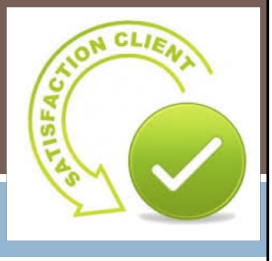
HỆ THỐNG GỢI Ý VÀ CÁC PHƯƠNG PHÁP ĐÁNH GIÁ



NỘI DUNG

- Nghi thức kiểm tra
- Các chỉ số đánh giá
  - Tiêu chí định lượng
  - Tiêu chí định tính
- Kết luận

3

NGHI THỨC ĐÁNH GIÁ

Nghi thức đánh giá

4

- sử dụng nghi thức k-fold :
  - chia tập dữ liệu thành k phần (fold) bằng nhau, lặp lại k lần, mỗi lần sử dụng k-1 folds để học và 1 fold để kiểm tra, sau đó tính trung bình của k lần kiểm tra
- nghi thức hold-out : lấy ngẫu nhiên 2/3 tập dữ liệu để học và 1/3 tập dữ liệu còn lại dùng cho kiểm tra, có thể lặp lại quá bước này k lần rồi tính giá trị trung bình

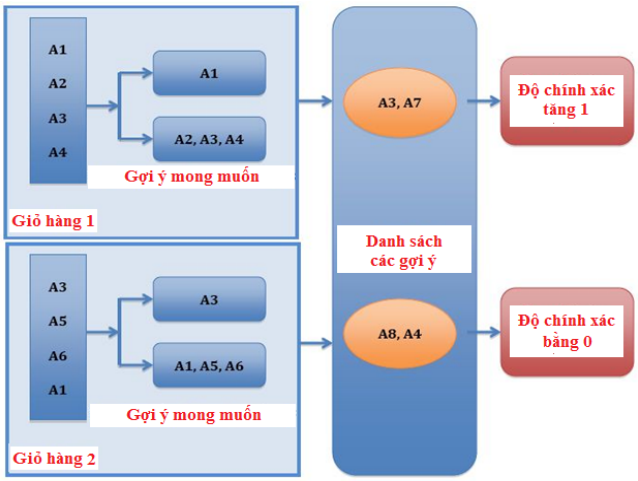
Nghi thức đánh giá

5

- **Given-N** [J.S Breese98]
  - Thường được sử dụng để đánh giá trong các lĩnh vực thương mại điện tử
  - Là một mở rộng của **k-fold** nhưng thực hiện trên từng giao dịch thay vì toàn bộ dữ liệu
  - Giao dịch sử dụng để đánh giá phải có ít nhất **N+1** mục dữ liệu
- **Phương pháp**
  - Chia danh sách các sản phẩm trong giỏ hàng thành 2 phần: tập được gọi là « Given » và 1 tập « Test »
  - So sánh các gợi ý thực tế (Test) và những sản phẩm gợi ý đề nghị bởi hệ thống, độ chính xác của hệ thống sẽ tăng lên 1 đơn vị hay bằng 0

Phương pháp đánh giá

6



Phương pháp đánh giá

7

"All But One" trường hợp đặc biệt của Given-N

- Tập "given" : số lượng các sản phẩm của giỏ hàng - 1 (ít nhất 1 sản phẩm)
- Tập « test » : luôn luôn bằng 1
- Ưu điểm của phương pháp " All But One " cho phép đánh giá các giỏ hàng có kích thước lớn hơn 1 sản phẩm

8

CHỈ SỐ ĐÁNH GIÁ

Tiêu chí định lượng

9

- **Đánh giá độ chính xác của các dự đoán**

movielens
helping you find the *right* movies

- Hệ thống gợi ý phim "MovieLens" [Dahlen et al. 1998] **dự đoán số ngôi sao** cho các bộ phim và hiển thị chúng cho người sử dụng hệ thống. Độ chính xác của dự đoán được đánh giá bởi số lượng phim mà người dùng và hệ thống có cùng số lượng ngôi sao cho cùng bộ phim. Ví dụ "City of God"
- Ngay cả khi một hệ thống gợi ý có khả năng cung cấp tên bộ phim cho người sử dụng nhưng hệ thống vẫn xem là không hiệu quả nếu số lượng ngôi sao đánh giá cho bộ phim của người dùng và hệ thống là không giống nhau.

Tiêu chí định lượng

10

- **Đánh giá độ chính xác của các dự đoán**
  - Đánh giá sự chính xác số điểm(ngôi sao) cho các mục dữ liệu (item) mà hệ thống tính toán được so với số điểm (ngôi sao) thực tế mà người sử dụng thực tế sẽ cho
  - Một số hệ thống
    - MovieLens 100K-10M ratings
    - Netflix 100M ratings

movielens
helping you find the *right* movies

Tiêu chí định lượng

Đánh giá độ chính xác của các dự đoán

- Các chỉ số thường dùng: **MSE** – Mean Square Error, **RMSE** – Root Mean Square Error, **MAE** – Mean Absolute Error
- Đo lường mức độ sai số của các dự đoán. Các giá trị đo lường này **bằng 0 khi hệ thống đạt được hiệu quả tốt nhất**. Giá trị này càng cao thì hiệu quả của hệ thống càng thấp.
- MAE là chỉ số được sử dụng nhiều nhất vì khả năng giải thích trực tiếp của nó.

Tiêu chí định lượng

12

Đánh giá độ chính xác của các dự đoán

- Tính chính xác của dự đoán được đo trên **n** quan sát trong đó $\mathbf{p_i}$ là giá trị dự đoán đánh giá của item **i**,

  $\mathbf{r_i}$ là giá trị đánh giá thực tế của item **i**
- Mean Absolute Error (*MAE*) (sai số trung bình tuyệt đối) tính toán độ lệch giữa dự đoán xếp hạng và xếp hạng thực tế

$$MAE = \frac{1}{n}\sum_{i=1}^{n} | p_i - r_i |$$

Tiêu chí định lượng

- **Đánh giá độ chính xác của các dự đoán**
  - Mean Square Error (*MSE*) (sai số bình phương trung bình)…..

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(p_i - r_i)^2$$

  - Root Mean Square Error (*RMSE*) (sai số trung bình toàn phương) tương tự như MAE nhưng chú trọng tới những giá trị có độ lệch lớn

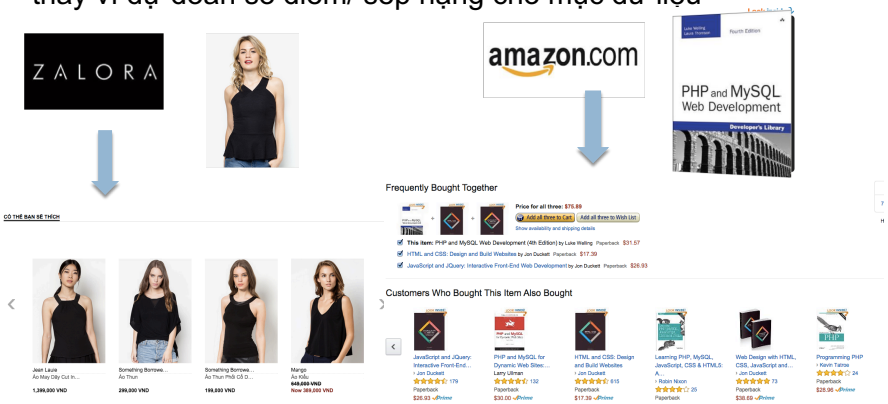
$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(p_i - r_i)^2}$$

Tiêu chí định lượng

14

Đánh giá việc sử dụng các dự đoán

- Hệ thống gợi ý đưa ra một danh sách các sản phẩm gợi ý thay vì dự đoán số điểm/ sếp hạng cho mục dữ liệu



Tiêu chí định lượng

15

Đánh giá việc sử dụng các dự đoán

- Đánh giá sự phù hợp gợi ý đối với người dùng. Gợi ý được xem là phù hợp khi người dùng chọn mục dữ liệu từ danh sách những đề nghị đã được gợi ý cho người dùng.
- Các chỉ số thường dùng:
  - Precision
  - Recall
  - F_{score}
  - Bresse score

Tiêu chí định lượng

16

Đánh giá việc sử dụng các dự đoán

- Precision là tỷ lệ giữa số lượng các gợi ý phù hợp và tổng số các gợi ý đã cung cấp (đã tạo ra). Precision bằng 100% có nghĩa là tất cả các kiến nghị đều phù hợp

$$Precision = \frac{Số\ lượng\ gợi\ ý\ phù\ hợp}{Số\ lượng\ gợi\ ý\ tạo\ ra}$$

- Recall được sử dụng để đo khả năng hệ thống tìm được những mục dữ liệu phù hợp so với những gì mà người dùng cần

$$Recall = \frac{Số\ lượng\ gợi\ ý\ phù\ hợp}{Số\ lượng\ sản\ phẩm\ mua\ bởi\ người\ dùng}$$

Tiêu chí định lượng

17

Đánh giá việc sử dụng các dự đoán

- Precision và Recall được xem là hữu ích trong việc đánh giá một gợi ý. Tuy nhiên, trong một số trường hợp thì precision và recall có giá trị tỉ lệ nghịch với nhau.
- ***F-score*** được sử dụng để đánh giá hiệu quả tổng thể của hệ thống bằng cách kết hợp hài hòa hai chỉ số Recall và Precision.

$$F - score = \frac{2 \times Precison \times Recall}{Precision + Recall}$$

Tiêu chí định lượng

18

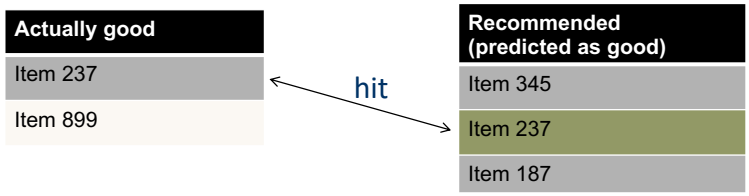
Đánh giá việc sử dụng các dự đoán

- **$Rank_{score}$ hay Breese score** [John S. Breese(1998)] cũng là một trong những chỉ số đánh giá khả năng sử dụng dự đoán nhưng chỉ số này chính xác đến thứ tự của các gợi ý được xây dựng
- Ví dụ, một hệ thống gợi ý cho người dùng 10 sản phẩm sắp xếp theo thứ tự ưu tiên từ cao đến thấp. Nếu người dùng chọn sản phẩm đầu tiên trong danh sách thì hệ thống gợi ý hiệu quả hơn khi người dùng chọn sản phẩm có thứ tự thứ 10

Tiêu chí định lượng

Đánh giá việc sử dụng các dự đoán

Ví dụ người dùng U



- **Rank Score** là mở rộng của giá trị "recall" để tìm được vị trí của item đúng trong danh sách có thứ tự các gợi ý.
  - Các gợi ý có độ ưu tiên thấp (lower ranked) có thể bị bỏ qua bởi người dùng.
  - Các item thích hợp hữu ích hơn khi nó xuất hiện "sớm" – độ ưu tiên cao trong danh sách khuyến nghị.

Tiêu chí định lượng

- **Đánh giá việc sử dụng các dự đoán**

$$R_i = \sum_j \frac{\delta(i,j)}{2^{(j-1)/(\alpha-1)}}$$

Trong đó:
- j là thứ tự của sản phẩm trong danh sách gợi ý
- **δ(I,j)** =1 nếu người dùng i chọn sản phẩm j, ngược lại **δ(I,j)** = 0
- **α** là ranking half life - xác suất mà mục dữ liệu trong danh sách gợi ý được chọn là 50%

Tiêu chí định tính

21

- Tính mới của các gợi ý
  - Ví dụ như hệ thống gợi ý "sữa tươi" cho khách hàng trong một siêu thị ở châu Âu. Đề nghị này là chính xác bởi vì hầu như tất cả các khách hàng đều mua sữa, nhưng nó không phải là hữu ích cho tất cả người dùng đã quen thuộc với sản phẩm này
  - Giả sử một người dùng mua quà tặng trước Giáng sinh hai tuần. Trong trường hợp này, việc xây dựng danh sách các gợi ý căn cứ vào mặt hàng phổ biến của các năm trước thì không phù hợp bởi vì người dùng thường tìm cách tặng các sản phẩm mới cho người thân.

Tiêu chí định tính

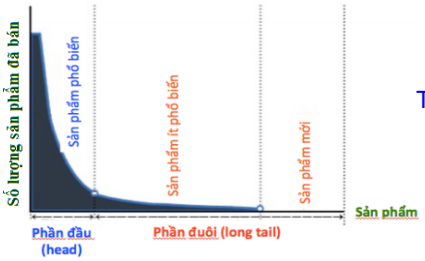
22

- Tính mới của các gợi ý
  - Thuộc tính mới được nhấn mạnh như là một chỉ số cần thiết để đánh giá tính hiệu quả của hệ thống gợi ý. G. Shani và cộng sự cùng với những nghiên cứu của mình đã chỉ ra 3 điểm quan trọng liên quan đến hiệu quả của hệ thống gợi ý.
  - (1) tính chính xác và tính mới lạ phải được tính đến để xây dựng được các gợi ý hiệu quả
  - (2) yếu tố thời gian là điều cần thiết trong việc đánh giá tính mới của mục dữ liệu
  - (3) danh sách gợi ý phù hợp nhất phải kết hợp một tỉ lệ các mục dữ liệu mới và các gợi ý phù hợp khác.

Tiêu chí định tính

23

- Tính mới của các gợi ý
  - Khái niệm "sản phẩm mới" ?
    - Tính mới của mục dữ liệu theo quan điểm thời gian (trong trường hợp xuất hiện một sản phẩm mới)
    - Liên quan đến lịch sử của người sử dụng (một sản phẩm mà chưa bao giờ được mua).



Tổ chức đồ các sản phẩm sắp xếp theo sự phổ biến của sản phẩm

Tiêu chí định tính

24

- Tính đa dạng (diversity)
  - Sự đa dạng của hệ thống gợi ý đo lường khả năng cung cấp một danh sách các mục dữ liệu được phân phối từ nhiều loại khác nhau.
  - Phân loại đa dạng
    - sự đa dạng cá nhân
    - đa dạng tổng thể
  - Ví dụ như sự đang dạng của các điểm tham quan cho các kỳ nghỉ lễ trong hệ thống gợi ý các địa điểm du lịch

Tiêu chí định tính

25

- Tính đa dạng (diversity)
  - Đa dạng cá nhân quan tâm đến các khái niệm về đa dạng từ quan điểm của người sử dụng. Chỉ số này được tính toán dựa trên trung bình sự khác nhau giữa tất cả các cặp mục dữ liệu đã gợi ý.
  - Sự đa dạng tổng thể là quan tâm đến các mục dữ liệu đã gợi ý hơn là quan tâm đến người dùng
  - Nếu sự đa dạng tổng thể của hệ thống giới thiệu là lớn, thì sự đa dạng của các gợi ý cá nhân cũng là rất lớn, nhưng điều này không đúng cho chiều ngược lại
  - Ví dụ, hệ thống cung cấp 3 sản phẩm gợi ý A,B,C cho tất cả người dùng, thì sự đa dạng cá nhân là tương đối cao nhưng sự đa dạng tổng thể là rất thấp.

Tiêu chí định tính

26

- Tính đa dạng (diversity)
  - Yếu tố có tác động trực tiếp đến sự đa dạng
    - các thuật toán sử dụng để xây dựng hệ thống
    - đặc tính của cơ sở dữ liệu
  - Kết quả nghiên cứu của G. Adomavicius và Y. Kwon về sự tương quan của độ chính xác và tính đa dạng

	Độ chính xác (precision)	Sự đa dạng
Sản phẩm nằm ở phần đầu (head)	82%	49 sản phẩm khác nhau
Sản phẩm nằm ở phần đuôi (tail)	68%	695 sản phẩm khác nhau

Bảng 5-1. Sự tương quan giữa độ chính xác và tính đa dạng.

Tiêu chí định tính

27

- Tính đa dạng (diversity)
  - Yếu tố có tác động trực tiếp đến sự đa dạng
    - các thuật toán sử dụng để xây dựng hệ thống
    - đặc tính của cơ sở dữ liệu
  - Sự đa dạng không phải là chỉ tiêu quan trọng nhất, độ chính xác cũng phải được tính đến

=> phải có một sự thỏa hiệp giữa độ chính xác và sự đa dạng (tức là giữa định lượng và định tính)

  - Ví dụ, giải thuật đề nghị của Adomavicius và Y. Kwon đánh giá trên cơ sở dữ liệu MovieLens, thì hệ thống tăng thêm 20% tính đa dạng nhưng chỉ làm mất đi 1% độ chính xác

Tiêu chí định tính

28

- Tính đa dạng (diversity)
  - Tính đa dạng có thể được quan tâm trực tiếp khi xây dựng danh sách các gợi ý danh sách gợi ý này sẽ được tính lại thứ tự sắp xếp.
  - Một vài chỉ số để tính lại thứ tự của các mục dữ liệu
    - sự phổ biến của các mục dữ liệu
    - trung bình các đánh giá cho mỗi mục dữ liệu
    - phần trăm người dùng đã có cùng một đánh giá cho một mục dữ liệu

Các giải thuật này chứng tỏ được tính hiệu quả ở sự đa dạng nhưng không làm giảm đáng kể độ chính xác của hệ thống.

Tiêu chí định tính

29

- Độ bao phủ (coverage)
  - Là thước đo số lượng lĩnh vực mà danh sách các sản phẩm gợi ý được tạo ra thuộc về chúng, số lĩnh vực này có bao trùm được hệ thống hay không.
  - Độ bao phủ của các gợi ý thấp thì thường ít được đánh giá cao bởi người dùng (người dùng bị giới hạn thông tin về các lĩnh vực của hệ thống và họ cần được tư vấn đa lĩnh vực)
  - Độ bao phủ (Coverage) thường được kết hợp với chỉ số "accuracy", vì không thể tăng độ bao phủ mà không quan tâm đến việc tạo ra những gợi ý không chính xác

Tiêu chí định tính

30

- Độ bao phủ (coverage)
  - Hầu hết độ bao phủ được đo bằng số các mặt hàng mà dự đoán có thể được hình thành như là một tỷ lệ phần trăm của tổng số các mặt hàng.
  - Một cách khác để tính độ bao phủ là chỉ xem xét độ bao phủ trên những mặt hàng mà người dùng quan tâm. Độ bao phủ tính theo cách này không được đo trên toàn bộ các sản phẩm mà chỉ quan tâm đến những sản phẩm mà khách hàng đã biết hay đã từng xem qua. Ưu điểm của cách tính này là nó đáp ứng tốt nhu cầu của người dùng.

31

#	Critère	Formule	Système appliqué
1	MAE	$\frac{1}{n}\sum_{i=1}^{n} p_i - r_i$	MovieLens
2	MSE	$\frac{1}{n}\sum_{i=1}^{n} (p_i - r_i)^2$	Netflix
3	RMSE	$\sqrt{\frac{1}{n}\sum_{i=1}^{n} (p_i - r_i)^2}$	BookCrossing
4	Précision	$\frac{nombreRecommandationsPertinentes}{nombreTotalRecommandationsgénérée}$	EachMovie,
5	Rappel	$\frac{nombreRecommandationsPertinentes}{nombreTotalItemAcheté}$	un grand magasin en Corée [Cho Yeong Bin(2004)]
6	F-mesure	$\frac{2*Precision*Rappel}{Precision+Rappel}$	MovieLens[B. Sarwar(2000a)]
7	Score du Breese	$R_i = \sum_j \frac{\delta(i,j)}{2^{(j-1)/\alpha}}$	TaFeng, B&Q [John S. Breese(1998), Chun-Nan Hsu(2004)]

TABLE 3.1: Résumé des critères d'évaluation quantitatifs

Kết luận

- Các nghi thức kiểm tra hệ thống
- Các tiêu chí đánh giá hệ thống
  - định lượng
  - định tính

=> để xây dựng hệ thống gợi ý chính xác và hữu dụng, chúng ta cần quan tâm đến việc chọn phương pháp và các chỉ số đánh giá hệ thống sao cho phù hợp, hiệu quả đối với từng bài toán cụ thể