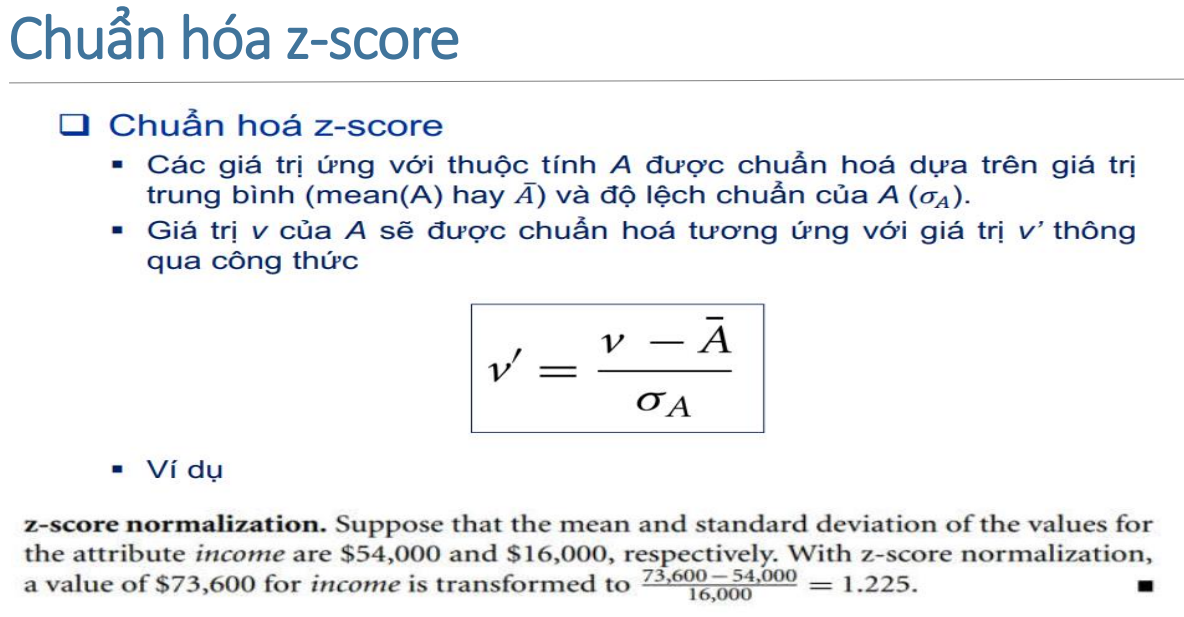
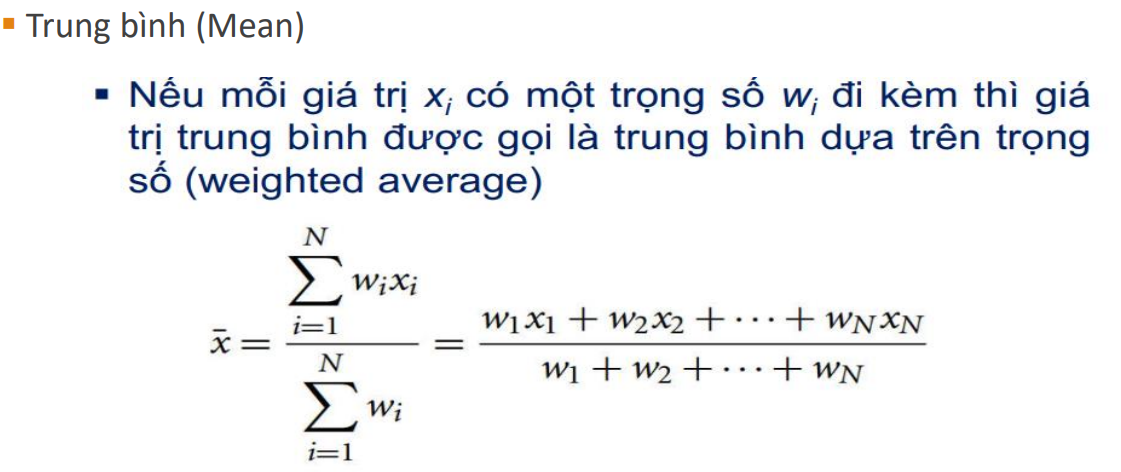
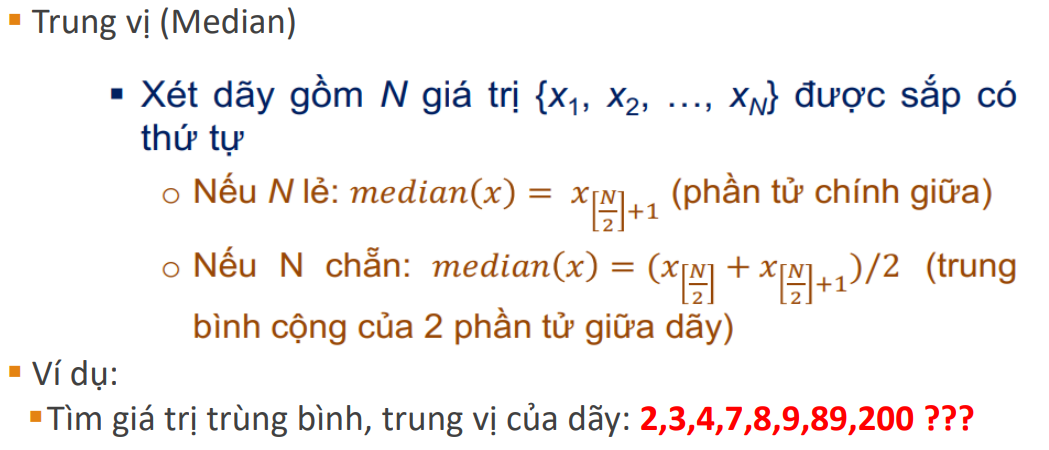
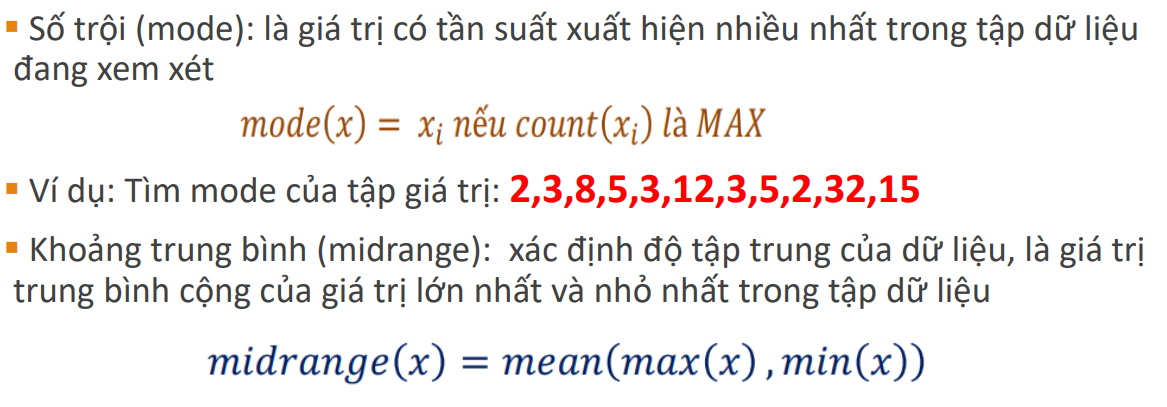


Chuẩn hóa z-core là số độ lệch chuẩn mà nó nằm trên hoặc dưới mức trung bình của tổng thể.

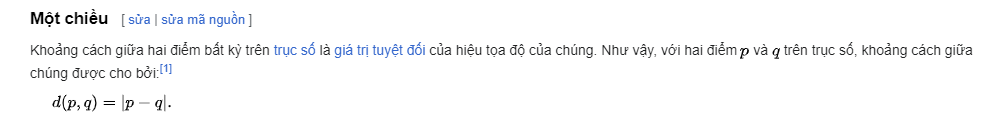






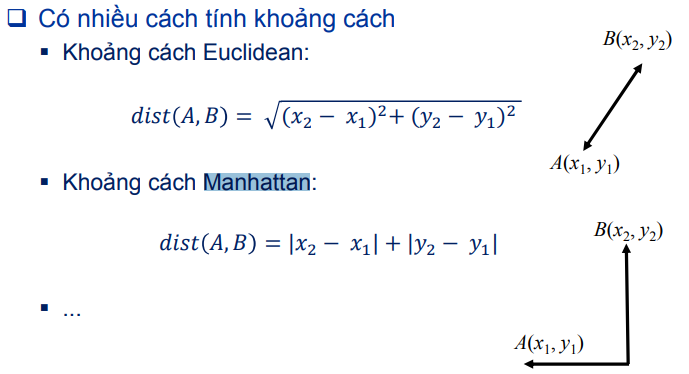


Tính khoảng cách euclid



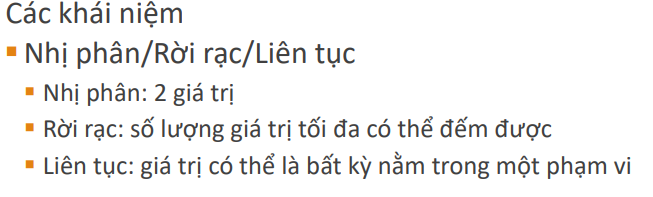
Tính khoảng cách manhattan

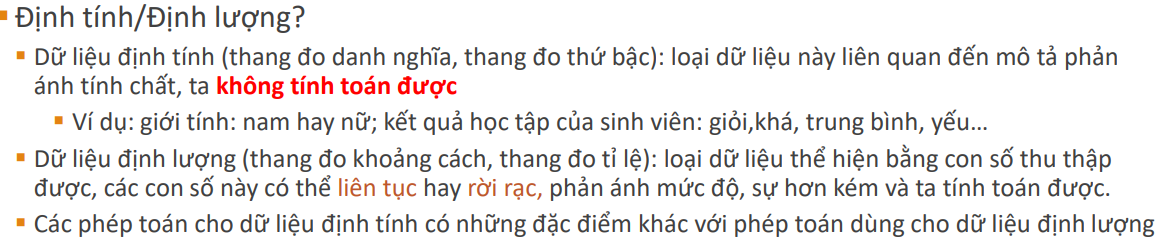
là 1 dạng khoảng cách giữa 2 điểm trong**không gian Euclid và hệ tọa độ Descartes**và được tính bằng tổng **chiều dài của hình chiếu của đường thẳng nối 2 điểm này trong hệ trục tọa độ Descartes .**

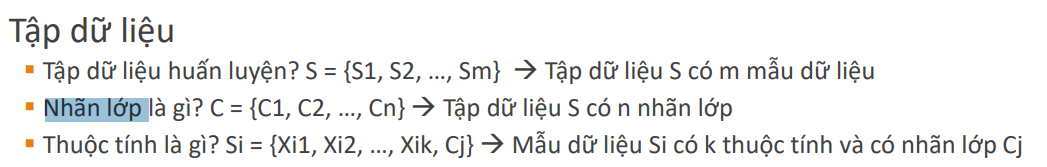


Dữ liệu, thuộc tính, nhãn lớp là gì?

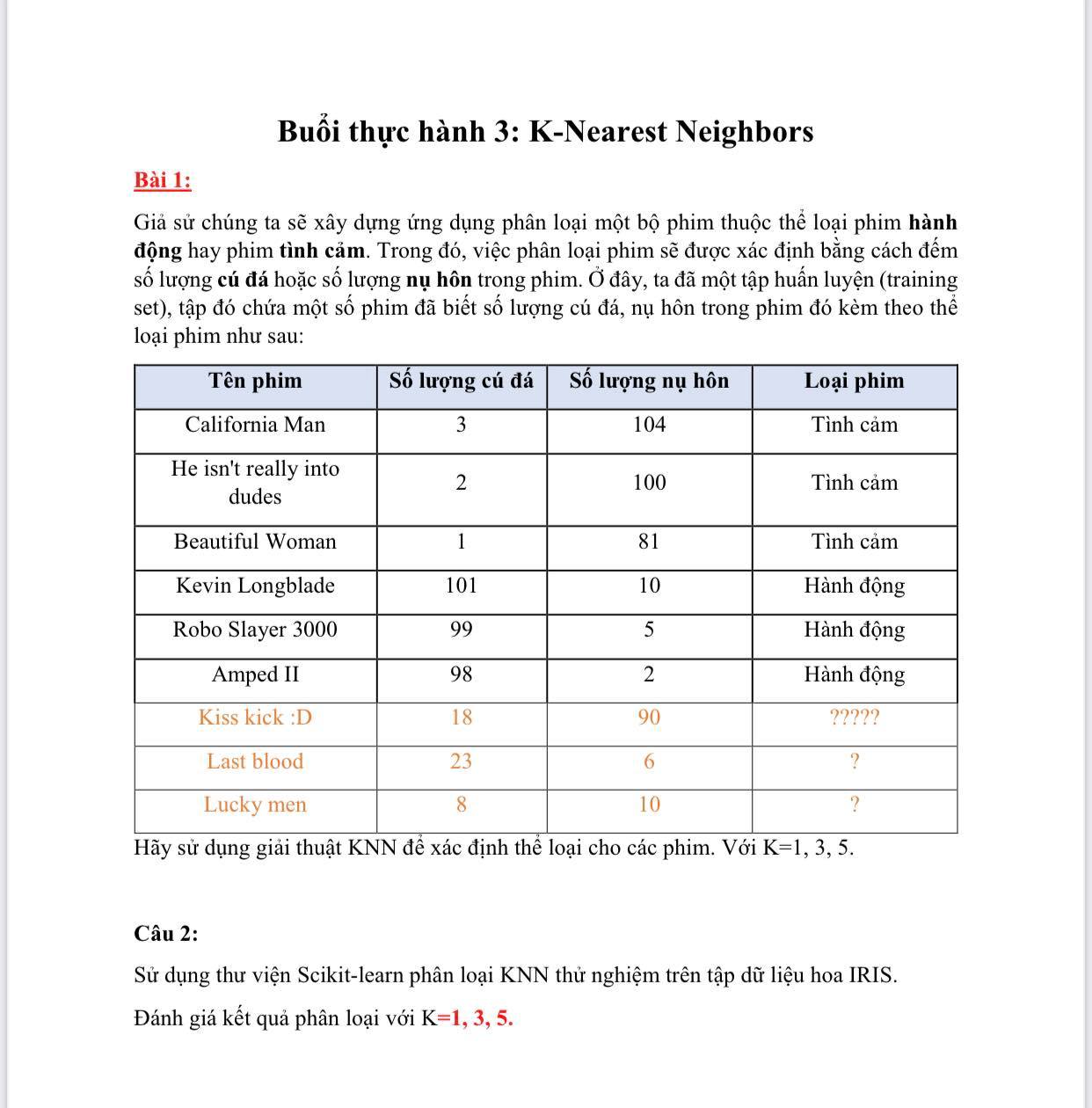
Dữ liệu là thông tin dưới dạng ký hiệu,chữ viết, chữ số, hình ảnh, âm thanh hoặc dạng tương tự. bao gồm những mệnh đề phản ánh thực tại, thường đo đạc hay quan sát về một đại lượng biến đổi. Các mệnh đề đó có thể bao gồm các số, từ hoặc hình ảnh

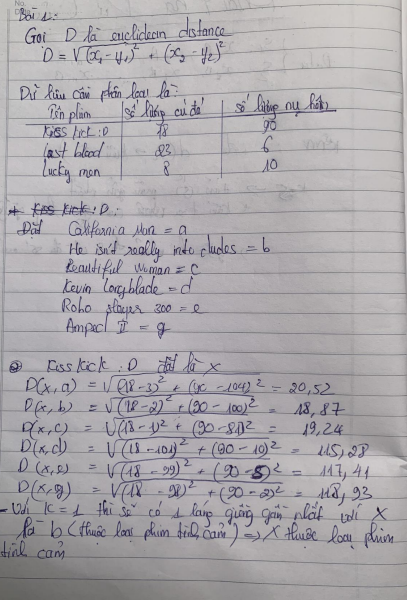


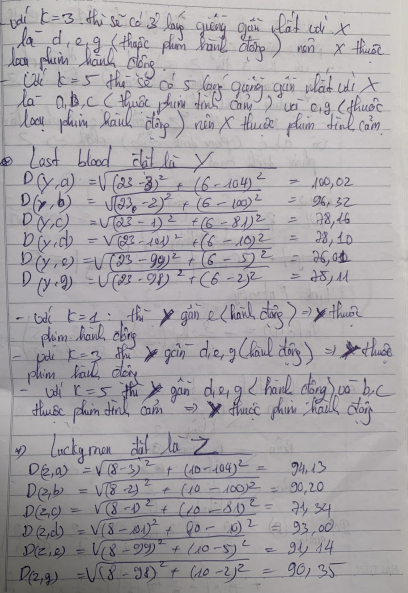


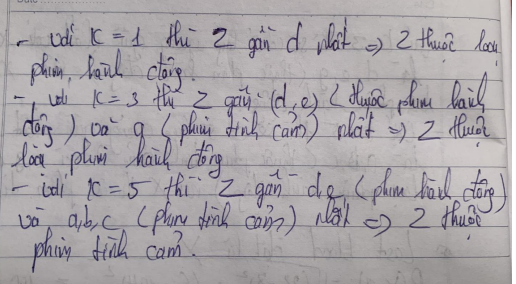


Phân lớp KNN

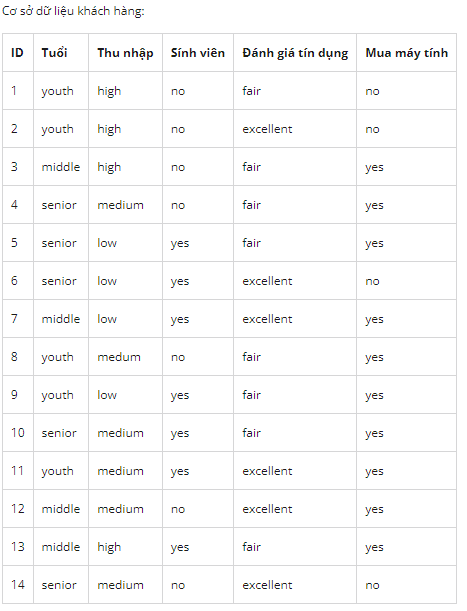


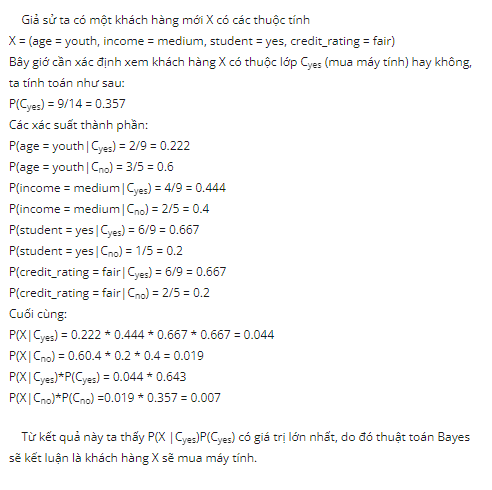




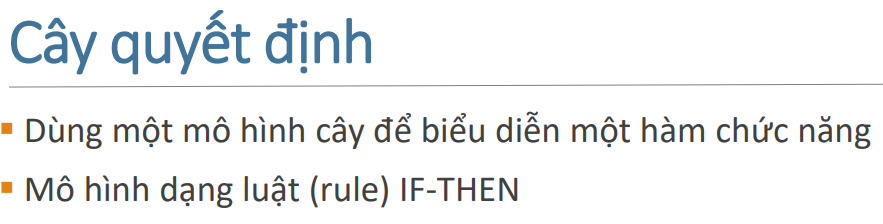


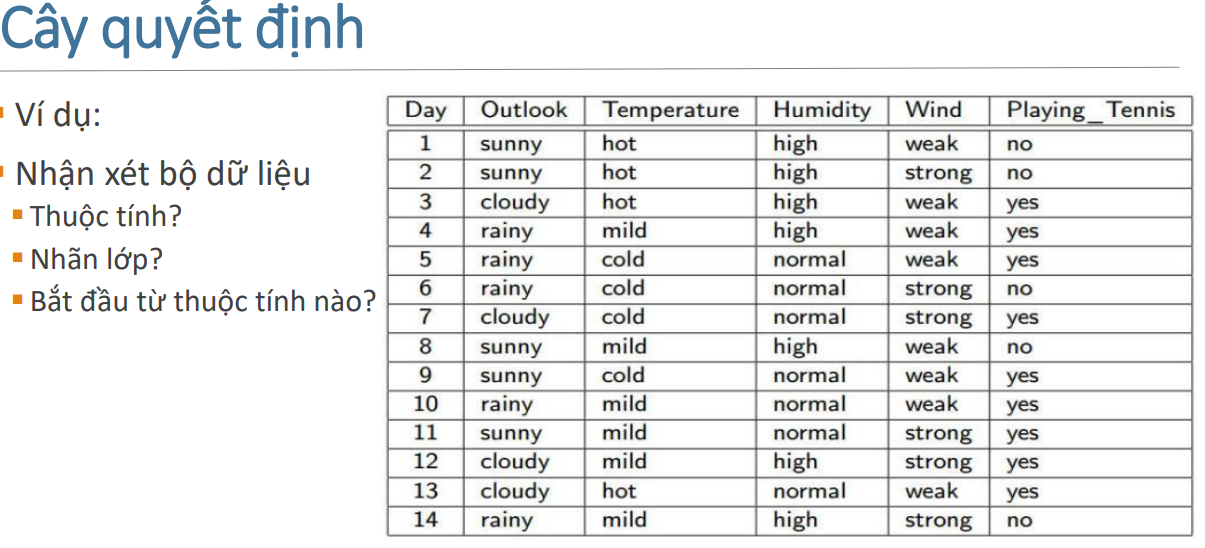
Phân lớp Naive bayes

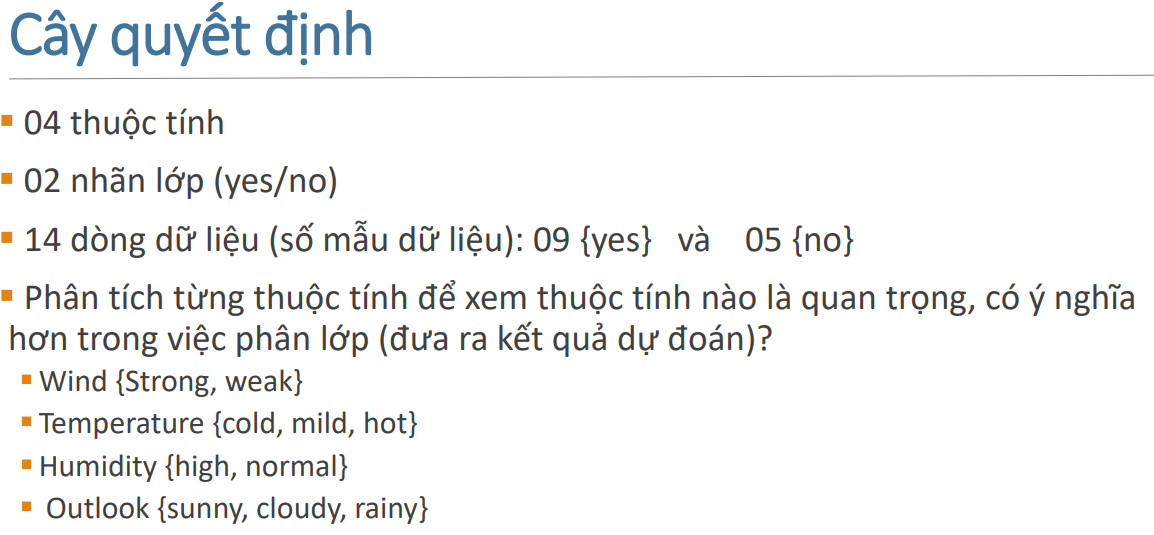


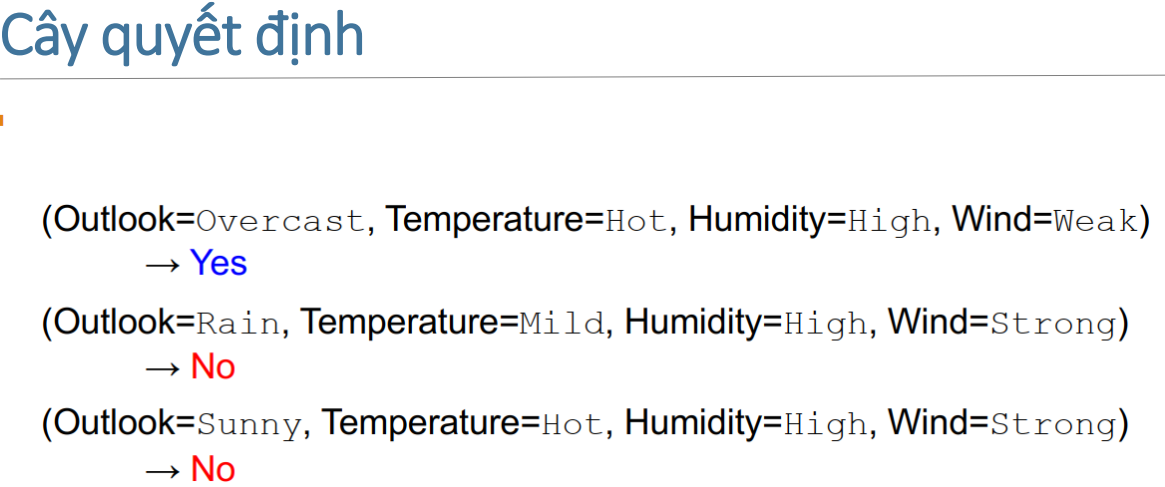


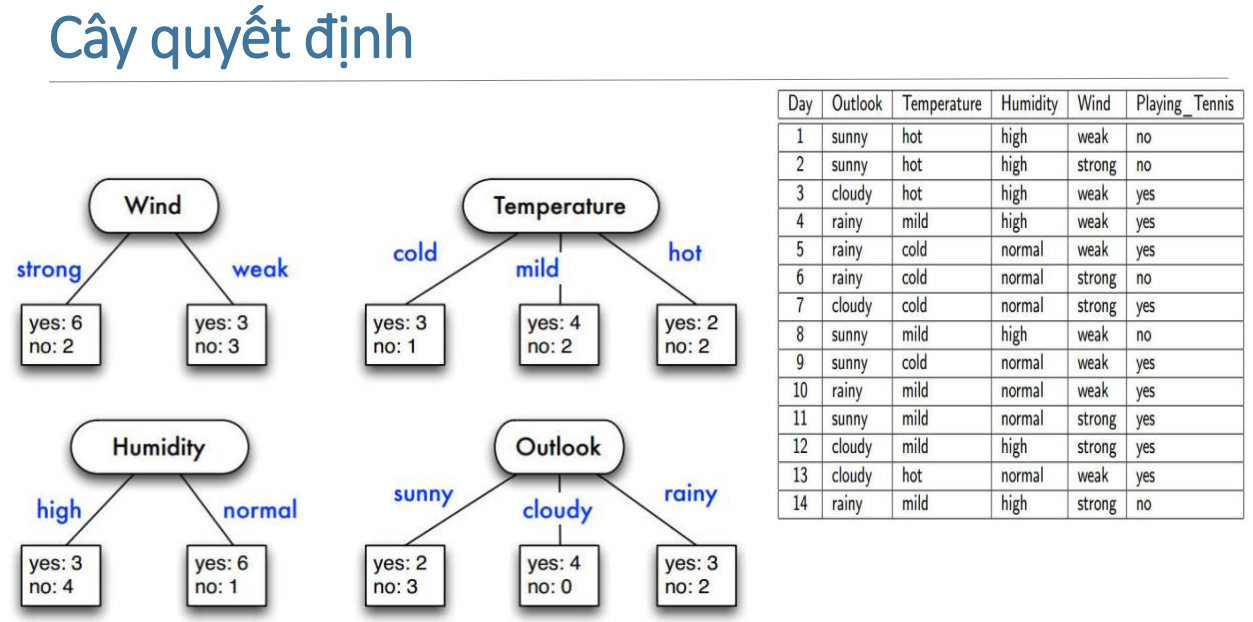
Phân lớp cây quyết định

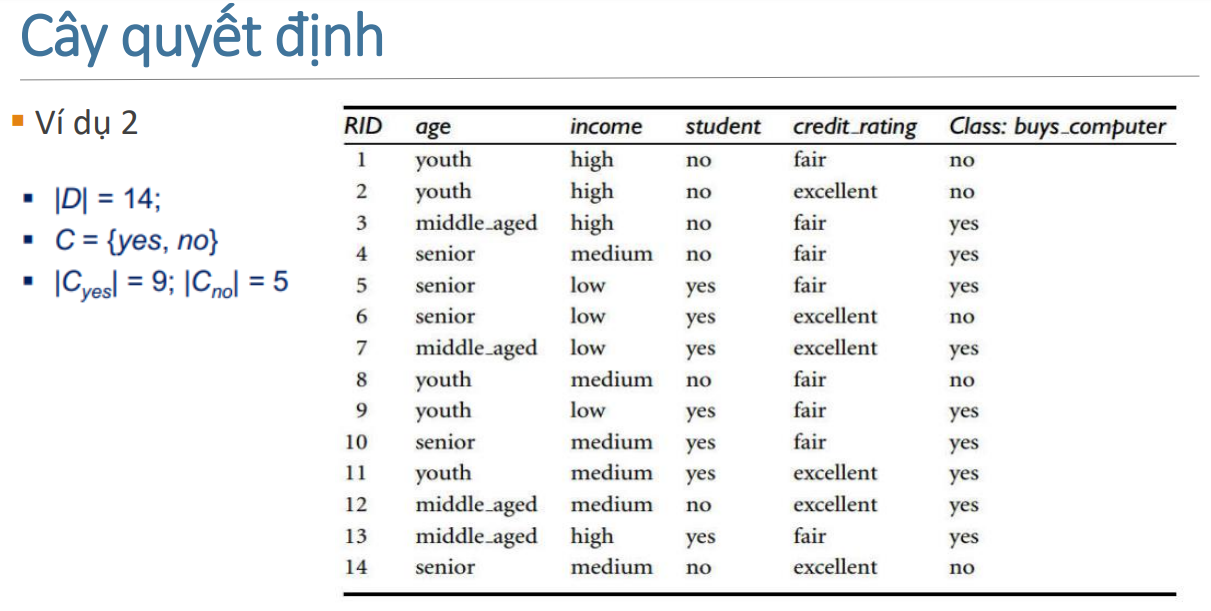


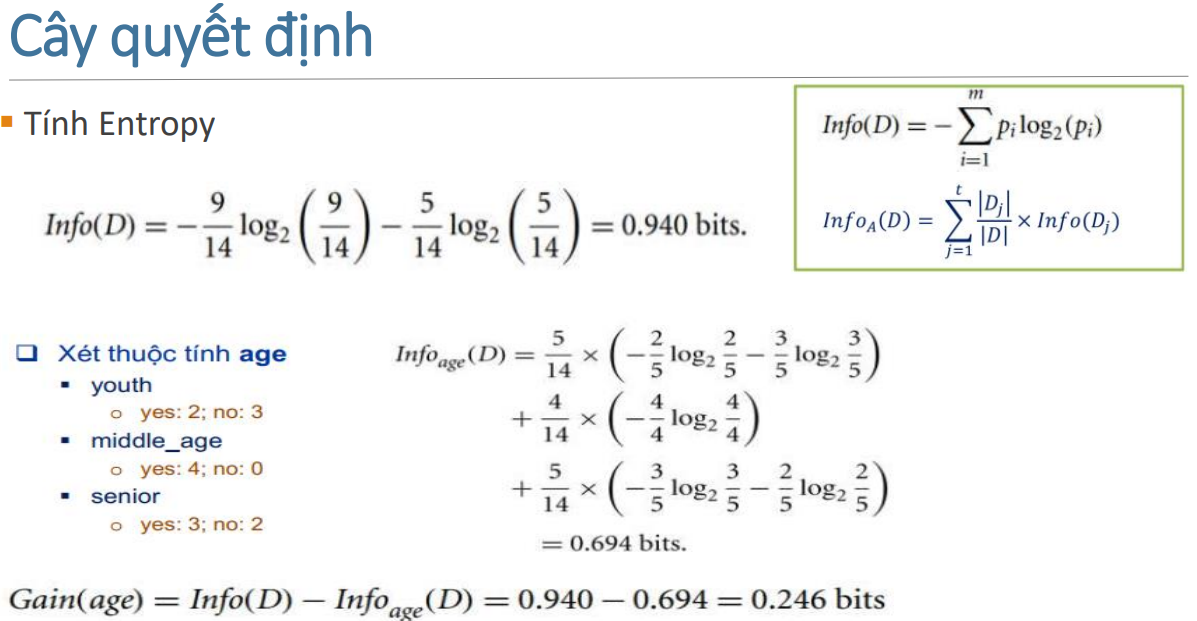


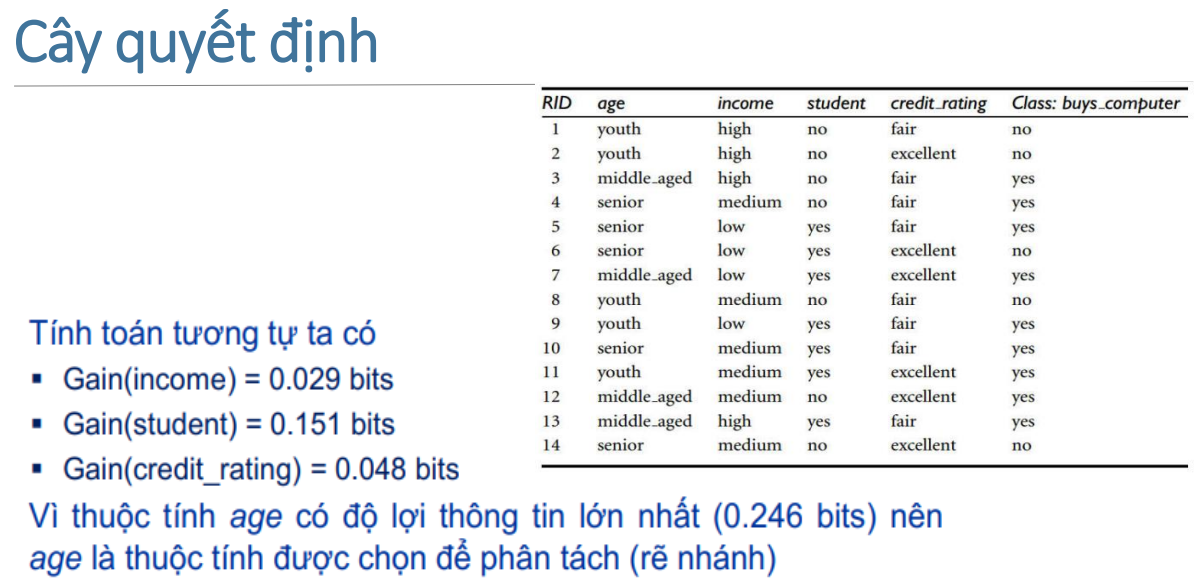


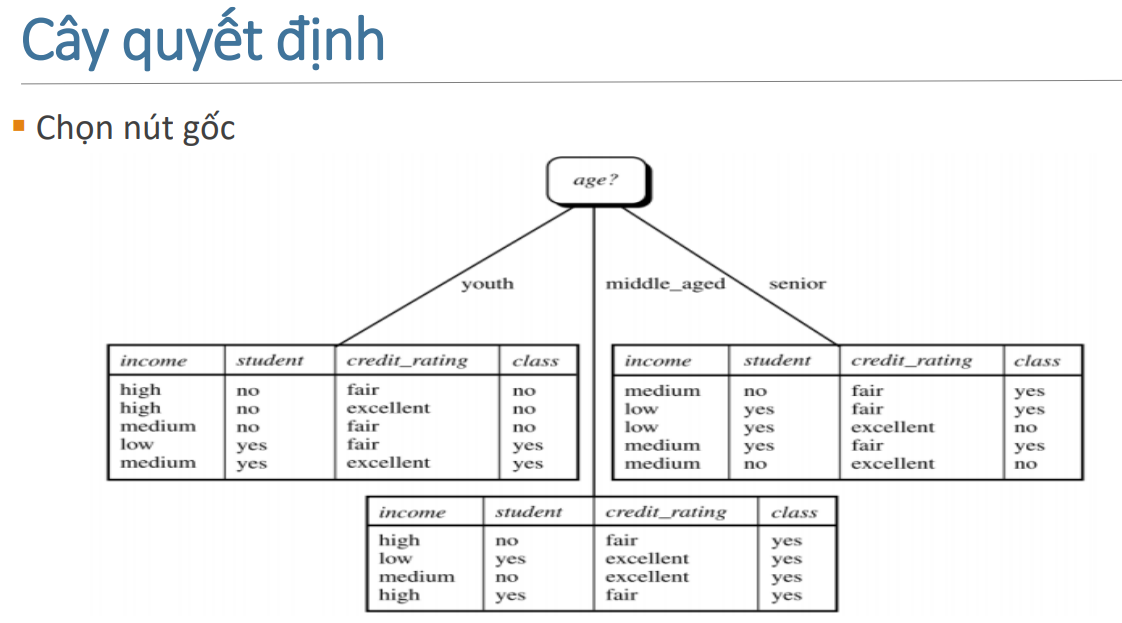


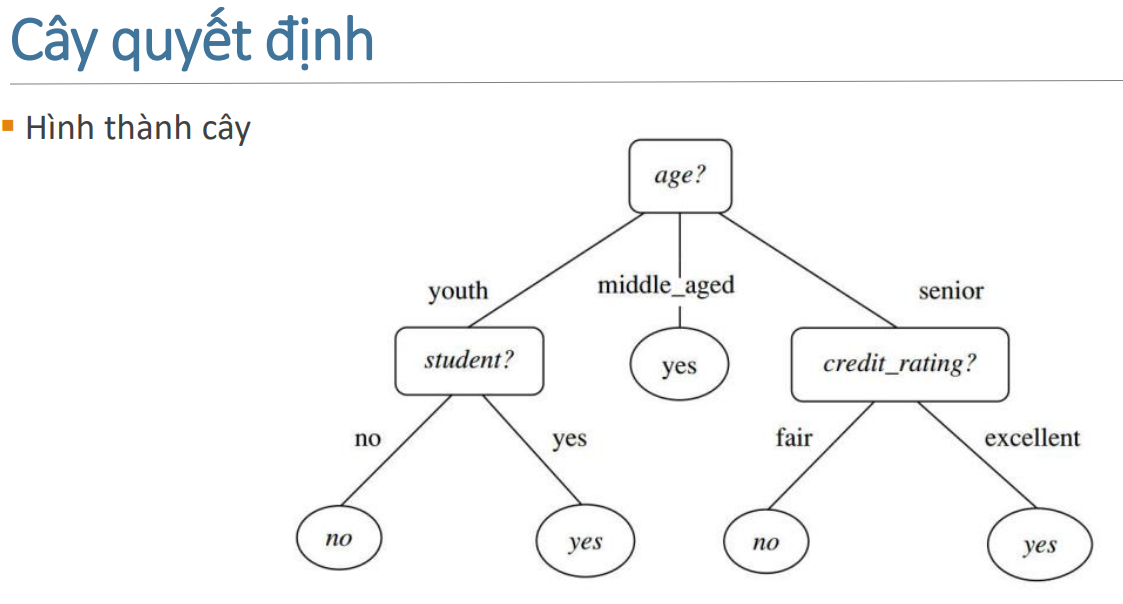




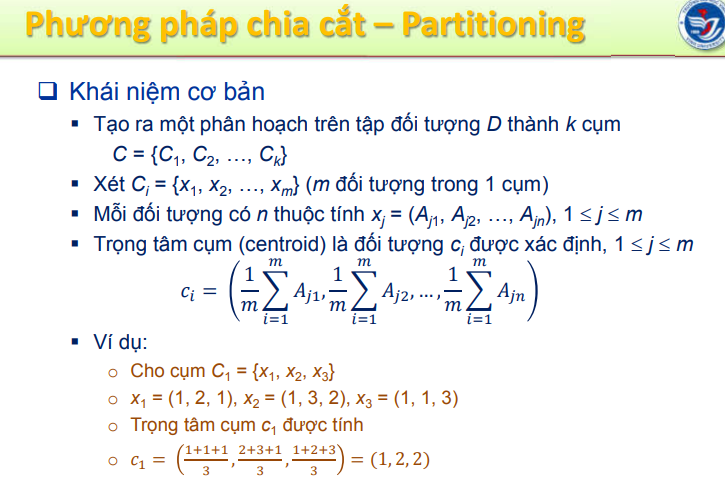


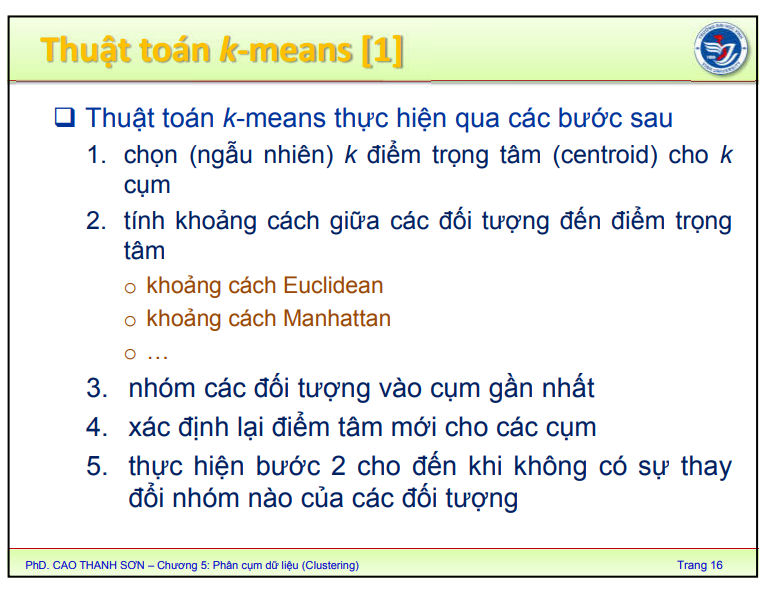


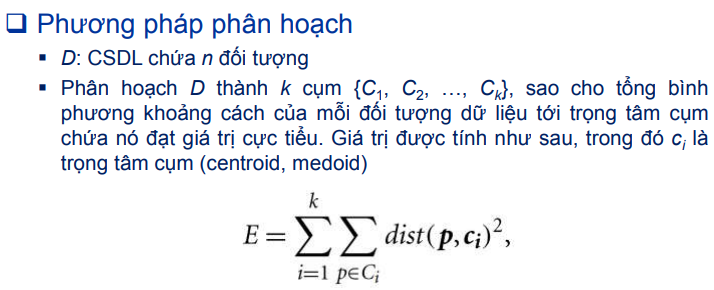


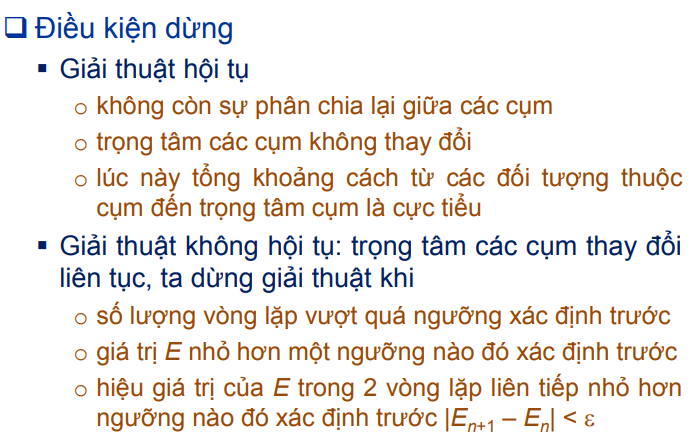


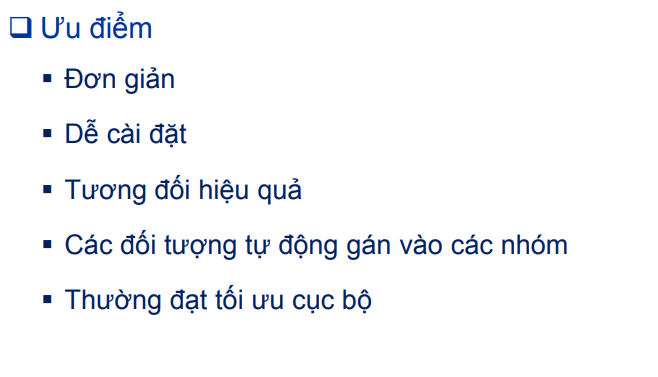
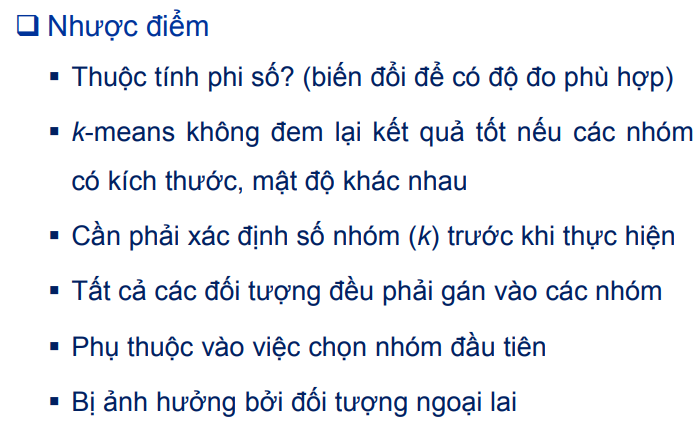
Phân cụm K-means

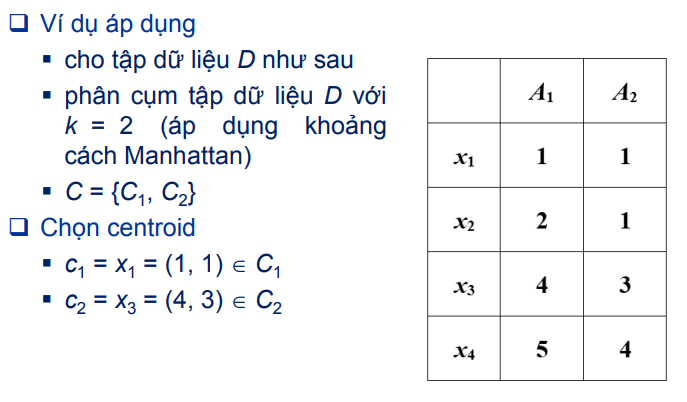


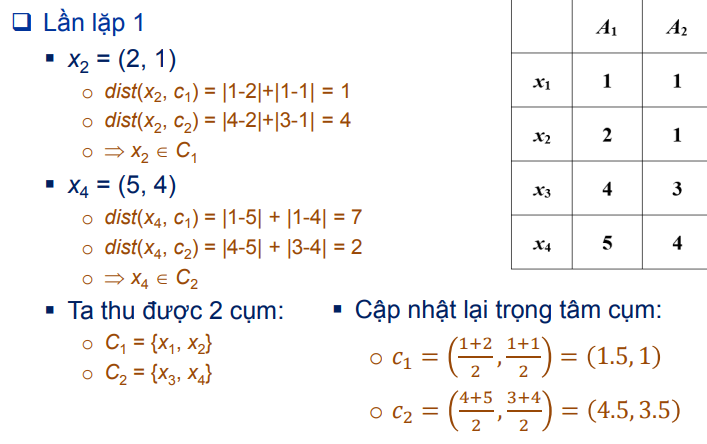




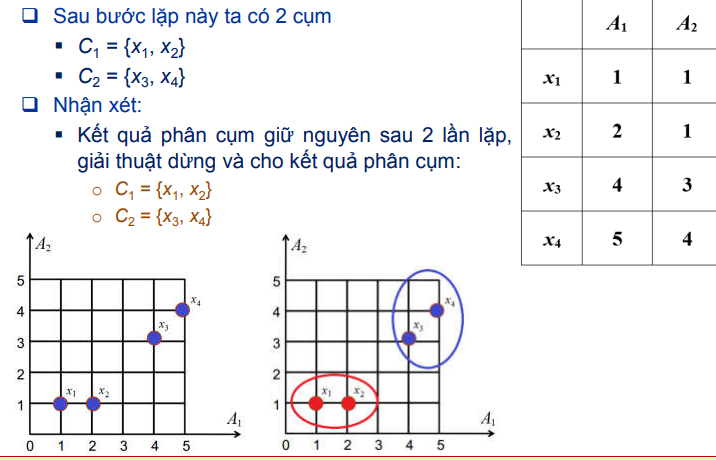










**Mô hình phân lớp Naive Bayes**

Một mô hình phân lớp là một mô hình Machine Learning dùng để phân loại các vật mẫu dựa trên các đặc tính đã xác định.

Naive Bayes là một thuật toán phân lớp được mô hình hoá dựa trên định Bayes trong xác suất thống kê.

A picture containing text

Description automatically generated

P(y|X) gọi là posterior probability: xác suất của mục tiêu y với điều kiện có đặc trưng X

P(X|y) gọi là likelihood: xác suất của đặc trưng X khi đã biết mục tiêu y

P(y) gọi là prior probability của mục tiêu y

P(X) gọi là prior probability của đặc trưng X

Ví dụ: Xét một bộ dữ liệu đơn giản về việc đi làm muộn của một bạn nhân viên

Table

Description automatically generated

Table

Description automatically generated

Table

Description automatically generated

Table

Description automatically generated

Giả sử, để dự đoán cho một ngày X=(Muộn, Xấu, Mưa), cần tính:

P(Muộn|X) ∝ P(Muộn|Muộn) \* P(Xấu|Muộn) \* P(Mưa|Muộn) \* P(Muộn) = (3/5) \* (2/5) \* (1/5) \* (5/10) = 0.024

P(Không muộn|X) ∝ P(Muộn|Không muộn) \* P(Xấu|Không muộn) \* P(Mưa|Không muộn) \* P(Không muộn) = (0/5) \* (3/5) \* (1/5) \* (5/10) = 0

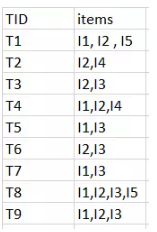
=> y = argmax{ P(X|y)P(y) } = Muộn

**Thuật toán Apriori:**

Để tìm các tập phổ biến trong một bộ dữ liệu lớn.

Tất cả các tập con của một tập hợp thường xuyên phải là thường xuyên (thuộc tính Apriori).Trong một vật phẩm không thường xuyên, tất cả các tập cha của nó sẽ không thường xuyên.

Ví dụ:



min support là 2 min confidence là 60%

Bước 1: K = 1 (I) Tạo bảng chứa số support của từng mục có trong tập dữ liệu - Được gọi là C1 (tập ứng cử viên)

Table

Description automatically generated

(II) so sánh số support của tập các ứng cử viên với số lượng hỗ trợ tối thiểu (ở đây min\_support = 2 nếu support\_count của tập ứng cử viên nhỏ hơn min\_support sẽ xóa các tập đó). Điều này cung cấp cho chúng ta mục L1.

Table

Description automatically generated

Bước 2: K = 2 Tạo tập ứng viên C2 bằng L1 (đây được gọi là bước kết hợp). Điều kiện để có thể kết hợp Lk-1 với Lk-1 là hai tập cha đó phải có K-2 (trong trường hợp này là 0) yếu tố chung . Duyệt qua các tập cha của C2, nếu tập cha nào không đạt chuẩn thường xuyên thì tập con đó sẽ bị xóa. (Ví dụ tập hợp con của {I1, I2} là {I1}, {I2} để kiểm tra độ thường xuyên thường xuyên. Kiểm tra cho từng mục) Bây giờ tính độ thường xuyên của các tập con mới được tạo.

Table

Description automatically generated

Tiếp tục kiểm tra độ thường xuyên của các tập trong C2, nếu tập nào không thỏa mãn min\_support thì xóa đi. Ta sẽ nhận được kết quả là tập L2.

Table

Description automatically generated

Bước 3: K = 3

Lặp lại quy trình như bước 2 ta được hai tập.

C3:

Table

Description automatically generated

L3:

Table

Description automatically generated

Cứ tiếp tục lặp lại đến khi tập Cn không còn phần tử nào nữa.

Từ đó, chúng ta đã phát hiện ra tất cả các tập vật phẩm thường xuyên. Bây giờ tính mạnh mẽ, bền chặt của một tập vật phẩm được chú ý tới. Cho rằng chúng ta cần tính toán sự tự tin của từng tập.

Sự tự tin - Độ tin cậy 60% có nghĩa là 60% khách hàng mà mua sữa và bánh mì cũng sẽ mua bơ.

Confidence(A->B)=Support\_count(A∪B)/Support\_count(A) Ở đây tôi sẽ lấy ví dụ với các tập thường xuyên ta tìm được ở trên

[I1^I2]=>[I3] //confidence = sup(I1^I2^I3)/sup(I1^I2) = 2/4\*100=50%

[I1^I3]=>[I2] //confidence = sup(I1^I2^I3)/sup(I1^I3) = 2/4\*100=50%

[I2^I3]=>[I1] //confidence = sup(I1^I2^I3)/sup(I2^I3) = 2/4\*100=50%

[I1]=>[I2^I3] //confidence = sup(I1^I2^I3)/sup(I1) = 2/6\*100=33%

[I2]=>[I1^I3] //confidence = sup(I1^I2^I3)/sup(I2) = 2/7\*100=28%

[I3]=>[I1^I2] //confidence = sup(I1^I2^I3)/sup(I3) = 2/6\*100=33%

=>nếu độ tin cậy tối thiểu là 50% thỉ 3 quy tắc trên là mạnh mẽ