Personalized Regression for Recommender System

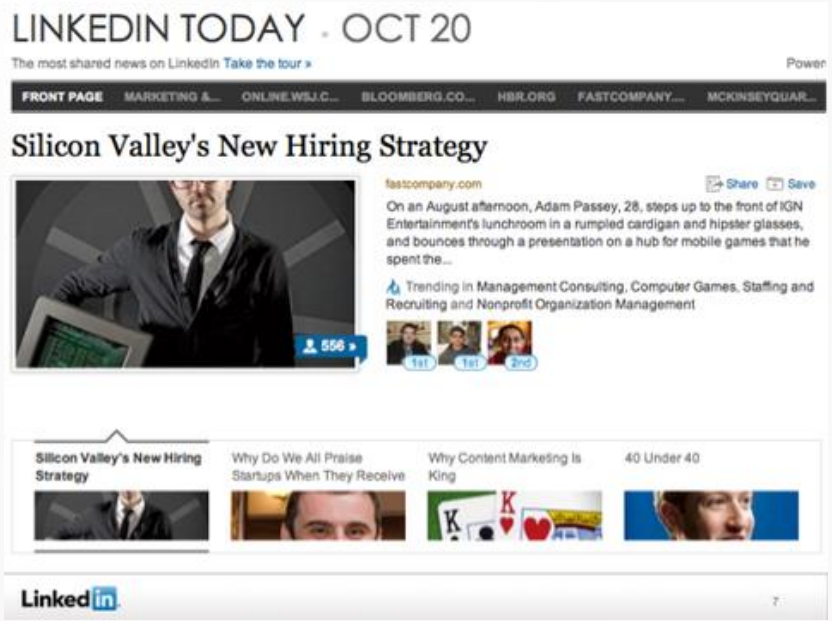
서울대학교 통계학과
최 세 민, 김 용 대

Contents

1. 추천 시스템 소개

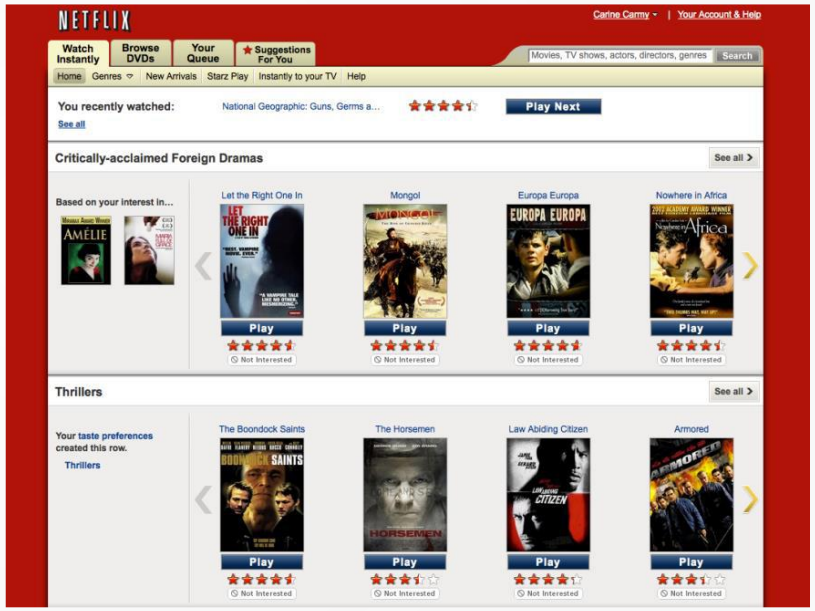
추천 시스템 소개

- News Articles, Tags, Online Mates, Restaurants
- Courses in e-Learning, Movies, Books, Various Goods
- Research Papers, Citations, …



추천 시스템 소개


amazon.com
The All-New Kindle Family
Kindle $79
Kindle Touch $99
Kindle Fire $199
Textbooks
Kindle Fire
Movies, music, games, web & reading
Only $199
Accessories for Kindle
Covers, skins, and more
Returns Are Easy
PSVITA
Console-style gaming in the palm of your hand
Flexible Tripod Goes Where You Go
More Items to Consider


NETFLIX
You recently watched:
Play Next
Critically-acclaimed Foreign Dramas
Based on your interest in...
Thrillers
Your taste preferences created this row.

2. 협력적 정화

협력적 정화방법(Collaborative Filtering)

- 협력적 정화방법이란?
  - 개인화된 추천을 위한 통계적 방법
    - 개인의 선호도와 과거 상품 구매이력 등을 분석하여 개인에게 최적인 상품을 추천.
  - 기본 아이디어
    - 주어진 고객과 상품들에 대한 선호도가 비슷한 고객을 조사
    - 선호도가 비슷한 고객들이 좋아하는 상품 중에 주어진 고객이 모르고 있는 상품을 추천.
  - 종류
    - 고객 중심의 협력적 정화방법(User-based)
    - 상품 중심의 협력적 정화방법(Item-based)

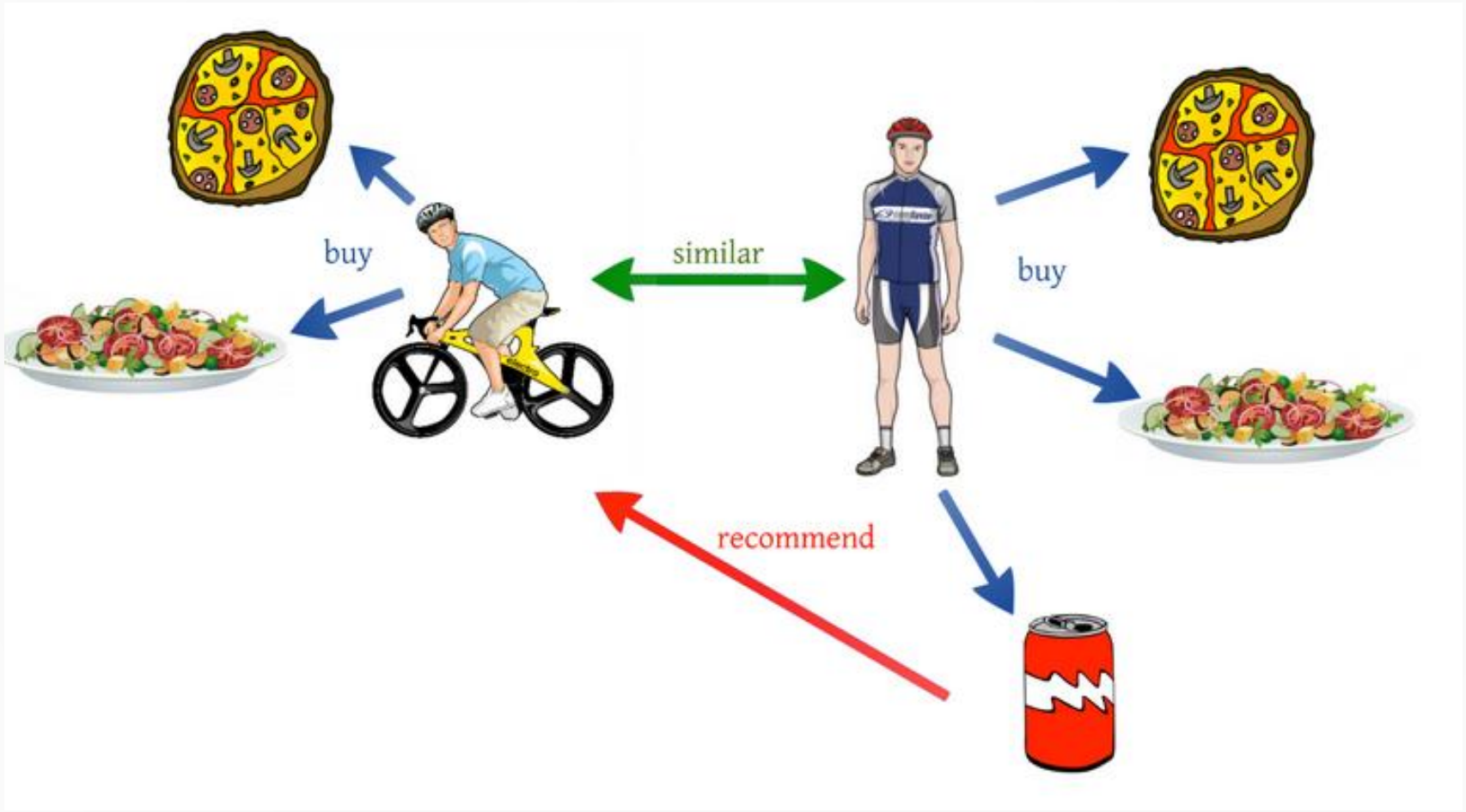
협력적 정화방법(Collaborative Filtering)

- 선호도 데이터의 예

	The Matrix	Titanic	Die Hard	Forrest Gump	Wall-E
John	5	1	2	2	
Lucy	1	5	2	5	5
Eric	2	?	3	5	4
Diane	4	3	5	3	

평점을 몇 점 부여하게 될까?

고객 중심의 협력적 정화방법



고객 중심 협력적 정화방법

- 관측되지 않은 선호도를 추정하는 방법

1. 고객들 사이의 선호도 패턴의 유사성을 측정
   - 표기법
     - r_{ui}는 u번째 고객의 i번째 상품에 대한 선호도(관측되지 않은 선호도)
     - O_{uv}는 고객 u와 v에서 동시에 선호도를 관측한 상품들의 집합
     - $\overline{r_u}$와 $\overline{r_v}$는 각각 고객 u와 고객 v에서 관측된 선호도들의 평균
   - 사용자 u와 v의 유사도(similarity) $s(u,v)$는 주로 Pearson correlation coefficient 또는 cosine similarity를 사용.
     - $s^U(u,v) = \frac{\sum_{i \in O_{uv}} (r_{ui} - \bar{r}_u)(r_{vi} - \bar{r}_v)}{\sqrt{\sum_{i \in O_{uv}} (r_{ui} - \bar{r}_u)^2} \sqrt{\sum_{i \in O_{uv}} (r_{vi} - \bar{r}_v)^2}}$ Pearson Correlation coefficient
     - $s^U(u,v) = \frac{\sum_{i \in O_{uv}} r_{ui} r_{vi}}{\sqrt{\sum_{i \in O_{uv}} r_{ui}^2} \sqrt{\sum_{i \in O_{uv}} r_{vi}^2}}$ Cosine similarity

고객 중심 협력적 정화방법

- 관측되지 않은 선호도를 추정하는 방법

2. 유사성을 이용하여 선호도 측정
   - 고객 u와 유사성이 높은 고객을 이용하여 선호도 추정
   - 표기법
     - $R = \{(u, i): r_{ui} \text{ is observed}\}$
     - $R_U(i) = \{u : r_{ui} \; is \; observed\}$
     - $R_U^k(u : i) : R_U(i)$에 속한 고객 중에 고객 u와 유사성이 큰 k명의 고객들의 집합

고객 중심 협력적 정화방법

- 관측되지 않은 선호도를 추정하는 방법

2. 유사성을 이용하여 선호도 측정 (Cont.)
   - 선호도 추정
     - $\widehat{r_{ui}} = \frac{\sum_{v \in R_U^k(u:i)} r_{vi}}{|R_U^k(u:i)|}$ 또는 $\widehat{r_{ui}} = \mu_{ui} + \frac{\sum_{v \in R_U^k(u:i)} (r_{vi} - \mu_{vi})}{|R_U^k(u:i)|}$, $(\mu_{ui} = \mu_0 + \mu_u^U + \mu_i^I)$
     - $\widehat{r_{ui}} = \frac{\sum_{v \in R_U^k(u:i)} s^U(u,v) r_{vi}}{\sum_{v \in R_U^k(u:i)} |s^U(u,v)|}$ 또는 $\widehat{r_{ui}} = \mu_{ui} + \frac{\sum_{v \in R_U^k(u:i)} s^U(u,v)(r_{vi} - \mu_{vi})}{\sum_{v \in R_U^k(u:i)} |s^U(u,v)|}$

전통적인 협력적 정화방법의 문제점

- 자료의 sparsity로 유사성의 측정이 어렵다.
- 고객의 demographic정보나 상품의 내용정보를 분석에 사용하기가 어렵다.
- 새로운 고객이나 새로운 상품에 대한 추천이 어렵다.
  - ✓ Cold start problem

✓ 대안: 회귀모형을 이용한 협력적 정화방법

회귀모형을 이용한 협력적 정화방법

✓ 설명의 편의를 위해 상품 중심의 협력적 정화 방법에 대해서만 소개.

- The global CF model
  - 기존의 협력적 정화 방법에서의 평점 예측 식은 다음과 같다.
    - $\widehat{r_{ui}} = \mu_{ui} + \frac{\sum_{j \in R_I^k(i:u)} s^I(i,j)(r_{uj} - \mu_{uj})}{\sum_{j \in R_I^k(i:u)} |s^I(i,j)|} = \mu_{ui} + \sum_{j \in R_I^k(i:u)} w_{ij}^u (r_{uj} - \mu_{uj})$
  - 이를 보다 간단하게 하기 위해 $R_I^k(i:u)$를 $R_I(u)$로 바꾸고, w_{ij}^u를 w_{ij}로 바꾸면 다음과 같다.

$$\widehat{r_{ui}} = \mu_{ui} + \sum_{j \in R_I(u)} w_{ij} (r_{uj} - \mu_{uj})$$

회귀모형을 이용한 협력적 정화방법

- The weighted global CF model
  - 앞에서 언급된 global CF model에서 약간 변형된 형태.
  - 실험적으로 더 좋은 성능을 내는 것이 입증.

$$\widehat{r_{ui}} = \mu_{ui} + |R_I(u)|^{-1/2} \sum_{j \in R_I(u)} w_{ij}\,(r_{uj} - \mu_{uj})$$

  - 모수 추정 방법

$$min_{\mu_0, \mu_i^U, \mu_i^I, w_{ij}, u \in U, i,j \in I} \sum_{(u,i) \in R} \{r_{ui} - \hat{r}_{ui}\}^2 + \lambda \left(\sum_u {\mu_u^U}^2 + \sum_i {\mu_i^I}^2 + \sum_{i,j} w_{ij}^2 \right)$$

회귀모형을 이용한 협력적 정화방법의 문제점

- 모수의 수가 많다 (Item수의 제곱)
- 자료의 sparsity로 모수의 추정이 어렵다.

✓ 행렬분해 (matrix factorization) 알고리즘의 탄생

- Netflix Prize 수상자가 사용한 방법
- 추천 시스템에서 최근에 가장 널리 쓰임.

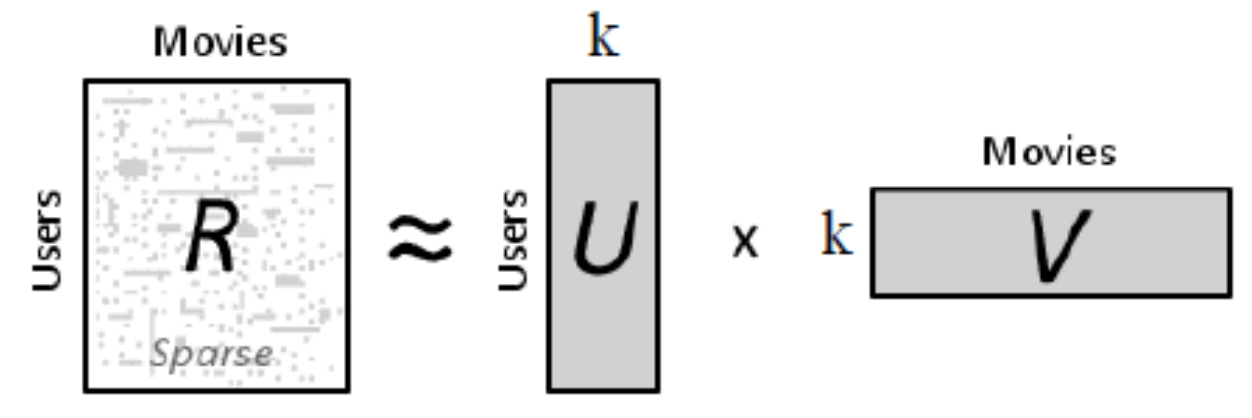
3. 행렬분해 방법

행렬분해 방법(Matrix factorization)

- 고객과 상품 사이의 관계가 직접적이 아닌 잠재 요인을 통해서 존재.

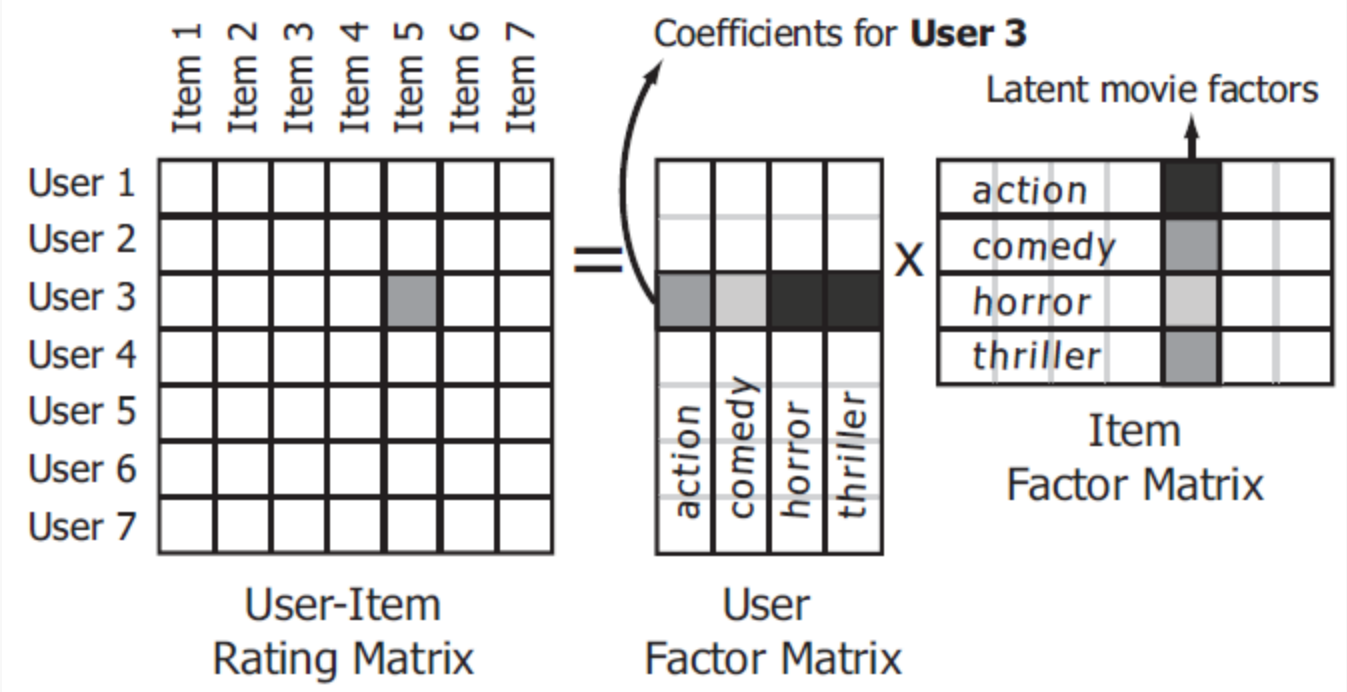
- 예 : 영화평을 예측하기 위한 경우.
  목적 : 영화 i에 대한 고객 u의 평점 r_{ui}예측.
  $\phi_u^U (\in \mathbb{R}^k)$: k 가지의 영화 장르에 대한 사용자 u의 선호도.
  $\phi_i^I (\in \mathbb{R}^k)$: k 가지의 영화 장르에 대한 영화 i의 구성 정도.
  - ✓ k 가지의 영화 장르를 잠재 요인이라 함.

- $\widehat{r_{ui}} = \phi_u^{U'} \phi_i^I$으로 평점을 예측.
- 좀 더 일반적으로, $\widehat{r_{ui}} = \mu_{ui} + \phi_u^{U'} \phi_i^I$로 예측하기도 함.



행렬분해 방법(Matrix factorization)

- 예제



행렬분해 방법(Matrix factorization)

- 고객과 상품에 대한 요인 벡터를 추정하기 위해 벌점 함수가 추가된 제곱합 함수 또는 교차 무질서도 함수를 최소로 하는 방법을 주로 사용.

$$\mathbf{argmin}_{\mu_{ui},\phi_u^U,\phi_i^I,u\in U,i\in I} \sum_{(u,i)\in R} \left(r_{ui} - \mu_{ui} - \phi_u^{U'}\phi_i^I\right)^2 + J_\lambda(\phi)$$

✓ 벌점함수의 예 – 제곱합 함수 : $J_\lambda(\phi) = \lambda_U \sum_{u\in U} \left\|\phi_u^U\right\|^2 + \lambda_I \sum_{i\in I} \left\|\phi_i^I\right\|^2$

4. 개인화 회귀분석

개인화 회귀분석

- R_{ui} : 고객 u의 상품 i에 대한 선호도 확률변수
- r_{ui} : R_{ui}의 관측값
- $R_u = (R_{ui} : i = 1, \dots, I)$: 고객 u의 선호도 확률벡터
- R_u는 평균이 μ_u이고 분산이 Σ_u인 다변량정규분포를 따르고, 서로 독립이라고 가정.

- 이 때, $E(R_{ui}|R_{uj} = r_{uj}, (u,j) \in R) = \mu_{ui} + c_{ui}'\Sigma_{ui}^{-1}(r_{u(-i)} - \mu_{u(-i)})$로 관측되지 않은 선호도를 추정함.
  - $c_{ui} = (\sigma_{uij}, (u,j) \in R, j \neq i)$, $\Sigma_{ui} = (\sigma_{ujk}, j \in R_u^U, k \in R_u^U, j \neq i, k \neq i)$
  - $r_{u(-i)} = (r_{uj}, j \in R_u^U, j \neq i), \mu_{u(-i)} = (\mu_{uj}, j \in R_u^U, j \neq i)$

- 즉, μ_u와 Σ_u를 추정함으로써 관측되지 않은 모든 선호도를 추정할 수 있다.

개인화 회귀분석

- **Method of Moment approach 모형식**

$$R_u \sim N_I(\mu_u, \Sigma_u),\ R_u\text{'s are independent}$$
$$\mu_{ui} = \mu_0 + \mu_i^I + \mu_u^U,\ \Sigma_u = \sigma_u^2 \Phi$$

- **Method of Moment approach 모수 추정 방법**

- μ_{ui} 추정 : $\mathrm{argmin}_{\mu_0, \mu_i^I, \mu_u^U} \sum_{(u,i)\in R} \left(r_{ui} - \mu_0 - \mu_i^I - \mu_u^U\right)^2 + \lambda_U \sum_u {\mu_u^U}^2 + \lambda_I \sum_i {\mu_i^I}^2$
- σ_u^2 추정 :
  - 고객 u의 선호도의 표본분산 $\widehat{\sigma_u^2} = \sum_{j\in R_u^U} \left(r_{uj} - \mu_{uj}\right)^2 / \left|R_u^U\right|$
  - shrinkage estimator $\widehat{\sigma_u^2} = \frac{\sum_{j\in R_u^U} \left(r_{uj} - \mu_{uj}\right)^2 + q_\sigma \widehat{\sigma^2}}{\left|R_u^U\right| + q_\sigma}$
    - $\widehat{\sigma^2} = \sum_u \sum_{j\in R_u^U} \left(r_{uj} - \bar{r}\right)^2 / \sum_u \left|R_u^U\right|$, $\bar{r} = \sum_u \sum_{j\in R_u^U} r_{uj} / \sum_u \left|R_u^U\right|$

개인화 회귀분석

- Φ 를 추정하기 앞서, 아래의 covariance 추정량 $\widehat{cov}_{jk}$를 제안한다.
  - $\widehat{cov}_{jk} = \frac{\sum_{u \in R_i^I \cap R_k^I} \frac{(r_{uj}-\mu_{uj})(r_{uk}-\mu_{uk})}{\widehat{\sigma_u^2}}}{\sum_u I(j,k \in R_u^U)}$
  - $\widehat{cov}_{jk}^{soft} = \left(\widehat{cov}_{jk} - \frac{\lambda}{\sqrt{n_{jk}}}\right)_+ , \quad (n_{jk} = \sum_u I(j,k \in R_u^U))$
- Φ 추정 :
  - $\widehat{\Phi}_{jk} = \frac{\widehat{cov}_{jk}}{\sqrt{\widehat{cov}_{jj}\widehat{cov}_{kk}}}$
- 위의 추정값들을 $\mu_{ui} + c_{ui}'\Sigma_{ui}^{-1}(r_{u(-i)} - \mu_{u(-i)})$에 대입하여 관측되지 않은 선호도를 모두 추정.
- 혹은, $\mu_{ui} + c_{ui}'(\Sigma_{ui} + \lambda I_{n_{ui}})^{-1}(r_{u(-i)} - \mu_{u(-i)})$에 대입. (shrinkage version.)
  - $n_{ui} = \sum_{j \neq i} I(j \in R_u^U)$

개인화 회귀분석

- Method of moment approach의 재해석 :
  - 한편, $r_{ui} - \mu_{ui} = \sum_{j \in R_u^U, j \neq i} \beta_{ij}^u (r_{uj} - \mu_{uj}) + \epsilon_{ui}$와 같은 회귀모형을 고려하면, β_{ij}^u의 최소제곱추정량은 $c_{ui}{}' \Sigma_{ui}^{-1}$의 MME와 같다.
  - 마찬가지로, 위 회귀모형의 ridge estimator는 $c_{ui}' (\Sigma_{ui} + \lambda I_{n_{ui}})^{-1}$의 MME와 같다.
  - 즉, 위의 회귀모형은 각각의 고객에 대하여 두 상품의 선호도 사이의 공분산을 모형화한 것으로 볼 수 있음 .
  - 위의 회귀모형을 Personalized regression이라 한다.
- 앞에서 소개된 협력적 정화 방법 review
  - Nearest neighborhood method

$$\widehat{r_{ui}} = \mu_{ui} + \sum_{j \in R_I^k(i;u)} w_{ij}^u (r_{uj} - \mu_{uj})$$

  - Global CF model

$$\widehat{r_{ui}} = \mu_{ui} + \sum_{j \in R_I(u)} w_{ij} (r_{uj} - \mu_{uj})$$

개인화 회귀분석

- **Personalized regression algorithm의 장점**
  - 반복적인 계산이 필요하지 않아서 거대자료에 쉽게 적용할 수 있다.
  - 특히, 분산 처리가 용이하여 계산 속도를 비약적으로 향상시킬 수 있다.
  - 앞에 소개된 모든 방법들에 비해 정확도가 우수하다. (실험 결과 참고)
  - 부가정보를 반영하기 쉽다.

개인화 회귀분석

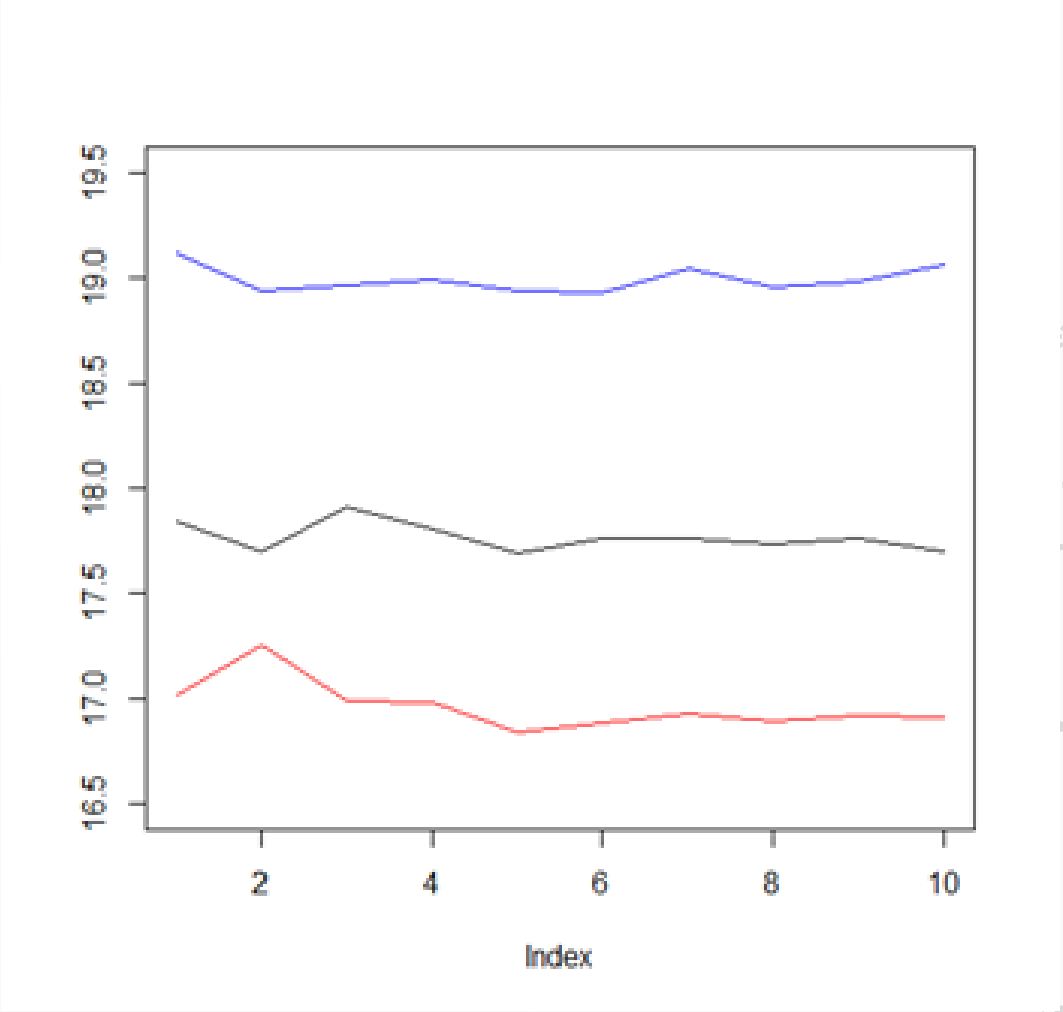
- Personalized regression algorithm과 기존 방법의 비교 실험
  - 자료 : Jester5k (R의 recommenderlab 패키지 안의 자료)
    - 5000명의 user, 100개의 item
    - 362106 ratings

	j1	j2	j3	j4	j5	j6	j7	j8	j9	j10	j11	j12	j13	j14	j15	j16	j17
u2841	7.91	9.17	5.34	8.16	-8.74	7.14	8.88	-8.25	5.87	6.21	7.72	6.12	-0.73	7.77	-5.83	-8.88	8.98
u15547	-3.20	-3.50	-9.56	-8.74	-6.36	-3.30	0.78	2.18	-8.40	-8.79	-7.04	-6.02	3.35	-4.61	3.64	-6.41	-4.13
u15221	-1.70	1.21	1.55	2.77	5.58	3.06	2.72	-4.66	4.51	-3.06	2.33	3.93	0.05	2.38	-3.64	-7.72	0.97
u15573	-7.38	-8.93	-3.88	-7.23	-4.90	4.13	2.57	3.83	4.37	3.16	-4.90	-5.78	-5.83	2.52	-5.24	4.51	4.37
u21505	0.10	4.17	4.90	1.55	5.53	1.50	-3.79	1.94	3.59	4.81	-0.68	-0.97	-6.46	-0.34	-2.14	-2.04	-2.57
u15994	0.83	-4.90	0.68	-7.18	0.34	-4.32	-6.17	6.12	-5.58	5.44	-4.85	-7.62	-6.65	-9.37	-6.07	-1.26	-7.43

- 비교 방법 : 7:3으로 training set과 test set을 설정하여 평균모델, 행렬분해방법과 본 방법을 test error로 비교한다. 평균모델이란, 관측되지 않은 선호도를 $\mu_{ui} = \mu_0 + \mu_i^I + \mu_u^U$ 만을 이용하여 예측하는 것을 뜻한다.

개인화 회귀분석

- Personalized regression algorithm과 기존 방법의 비교 실험
  - 파란색 : 평균모델
  - 검은색 : 행렬분해방법
  - 빨간색 : 개인화회귀분석 방법



THANK YOU