

Trần Nguyễn Minh Thư - KHMT 1

HỆ THỐNG GỢI Ý VÀ CÁC PHƯƠNG PHÁP ĐÁNH GIÁ




1

NỘI DUNG

2

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



2

3

NGHI THỨC ĐÁNH GIÁ

3

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

4

Nghi thức đánh giá

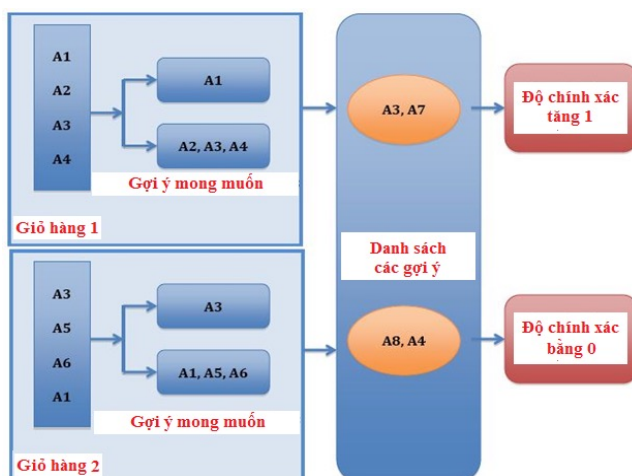
5

- **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

5

Phương pháp đánh giá

6



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

7

8

CHỈ SỐ ĐÁNH GIÁ

8

Tiêu chí định lượng

9

□ Đánh giá độ chính xác của các dự đoán

movielens
helping you find the right movies

(hide) Predictions for you %	Your Ratings	Movie Information	Wish List
★★★★	Not seen	Hotel Rwanda (2004)(Netflix) info[imdb] Drama, War	<input type="checkbox"/>
★★★★	Not seen	Kung Fu Hustle (Gong fu) (2004)(Netflix) info[imdb] Action, Comedy - Cloutiness	<input type="checkbox"/>
★★★★	Not seen	City of God (Cidade de Deus) (2002)(Netflix) DVD VHS info[imdb] Action, Crime, Drama, Thriller	<input type="checkbox"/>
★★★★	Not seen	Oldboy (2003)(Netflix) info[imdb] Action, Drama, Mystery, Thriller - Korean	<input type="checkbox"/>

- 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.

9

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



10

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.

11

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 **p_i** là giá trị dự đoán đánh giá của item **i** ,
 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

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

12

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

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

13

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/ xếp hạng cho mục dữ liệu

The image illustrates two different recommendation systems. On the left, Zalora's system suggests similar clothing items (dresses) to a user who viewed a specific dress. On the right, Amazon.com's system suggests related books, including 'Frequently Bought Together' and 'Customers Who Bought This Item Also Bought'.

14

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

15

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

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

- 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

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

16

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.

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

17

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

18

Tiêu chí định lượng

Đánh giá việc sử dụng các dự đoán

Ví dụ người dùng U

Actually good		Recommended (predicted as good)
Item 237	hit	Item 345
Item 899		Item 237
		Item 187

- 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.

19

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
- $\delta(i, j) = 1$ nếu người dùng i chọn sản phẩm j, ngược lại $\delta(i, j) = 0$
- α là ranking half life - xác suất mà mục dữ liệu trong danh sách gợi ý được chọn là 50%

20