Xây dựng hệ thống dự đoán sự đánh giá một bộ phim của người xem

Sinh viên thực hiện
Đặng Mạnh Cường Phan Thị Hồng Hạnh

Giáo viên hướng dẫn
TS.Trần Vĩnh Đức

Hà Nội, 09-1-2017

Mục Lục

Nội dung đề tài

- Xây dựng hệ thống có khả năng dự đoán sự đánh giá (rating) của người xem (user) đối với một bộ phim (movie)
- Dựa vào lịch sử đánh giá của người xem đối với các bộ phim - tập huấn luyện
  - Ma trận rating R_{mxn}, m là số lượng người xem, n là số lượng bộ phim
  - Nếu $R_{xi} \neq null$, người xem x đánh giá phim i với giá trị R_{xi} sao ($R_{xi} \in [0, 5]$)
  - Nếu $R_{xi} = null$, người xem x chưa đánh giá phim i

Nội dung đề tài

Đánh giá hiệu quả của hệ thống:

- Sử dụng tập Test: Tập T gồm các cặp người dùng, bộ phim mà hệ thống cần dự đoán

$$RMSE = \sqrt{\frac{\sum_{(i,x)\in T} (T_{xi} - \hat{T}_{xi})^2}{|T|}} \rightarrow MIN$$

  - $\hat{T}_{xi}$ là đánh giá của người xem x đối với bộ phim i do hệ thống dự đoán
  - T_{xi} là đánh giá thực của người xem x đối với bộ phim i

Mục Lục

Ý tưởng

Lọc cộng tác

- U là tập người xem
- M là tập các bộ phim
- Xét người dùng x và bộ phim i
- Trong tập $M(x)$ gồm những bộ phim mà người xem x đã rating, tìm tập $H(x, i)$ gồm những bộ phim có rating tương đồng nhất với i.
- Ước lượng rating của người xem x đối với bộ phim i dựa vào tập $H(x, i)$

Công thức đo độ tương đồng

Lọc cộng tác

- Gọi $sim(i,j)$ là độ tương đồng giữa i và j
- Sử dụng độ đo *cosine*: $sim(i,j) = cos(R_i, R_j) = \frac{R_i R_j}{||R_i||||R_j||}$

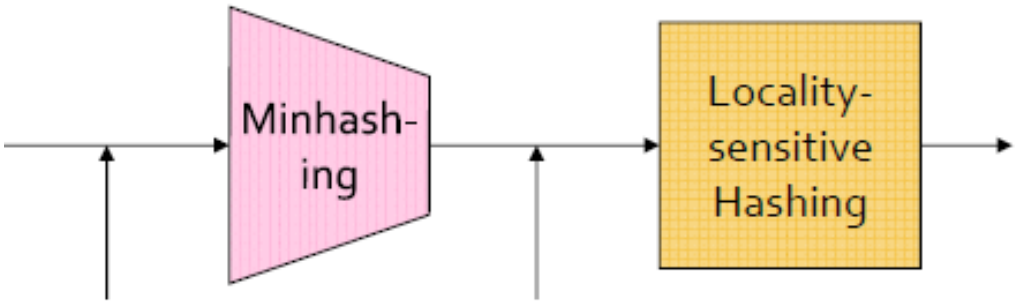
	HP1	HP2	HP3	TW	SW1	SW2	SW3
A	4			5	1		
B	5	5	4				
C				2	4	5	
D		3					3

$$sim(A,B) = \frac{4x5}{\sqrt{4^2+5^2+1^2}\sqrt{5^2+5^2+4^2}} = 0.38$$

- Khi $sim(i,j)$ càng lớn thì độ tương đồng giữa i và j càng cao

- Với mỗi bộ phim i, tìm tập $N(i) = \{j \in M | sim(i,j) \geq 0.2\}$
$$\Rightarrow H(x,i) = N(i) \cap M(x)$$
- Làm sao để tìm tập N(i)?
  - sử dụng thuật toán tầm thường $\Rightarrow$ độ phức tạp quá lớn
  - sử dụng Minhashing + Locality Sensitive Hashing

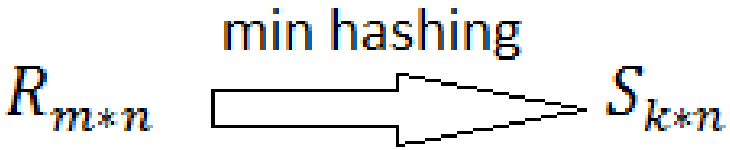
Tìm tập tương đồng



Candidate pairs : those pairs of signatures that we need to test for similarity.

Matrix rating R: The rows present users, the columns present movies

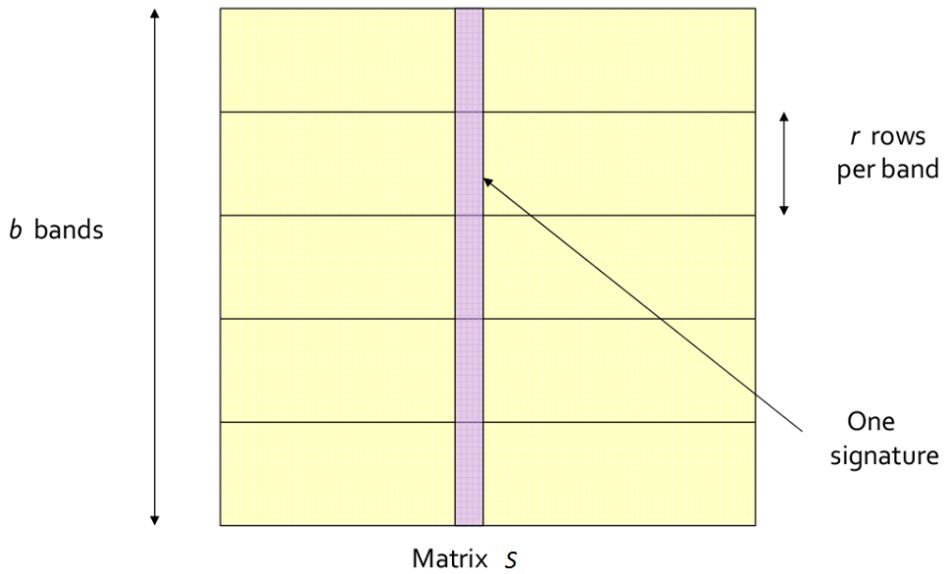
Signatures : short integer vectors that represent the sets, and reflect their similarity



- R_i là vector rating của bộ phim i ($|R_i| = m$)
- Lấy 1 tập ngẫu nhiên các vector $\{v_1, v_2, ..., v_k\}$ kích thước m, các vector chỉ chứa 2 giá trị -1 hoặc 1
- Xây dựng vector chữ ký S_i cho bộ phim i như sau:

$$\forall j \in [1, k], S_{ji} = \begin{cases} 1 & \text{if } R_i v_j > 0 \\ 0 & \text{if } R_i v_j \leq 0 \end{cases}$$

LSH



- Chia ma trận S thành b băng, mỗi băng r hàng
- Với mỗi băng, tiến hành băm phần của cột chữ ký thuộc băng đó vào bảng băm gồm c giỏ
- Một cặp ứng cử viên là một cặp được băm vào cùng 1 giỏ trong 1 băng bất kì
- Với mỗi cặp ứng cử viên, tính độ tương đồng và xây dựng tập $N(i)$

Ước lượng rating

$$\hat{R_{xi}} = b_{xi} + \frac{\sum_{j \in H(x,i)} sim(i,j) x (R_{xj} - b_{xj})}{\sum_{j \in H(x,i)} sim(i,j)}$$

Trong đó:

- $b_{xi} = \mu + b_x + b_i$
- μ là giá trị trung bình rating trên toàn ma trận R
- b_x là chênh lệch giữa giá trị rating trung bình của người xem x với μ
- b_i là chênh lệch giữa giá trị rating trung bình của bộ phim i với μ

Mục Lục

Nhân tố ẩn

- Gọi k là số nhân tố ẩn
- P_{kxm} là ma trận đặc tính tiềm ẩn của người xem
- Q_{kxn} là ma trận đặc tính tiềm ẩn của các bộ phim
- Cần tìm P, Q sao cho $P^T Q \approx R$

Nhân tố ẩn

Cực tiểu hàm mục tiêu:

$$E = \sum_{(x,i)\in R} (R_{xi} - \hat{R_{xi}})^2 + \frac{\lambda}{2}[\sum_x ||P||^2 + \sum_i ||Q||^2 + \sum_x ||U||^2 + \sum_i ||I||^2]$$

Trong đó:

- $\hat{R_{xi}} = \mu + U_x + I_i + P_x^T Q_i$
  - μ là giá trị trung bình rating của ma trận R
  - U_x là giá trị bias của người xem x
  - I_i là giá trị bias của bộ phim i
- λ là tham số điều khiển

Stochastic gradient descent

Đặt $W = \frac{\lambda}{2}[\sum_x ||P||^2 + \sum_i ||Q||^2 + \sum_x ||U||^2 + \sum_i ||I||^2]$

Xét người xem x và bộ phim i:

$$E_{xi} = (R_{xi} - \hat{R_{xi}})^2 + W = [R_{xi} - (\mu + U_x + I_i + \sum_{t=1}^{k} P_{tx} Q_{ti})]^2 + W$$

$$\varepsilon_{xi} = 2 * (R_{xi} - \mu - U_x - I_i - \sum_{t=1}^{k} P_{tx} Q_{ti})$$

Cập nhật tham số:

- $P_{tx} = P_{tx} - \alpha \frac{\partial E_{xi}}{\partial Ptx} = P_{tx} - \alpha(-\varepsilon_{xi} Q_{ti} + \lambda P_{tx})$
- $Q_{ti} = Q_{ti} - \alpha \frac{\partial E_{xi}}{\partial Qti} = Q_{ti} - \alpha(-\varepsilon_{xi} P_{tx} + \lambda Q_{ti})$
- $U_x = U_x - \alpha \frac{\partial E_{xi}}{\partial U_x} = U_x - \alpha(-\varepsilon_{xi} + \lambda U_x)$
- $I_i = I_i - \alpha \frac{\partial E_{xi}}{\partial I_i} = I_i - \alpha(-\varepsilon_{xi} + \lambda I_i)$

Algorithm 1 Stochastic gradient descent

1: *Initialize* P, Q, U, I
2: **for** $(x, i) \in$ *training* **do**
3: $\quad \varepsilon_{xi} = 2 * (r_{xi} - \mu - U_x - I_i - P_x^T Q_i)$
4: $\quad P_x = P_x + \alpha * (\varepsilon_{xi} Q_i - \lambda P_x)$
5: $\quad Q_i = Q_i + \alpha * (\varepsilon_{xi} P_x - \lambda Q_i)$
6: $\quad U_x = U_x + \alpha * (\varepsilon_{xi} - \lambda U_x)$
7: $\quad I_i = I_i + \alpha * (\varepsilon_{xi} - \lambda I_i)$
8: **end for**

Bộ dữ liệu

- MovieLen - 100k: bộ dữ liệu gồm 100000 rating của 942 người xem trên 1692 bộ phim
- MovieLen - latest: bộ dữ liệu gồm hơn 24 triệu rating của 256000 người xem trên 40110 bộ phim

Kết quả

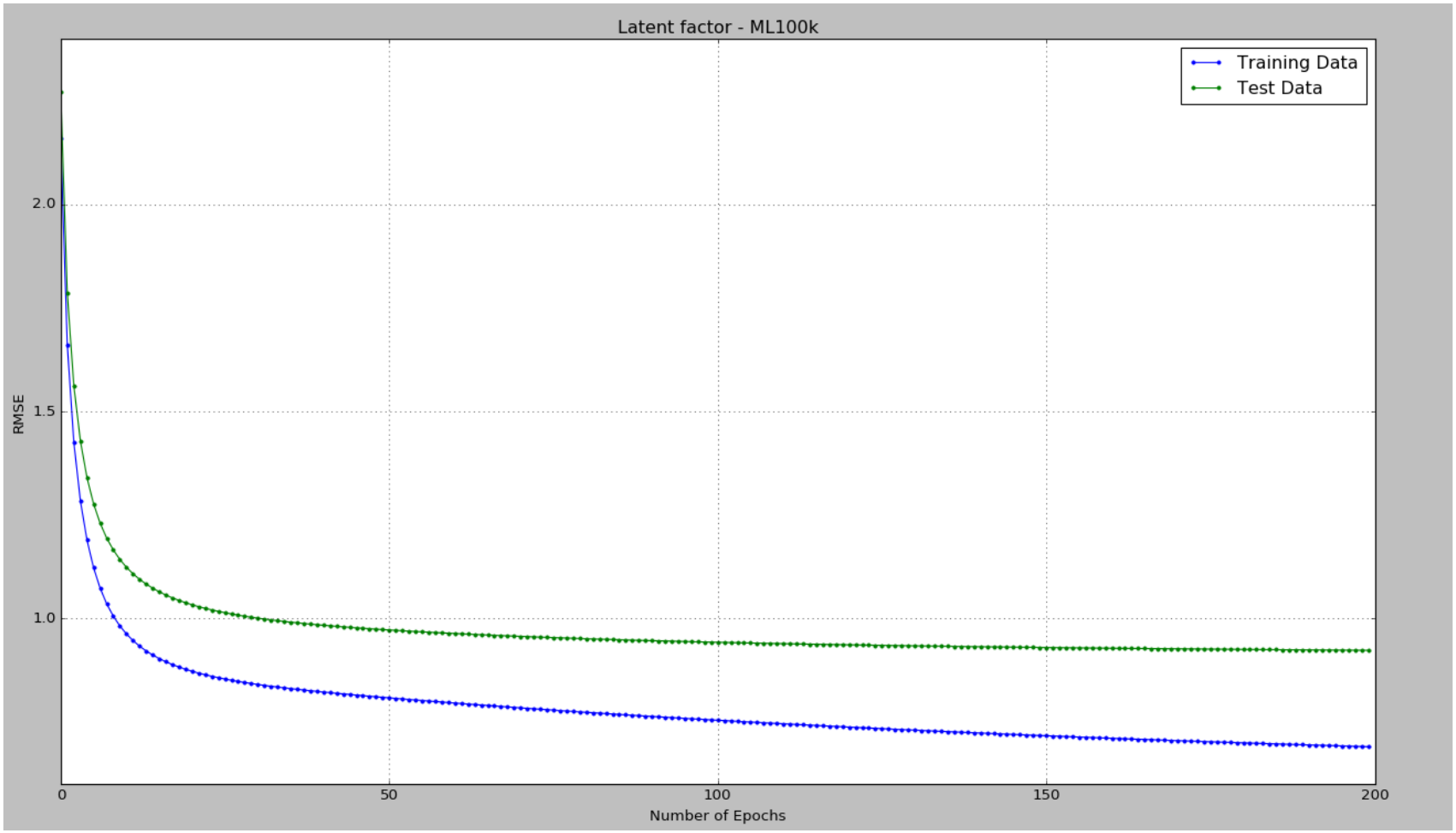
Bảng kết quả so sánh giá trị hàm RMSE giữa 2 mô hình lọc cộng tác (CF) và nhân tố ẩn (LF).

	ml-100k		ml-latest	
	Traing data	Test data	Training data	Test data
CF	0.85	0.95	0.98	1.03
LF	0.41	0.92	0.90	0.92

$\Rightarrow$ Mô hình nhân tố ẩn cho kết quả tốt hơn lọc cộng tác trên cả 2 bộ dữ liệu

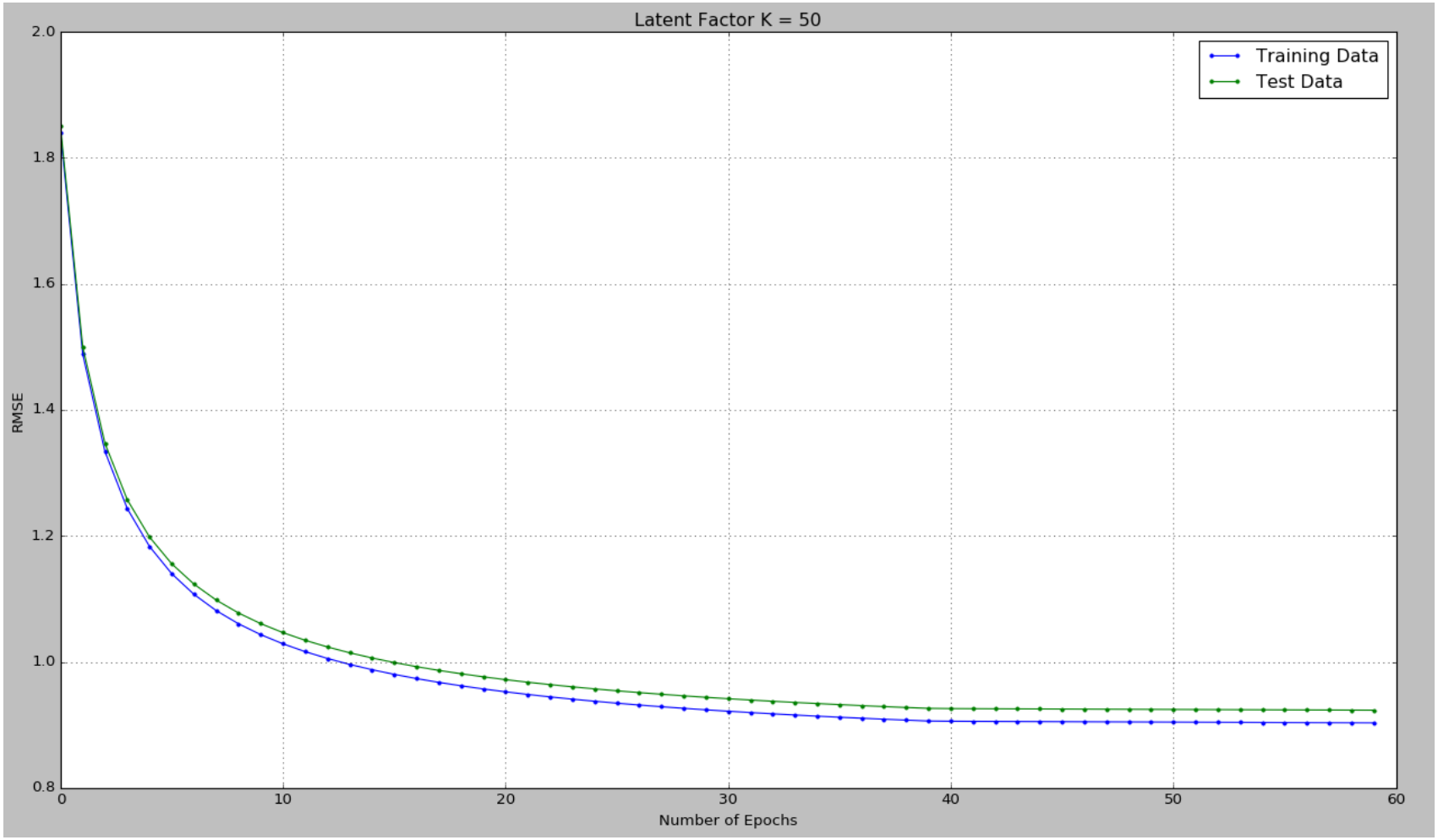
Kết quả

Đồ thị biểu diễn quá trình học LF trên bộ dữ liệu ml-100k



Kết quả

Đồ thị biểu diễn quá trình học LF trên bộ dữ liệu ml-latest



THANK YOU!