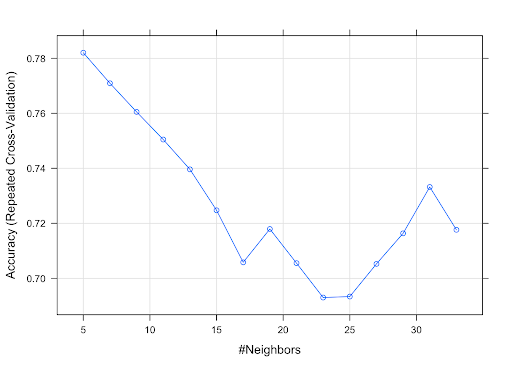
Đối với thuật toán Unsupervised , việc có dữ liệu đầu vào đồng đều , chọn thuật toán phù hợp và cơ số k phù hợp nó quan trọng hơn việc chúng ta adjust các chỉ số để cho ra output có ý nghĩa, vì vậy, phần toán học của thuật toán như hàm tổng quan , hàm mất mát nó không quan trọng bằng việc chúng ta chọn được cơ số K phù hợp và thuật toán phù hợp.

Vì vậy , tiết học này chúng ta sẽ không bàn đến những phân tích toán học như phân tích hàm mất mát hay làm thế nào để tối ưu hàm mất mát, dù những thứ này cũng quan trọng nhưng là chỉ khi bạn đã thuần thục trong việc xây dựng mô hình và chọn ra cơ số K phù hợp ( K fold testing). Các bạn có thể search đọc thêm trên mạng với keyword là Hàm Mất Mát và bài toán tối ưu của K Means clustering.

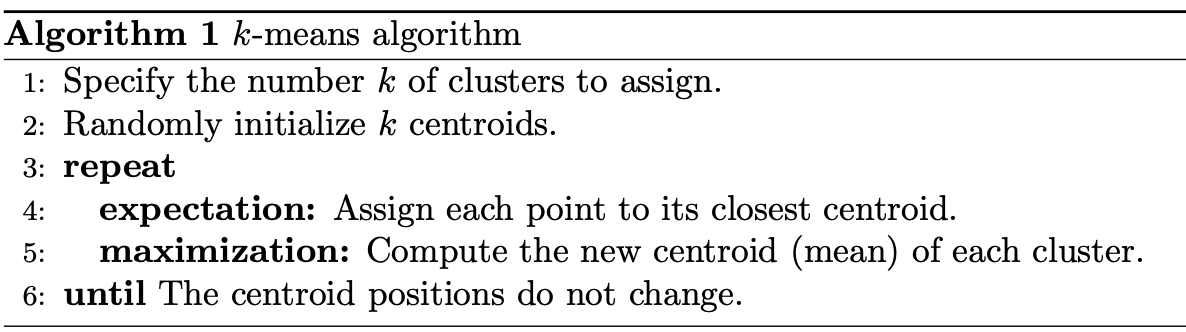
**K Fold testing**

Tương tự với K fold test validation của KNN , thuật toán K Means cũng có chọn ra cơ số K phù hợp cho thuật toán. Nếu ở KNN là chọn K Nearest Neighbours thì ở K Means cơ số K chính là số tâm cụm trong dữ liệu.

ở KNN, chúng ta chọn tâm cụm tại nơi xuất hiện elbow



Thì ở trong K Means , chúng ta chọn cơ số K phù hợp và hệ thống sẽ tự động tính toán khoảng cách từ tâm cụm đến các điểm trong từng cluster



Demo : [Link](https://colab.research.google.com/drive/1ALwSRstDyjXZntfhrFkuaJ5T2GR0Pn4-?usp=sharing)

1. **Hierarchical Clustering**

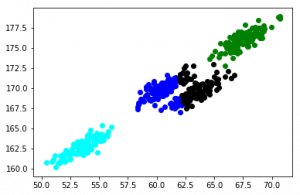
Vì thuật toán K Means có những hạn chế nhất định , đó là chúng ta phải xác định được cơ số K hợp lý , ngoài ra các cluster sẽ được sắp xếp sao cho có phân bổ đồng đều.

* Thuật toán Hierachical Clustering thì khắc phục điểm yếu về cơ số K , chúng ta không cần phải xác định cơ số K trước khi chạy model.

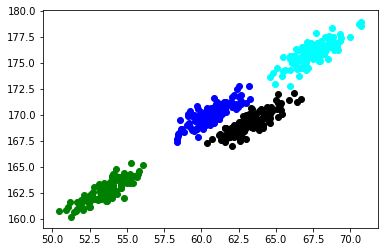
Nguyên lý hoạt động :

1. **Probabilistic Clustering**

Tương tự như Hierarchical Clustering , Probalilistic Clustering là thuật toán được sử dụng để khắc phục những hạn chế của K Means, đó là vấn đề phân bổ đồng đều và nhóm data của mỗi cluster. Ta có thể thấy thuật toán K Means sẽ cố gắng nhóm data theo dạng hình tròn, các data label sẽ tập trung xung quanh các tâm cụm với những khoảng cách, dẫn đến đôi khi các cluster sẽ không được phân cụm chính xác. Ví dụ như hình sau.

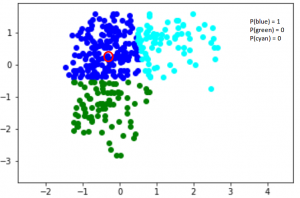


Nếu k = 3 thì K Means có thể phân được cluster 1 2 và 3 vì khoảng cách của chúng khá rõ rệt, tuy nhiên chúng ta lại nhìn thấy rõ ràng ở đây có 4 cluster, nhưng với k = 4 thì K Mean lại phân bố cluster 2 và 3 không đúng.



Vì vậy, chúng ta cần một thuật toán có thể phân cụm chính xác với tập dữ liệu này, đó là phân cụm dựa trên xác suất, chứ không phải dựa trên khoảng cách.

Để giải thích một cách dễ hiểu phân cụm dựa trên xác suất thì chúng ta sẽ có ví dụ sau :



Với hình ở trên, rõ ràng data lebel được khoanh đỏ nằm trong cluster màu xanh dương, vì vậy xác suất của cái điểm đó nằm trong cluster blue là 100% , xác suất điểm đó nằm trong cluster green và cyan là 0%.

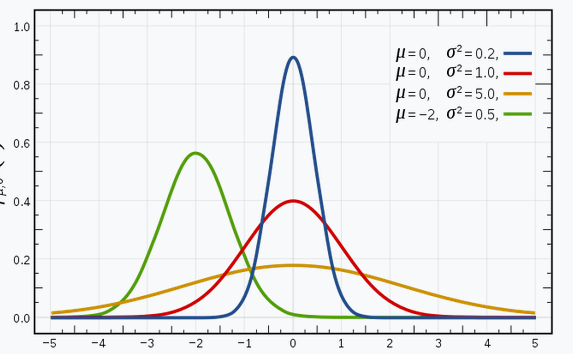
Chart, scatter chart

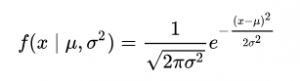
Description automatically generated

Tương tự, với điểm trên thì xác suất để nó nằm trong cluster green là 0%, và blue là 80% và cyan là 20%.

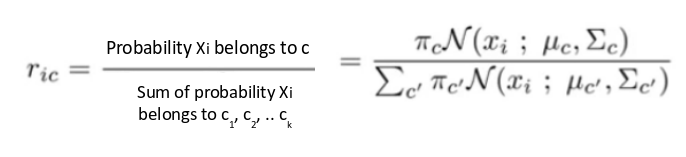
* Assign điểm đó vào cluster blue.

Còn làm cách nào để tính ra được 80% và 20% thì thuật toán sử dụng phân bổ Gaussian



Với hàm số tổng quan là 

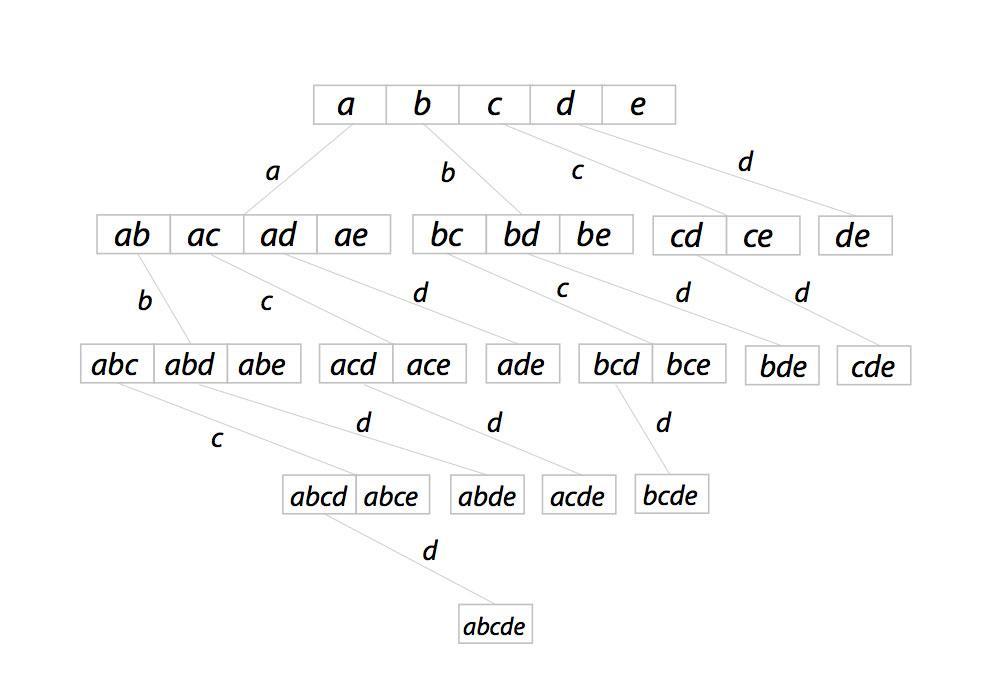
Với  μ is the là khoảng cách trung bình (mean) và σ2 là phương sai ( variance). Đương nhiên đó chỉ là giá trị 1D, đối với hàm có nhiều features , tức nhiều giá trị đầu vào thì sẽ có công thức phức tạp hơn. Ở đây chúng ta chỉ nói sơ qua về nguyên lý chứ không đi sâu vào Toán , nên nếu các bạn có hứng thú tìm hiểu thêm về phần này thì có thể search thêm.



Demo : [Link](https://colab.research.google.com/drive/1ALwSRstDyjXZntfhrFkuaJ5T2GR0Pn4-?usp=sharing)

1. **Apriori & ECLAT**

Apriori là thuật toán dùng để khám phá những quy luật của dữ liệu



Nhìn vào quá khứ và phân tích 1 việc xảy ra thì sẽ có tỉ lệ bao nhiêu phần trăm sự việc tiếp theo sẽ xảy ra

Chúng ta hiện chưa học đến phần này, bạn có thể đọc thêm tại : [Link](http://www.ungdung.khoa-hnvd.com/Hoc_thuat/Apriori.html)

1. **Eclat**

Tương tự như Apriori , Eclat là một thuật toán dùng để khai phá dữ liệu để tìm ra các quy luật, được phát triển dựa trên sự cải tiến của thuật toán Apriori

Bạn có thể đọc thêm tại : [Link](https://www.slideshare.net/wanaezwani/apriori-and-eclat-algorithm-in-association-rule-mining)

Những thuật toán này sẽ hữu dụng khi các bạn đụng đến phần data mining.