Naive Bayes

Xét bài toán phân lớp với C lớp từ 1,2,...C. Có 1 điểm dữ liệu là x = (x1,x2,...xn). Tính xác xuất để điểm dữ liệu này rơi vào lớp c. Nói cách khác cần tính p(y = c|x) hay p(c|x).

Với bài toán đã cho ta có 5 lớp :

+Lớp 1: từ 18-24 tuổi
+Lớp 2: từ 25-34 tuổi
+Lớp 3: từ 35-44 tuổi
+Lớp 4: từ 45-54 tuổi
+Lớp 5: từ 55 tuổi trở lên

Khi xét các ID tương ứng với 1 điểm dữ liệu là x = (ID1,ID2,..Idn) là những người thích các group có ID đấy sẽ ở độ tuổi bao nhiêu thì ta đưa ra xác xuất tính là:

c = max (p(ci|x)) (ci thuộc c1,c2,c3,c4,c5 tương đương với 5 lớp)

<=> c = max (p(x|c)p(c) / p(x)) (công thức Bayes) (đổi như này vì xác xuất ở trên khá khó tính khi code)

<=> c = max(p(x|c)p(c)) (mẫu ở dưới luôn là p(x) nên để tìm max chỉ cần tử max)

<=> c = max (p(c)*p(ID1|c)*(ID2|c)*...*(IDn|c) (ở đây ở sử rằng các ID độc lập với nhau)

<=> c = max (log(p(c)) + log(p(ID1|c)) + … + log(p(IDn|c))) (vì các xác suất rất bé chỉ từ $0 \rightarrow 1$ nên khi phân các xác suất như ở trên có thể dẫn đến sai số xác xuất ở thể trở thành 0, vì vậy ta chuyển nó thành tổng các log như này)

Như vậy sau khi tìm được c ứng với mỗi lớp c1,c2,c3,c4,c5 cái nào có xác suất lớn nhất thì ta dự đoán là người dùng thuộc lớp đấy.

Chú ý: khi test sẽ gặp trường hợp ID không xuất hiện dẫn đến xác suất bằng 0 có thể ảnh hưởng tới kết quả cuối, nên ta cần làm “trơn” nó bằng cách khi tính xác suất thay vì (a/b) ta sẽ viết thành (a+1/ b+1) số 1 ở đây là hệ số làm “trơn”.