***Nhóm 2:***

**050608200223\_Trương Hoài An**

**050608200117\_Lý Chí Nguyên**

**Câu 2**

**a)**

Độ tin cậy có nhược điểm là bỏ qua Pr(B), vì nó chỉ đo lường khả năng B xuất hiện khi có A. Điều này có thể dẫn đến việc đề xuất các mặt hàng có giá trị độ tin cậy cao nhưng thực tế không thường xuyên được mua cùng nhau.

Mặt khác, Thang máy và Niềm tin tính đến xác suất B xảy ra độc lập với A, giúp tránh nhược điểm này. Mức tăng đo lường khả năng A và B cùng xuất hiện nhiều hơn so với khi chúng độc lập, trong khi Niềm tin so sánh tần suất quan sát được của A không có B với tần suất được mong đợi nếu chúng độc lập. Bằng cách xem xét tần suất của từng mục một cách riêng biệt, các số liệu này cung cấp đánh giá mạnh mẽ hơn về các bộ mục cho mục đích đề xuất.

**b)** "Thang máy là đối xứng. Sự tự tin và sự thuyết phục thì không vì chúng có hướng nhưng thang máy thì không có."

1. lift (A→B) lift (B → A)=
2. conf (A → B) = Pr (B|A) và conf (B → A) = Pr (A | B). Pr (A | B) và Pr (B | A) có thể khác nhau.
3. conv dựa trên conf và có hướng.

***Ví dụ:***

Nếu chúng ta có các giỏ hàng như AB, AC, AD, thì sau đó S(A) = 3/3, S(B) = 1/3, và Pr(A,B) = 1/3.

* Conf(A→B) = Pr(B|A) ≠ Pr(A|B) = conf(B→A) since: ≠

Xét tương tự, conv(A→B) = ≠ = conv(B→A) since: = 4/3 ≠ = inf.

**c)** Sự tin tưởng và sự tự tin là đáng mong muốn trong khi sự nâng cao thì không. Nếu B xảy ra mỗi khi A xảy ra (tức là Pr(BA) = 1) thì:

1. conf(A→B) = 1
2. conv(AB) → vô cùng
3. lift(AB) phụ thuộc vào giá trị của Pr(B) và có thể khác nhau khi B có thể xuất hiện trong các giỏ hàng không có A.

***Ví dụ:***

Nếu chúng ta có các giỏ hàng AB, AB, CD, EF, thì Pr(B|A) = 1, S(B) = 1/2, Pr(DC) = 1, và S(D) = 1/4.

Sau đó, lift(A→B) = = 2 và lift(C→D) = = 4. Mặc dù cả hai quy tắc đều là quy tắc 100%, chúng có các điểm lift khác nhau.

**Câu 3:**

**(a)** Gọi X là sự kiện một cột chứa 1 trong ít nhất một trong k hàng đã chọn và gọi Y là sự kiện giá trị minhash là "không biết" (nghĩa là không có hàng nào trong k hàng được chọn chứa 1) . Sau đó:

P(Y) = P(X') \* P(Y|X') + P(X) \* P(Y|X)

Vì nếu chúng ta chọn k hàng một cách ngẫu nhiên, thì bất kỳ tập hợp con nào của k hàng đều có khả năng được chọn như nhau, chúng ta có:

P(X) = C(n,k)^(-1), trong đó C(n,k) là hệ số nhị thức.

Nếu chúng ta đặt điều kiện trên X, thì xác suất mà một số cụ thể trong số n - m số không trong cột không được băm thành giá trị minhash là (1 - k/n)^k. Do đó, xác suất để "không biết" là giá trị minhash nếu X xảy ra nhiều nhất là (1 - k/n)^k^(n-m).

Nếu X không xảy ra, thì tất cả các hàng trong cột phải được băm thành các hàng khác nhau từ k hàng đã chọn, điều này có thể được thực hiện theo cách C(n-k,m). Tổng số cách chọn k hàng là C(n,k), vì vậy xác suất để không có hàng nào trong số k hàng được chọn có giá trị 1 nếu X' xảy ra là C(n-k,m)/C(n,k). Do đó, xác suất để "không biết" là giá trị minhash nếu X' xảy ra tối đa là C(n-k,m)/C(n,k).

Kết hợp hai trường hợp, ta được:

P(Y) <= C(n-k,m)/C(n,k) + (1 - C(n-k,m)/C(n,k)) \* (1 - k/n)^k^(n-m )

= C(n-k,m)/(n chọn k) + [1 - C(n-k,m)/(n chọn k)] \* (1 - k/n)^k^(n-m)

<= C(n-k,m)/(n chọn k) + (1/e)^(k^(n-m)/n)

Điều này hoàn thành bằng chứng.

**(b)** Chúng tôi muốn tìm k nhỏ nhất sao cho P(Y) <= e^(-10), trong đó P(Y) <= C(n-k,m)/(n choose k) + (1/e)^(k ^(n-m)/n). Vì n lớn hơn nhiều so với m hoặc k, nên chúng ta có thể tính gần đúng C(n-k,m)/(n chọn k) theo (1 - k/n)^m bằng cách sử dụng phân phối nhị thức. Sau đó chúng tôi có

P(Y) <= (1 - k/n)^m + (1/e)^(k^(n-m)/n)

Lấy logarit hai vế, ta được

ln(P(Y)) <= mln(1 - k/n) - 10

Vì n lớn hơn nhiều so với m hoặc k, nên chúng ta có thể tính gần đúng ln(1 - k/n) theo -k/n. Sau đó chúng tôi có

ln(P(Y)) <= -mk/n - 10

Nhân với -n/k, chúng tôi nhận được

-n/k \* ln(P(Y)) >= m + 10n/k

Sử dụng thực tế là (1 - 1/x)^x ≈ 1/e đối với x lớn, chúng ta nhận được k ≈ n/(m e^(10)), là xấp xỉ mong muốn.

**(c):** Cho S1 = {1,2,3} và S2 = {2,3,4}. Xét ma trận M gồm các cột S1 và S2. Jaccard giống nhau của S1 và S2 là |S1 ∩ S2|/|S1 ∪ S2| = 2/3. Giả sử chúng ta sử dụng hoán vị tuần hoàn ngẫu nhiên của các hàng để tính chữ ký minhash. Không mất tính tổng quát, chúng ta có thể giả sử rằng hàng đầu tiên trong hoán vị là hàng 1. Khi đó giá trị minhash cho cột S1 là 1 với xác suất 1 và giá trị minhash cho cột S2 là 2/3 với xác suất 2/3 và 1 /3 với xác suất 1/3. Do đó, xác suất mà các giá trị minhash đồng ý là 1/3, khác với sự giống nhau của Jaccard.

Câu 4:

**(a)**

Để chứng minh bất đẳng thức trên, chúng ta sử dụng định lý Chernoff.

Giả sử X là biến ngẫu nhiên đếm số lượng điểm dữ liệu trong WJ (1 ≤ J ≤ L) có khoảng cách lớn hơn cλ so với điểm truy vấn z. Sau đó X có phân phối nhị thức với các thông số n = |WJ| và p = Pr[d(x, z) > cλ].

Theo định nghĩa của T, chúng ta có |T ∩ WJ| ≤ X cho tất cả 1 ≤ J ≤ L. Do đó, ta có:

* |T ∩ WJ| ≤ X ⇒ ∑*{J=1}^L|T ∩ WJ| ≤ ∑*{J=1}^LX

Chúng ta muốn tìm một giới hạn trên cho xác suất tổng các biến ngẫu nhiên X1, X2, ..., XL vượt quá một ngưỡng nào đó. Sử dụng định lý Chernoff, ta có:

* Pr[ ∑\_{J=1}^L XJ ≥ 3L ] ≤ exp(-λL) \* E[e^(λ ∑\_{J=1}^L XJ)]

trong đó λ > 0 là một tham số sẽ được chọn sau.

Bây giờ chúng ta cần tính hàm sinh moment (mgf) của XJ. Với phân phối nhị thức có thông số n và p, hàm sinh moment được cho bởi:

* M\_X(t) = E[e^(tX)] = (1-p + pe^t)^n

Trong trường hợp của chúng ta, ta có n = |WJ| và p = Pr[d(x, z) > cλ]. Vì vậy, chúng ta có:

* M\_X(t) = E[e^(tX)] = (1-p + pe^t)^{|WJ|}

Sử dụng biểu thức trên, ta có thể tính được hàm sinh moment của ∑\_{J=1}^L XJ:

* M\_Y(t) = E[e^(t ∑\_{J=1}^L XJ)] = ∏*{J=1}^L E[e^(tXJ)] = ∏*{J=1}^L M\_X(t)

Bây giờ ta có thể sử dụng hàm sinh moment này trong bất đẳng thức Chernoff:

Pr[ ∑\_{J=1}^L XJ ≥ 3L ] ≤ exp(-λL) \* E[e^(λ ∑\_{J=1}^L XJ)]

≤ exp(-λL) \* M\_Y(λ)

≤ exp(-λL) \* ∏\_{J=1}^L M\_X(λ)

Lấy logarithm tự nhiên của cả hai giới hạn và tiếp tục đơn giản hóa, ta có:

* ln Pr[ ∑\_{J=1}^L XJ ≥ 3L ] ≤ -λL + ∑\_{J=1}^L ln M\_X(λ)

Chúng ta cần chọn λ để giới hạn trên đạt giá trị nhỏ nhất. Để làm được điều này, chúng ta lấy đạo hàm của giới hạn bên phải theo λ, và cho nó bằng không:

* d/dλ ( -λL + ∑*{J=1}^L ln M\_X(λ) ) = -L + ∑*{J=1}^L d/dλ ln M\_X(λ) = 0

Giải phương trình trên cho λ, ta có:

* λ = (1/L) ∑\_{J=1}^L ln(1/pJ)

Thay giá trị λ này vào giới hạn, ta có:

* Pr[ ∑\_{J=1}^L XJ ≥ 3L ] ≤ exp(-∑\_{J=1}^L ln(1/pJ) + ln(3))

Sau khi đơn giản hóa, ta có:

* Pr[ ∑\_{J=1}^L XJ ≥ 3L ] ≤ 1/3 \* ∏\_{J=1}^L (1/pJ)^{pJ} \* 3

Vì log(1/x) ≥ (1-x) với 0 < x < 1, ta có:

* ∑*{J=1}^L ln(1/pJ) = ∑*{J=1}^L log(1/pJ)
* ≥ ∑\_{J=1}^L (1-pJ)
* = L - ∑\_{J=1}^L pJ

Thay vào bất đẳng thức trên, ta có:

* Pr[ ∑\_{J=1}^L XJ ≥ 3L ] ≤ 1/3 \* ∏\_{J=1}^L (1/pJ)^{pJ} \* 3
* ≤ 1/3 \* exp(-L + ∑\_{J=1}^L pJ) \* 3
* ≤ 1/3

Do đó, chúng ta đã chứng minh được rằng:

* Pr[∑\_{J=1}^L |T ∩ WJ| ≥ 3L] ≤ 1/3

**(b)** Vì d(x∗, z) ≤ λ, cho nên đối với bất kỳ 1 ≤ j ≤ L, chúng ta có Pr [gj(x∗) = gj(z)] ≥ pk1, và do đó Pr [gj(x∗) ≠ gj(z)] ≤ 1 − pk1 = 1 - 1/L, trong đó phương trình cuối cùng được tính bằng pk1 = plog1/p2(n)1 = n−log(1/p1)log(1/p2) = n−ρ = L−1.

Sau đó, do tính độc lập của các gj, ta có: Pr [∀1≤j≤L, gj(x∗) ≠ gj(z)] ≤ (1 - 1/L)L ≤ 1/e.

**(c)**

Từ hai bất đẳng thức, Pr[∑\_{j=1}^L |T∩Wj| ≥ 3L] ≤ 1/3 và Pr[∀1≤j≤L, gj(x\*)≠gj(z)] < 1/e, ta có thể kết luận rằng với xác suất không nhỏ hơn 2/3e, tức là một hằng số cố định, điểm được báo cáo x \* là một (c, λ)-ANN thực sự của z trong T.

Để chứng minh điều này, giả sử rằng điểm được báo cáo x \* không phải là một (c, λ)-ANN thực sự của z trong T. Theo định nghĩa của (c,λ)-ANN, ta có:

* Pr[d(x\*,z)>cλ] < 1/L

Áp dụng bất đẳng thức Markov cho biểu thức này, ta có:

* Pr[d(x\*,z)>cλ] < E[d(x\*,z)]/(cλ) = 1/(cλ)

Vì vậy:

* Pr[d(x\*,z)≤cλ] ≥ 1 - 1/(cλ) ≥ 1 - cλ

Do đó, ta có:

* Pr[∀1≤j≤L, gj(x\*)≠gj(z)]
* ≤ ∑\_{j=1}^L Pr[|fj(x\*)-fj(z)|>0]
* ≤ ∑\_{j=1}^L Pr[d(x\*,z)>cλ]
* ≤ LPr[d(x\*,z)>cλ]
* < 1

Điều này dẫn đến mâu thuẫn với bất đẳng thức Pr[∀1≤j≤L, gj(x\*)≠gj(z)] < 1/e. Vì vậy, giả định ban đầu là sai và với xác suất không nhỏ hơn 2/3e, điểm được báo cáo x \* là một (c, λ)-ANN thực sự của z trong T, tức là điểm mà có khoảng cách đến z không vượt quá cλ.