Coupon Purchase Prediction

2015.09 캐글뽀개기

임동권

팀 모집해놓고 팀활동 제대로 못해서 죄송합니다 ㅠㅠ

Overview

- What is this Competition?
- 데이터 설명
  - 실습
- Solutions
  - 실습
- Evaluation Metric 알기
  - 자유실습

What is this competition?

- https://www.kaggle.com/c/coupon-purchase-prediction

데이터 분석의 목표

- 각 유저가 Train기간(52주)동안 구입한 쿠폰 내역을 근거로, Test기간(1주)동안 구입할것으로 보이는 쿠폰을 유저당 10개씩 예측

데이터 설명

- 일단 다운로드 부터... (시간 오래걸림!)
  - 데이터는 캐글사이트에서 직접 받으셔도 되고 제 드랍박스에서 통째로 받으셔도 됩니다.
    https://www.dropbox.com/s/wzkx3tcchi9b1tg/input.zip?dl=0
  - 코드와 발표자료는 캐뽀 깃허브에서 받으실수있습니다
    https://github.com/KaggleBreak/problems
- 실습하실분은 R Studio도 설치해주세요

데이터 파일이 너무 많아요

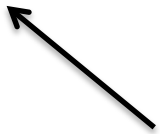
- user_list.csv
- coupon_list_train.csv
- coupon_list_test.csv
- coupon_area_train.csv
- coupon_area_test.csv
- coupon_detail_train.csv
- coupon_visit_train.csv
- prefecture_locations.csv
- sample_submissions.csv
- documentation/CAPSULE_TEXT_Translation.xlsx
- documentation/ERDiagram.xlsx

일단은 얘네만 쓰시면 됩니다

- user_list.csv (유저에 대한 정보)

- coupon_list_train.csv (Train기간동안 판매된 쿠폰에 대한 정보)
- coupon_detail_train.csv (Train기간동안 유저가 쿠폰을 구입한 정보)

- coupon_list_test.csv (Test기간동안 판매된 쿠폰에 대한 정보)
- coupon_detail_test.csv (Test기간동안 유저가 쿠폰을 구입한 정보)



이 정보를 예측하는게 목표

쿠폰에 관한 column들도 너무 많아요

- coupon_list_train

CAPSULE_TEXT, GENRE_NAME, PRICE_RATE, CATALOG_PRICE, DISCOUNT_PRICE, DISPFROM, DISPEND, DISPPERIOD, VALIDFROM, VALIDEND, VALIDPERIOD, USABLE_DATE_MON, USABLE_DATE_TUE, USABLE_DATE_WED, USABLE_DATE_THU, USABLE_DATE_FRI, USABLE_DATE_SAT, USABLE_DATE_SUN, USABLE_DATE_HOLIDAY, USABLE_DATE_BEFORE_HOLIDAY, large_area_name, ken_name, small_area_name, COUPON_ID_has

- coupon_detail_train

ITEM_COUNT, I_DATE, SMALL_AREA_NAME, PURSHASEID_has, USER_ID_hash, COUPON_ID_hash

일단은 얘네만 쓰시면 됩니다

- coupon_list_train

CAPSULE_TEXT, GENRE_NAME, PRICE_RATE, CATALOG_PRICE, DISCOUNT_PRICE, DISPFROM, DISPEND, DISPPERIOD, VALIDFROM, VALIDEND, VALIDPERIOD, USABLE_DATE_MON, USABLE_DATE_TUE, USABLE_DATE_WED, USABLE_DATE_THU, USABLE_DATE_FRI, USABLE_DATE_SAT, USABLE_DATE_SUN, USABLE_DATE_HOLIDAY, USABLE_DATE_BEFORE_HOLIDAY, large_area_name, ken_name, small_area_name, COUPON_ID_hash,

- coupon_detail_train

ITEM_COUNT, I_DATE, SMALL_AREA_NAME, PURSHASEID_has, USER_ID_hash, COUPON_ID_hash

그러면 남는 정보는

- 유저에 대한 정보
  - 어떤 쿠폰을 샀는가
- 쿠폰에 대한 정보
  - 장르 (GENRE_NAME)
  - 할인가격 (DISCOUNT_PRICE)
  - 판매기간 (DISPPERIOD)
  - 유효기간 (VALIDPERIOD)
  - 판매지역 (large_area_name, small_area_name)

실습

- 데이터로드 및 머지

modified_cos_sim.r 을 열어서
11행:

```
train <- merge(coupon_detail_train, coupon_list_train)
```

까지만 실행하면 데이터로드 및 머지 성공

R Studio 에서 일본어핸들링

- Tools -> Global Options
  - General 에서 Default text encoding을 UTF-8으로
  - Appearance 에서 Editor font를 MS Gothic으로

- read.csv가 에러날때는
  Sys.setlocale("LC_CTYPE", "C") 입력

- 콘솔창에서 일본어가 안보일때는
  - MAC의 경우는 Sys.setlocale("LC_CTYPE", "ja_JP") 실행하면 보일때도 있음
  - Windows의 경우는 그냥 포기하면 편함...

Solutions

- Recommendation System의 두가지 분류
  - Collaborative filtering
  - Content-based filtering

Collaborative Filtering

	쿠폰A	쿠폰B	쿠폰C	쿠폰D
유저1	o		o	o
유저2		o	o	
유저3	o	o		

나	o			o

나에게 추천할만한 쿠폰은?

Content-based Filtering

	장르	지역	가격
쿠폰A	고기집	서울	10000원
쿠폰B	피자할인	서울	14000원
쿠폰C	족발	경기	20000원

과거의 내가 산 쿠폰들

쿠폰D	피부관리	부산	76000원
쿠폰E	치킨	서울	15000원

아직 내가 사지않은 쿠폰들

나에게 추천할만한 쿠폰은?

그래서 Solution은?

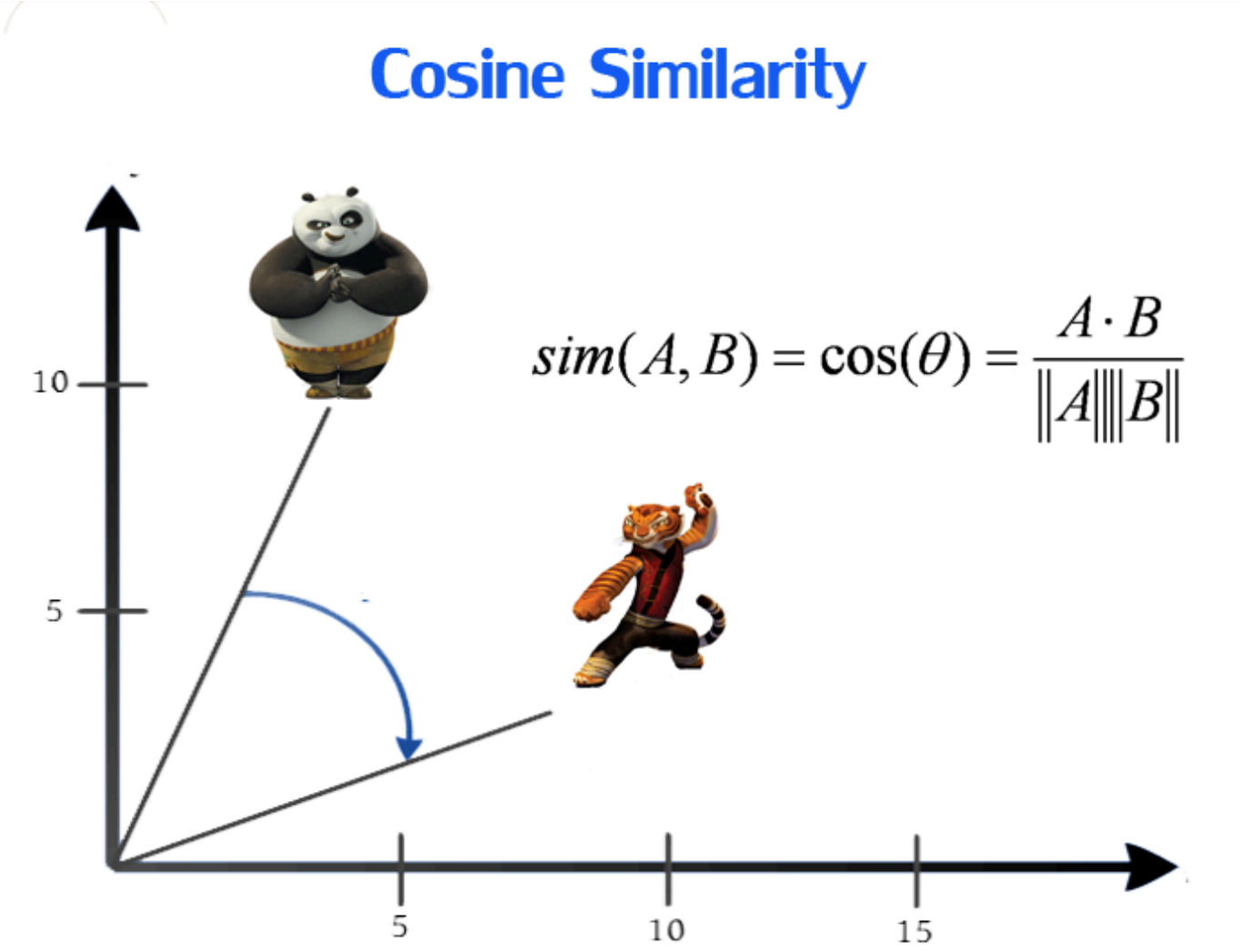
- Recommendation System의 두가지 분류
  - Collaborative filtering
  - Content-based filtering

무슨 알고리즘 쓰면 좋을지 모르겠어요

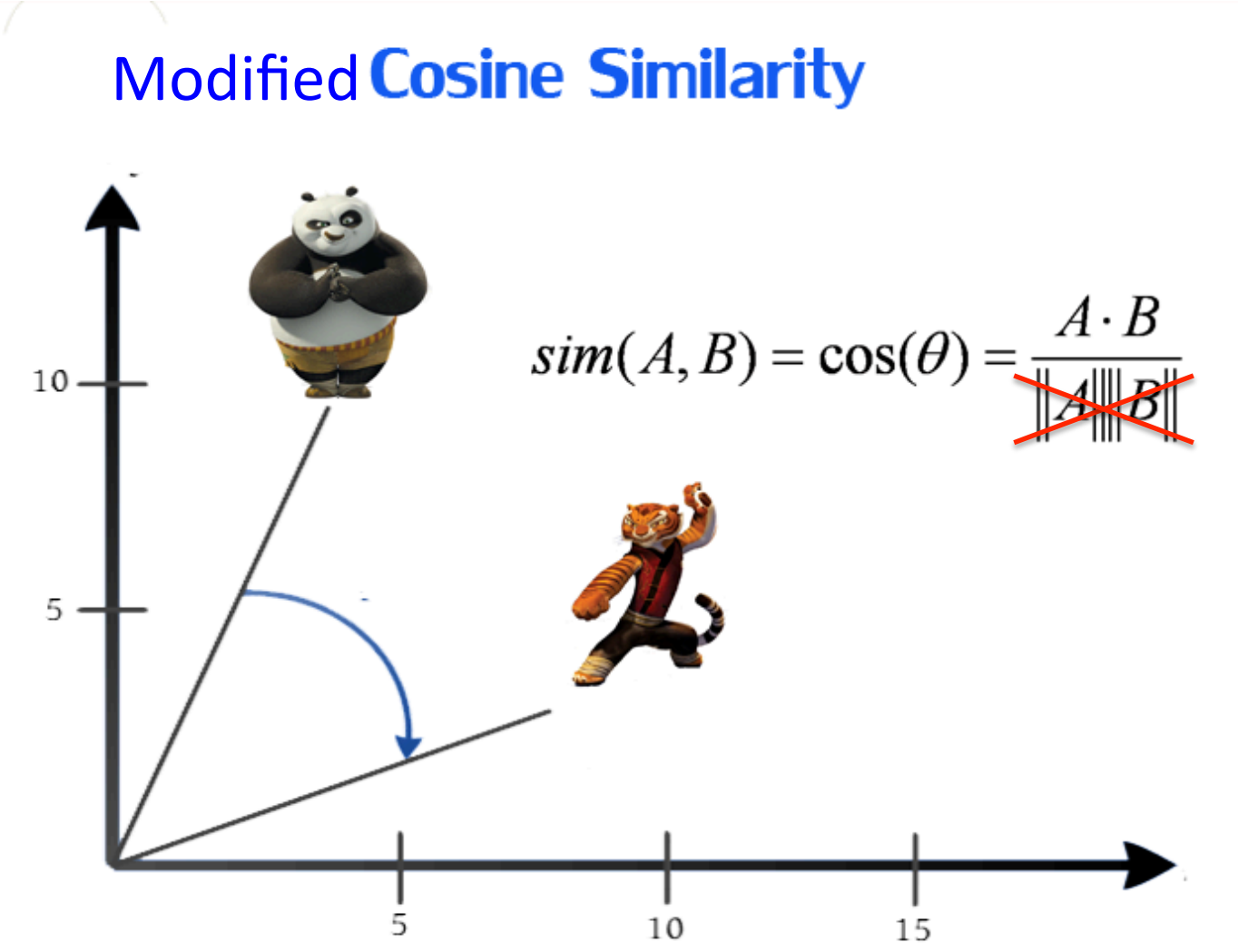
-> 포럼 뒤지면 다 나옵니다!

https://www.kaggle.com/c/coupon-purchase-prediction/scripts

Cosine Similarity



Modified Cosine Similarity



알고리즘의 전체적인 흐름

- 유저A가 Test기간동안 어떤 쿠폰을 구입할까 예측하려면

  1. 유저A가 Train기간동안 구입한 쿠폰들의 feature벡터의 평균을 구한다
     - 그 평균벡터가 유저A의 프로필벡터가 된다

  2. 유저A의 프로필벡터와 Test기간동안 판매된 쿠폰의 feature벡터의 코사인유사도를 구한다
     - 코사인유사도가 높을수록 유저A가 그 쿠폰을 구입할 가능성이 높다고 본다

  3. 코사인유사도가 높은 순으로 쿠폰을 정렬하면 끝!

실습

- 유저 X 쿠폰의 코사인유사도 구하기

- modified_cos_sim.r 을 끝까지 실행
- 빨리 끝나신분들은 further_modified_cos_sim.r 도 돌려보시면 좋아요

Evaluation Metric 알기

- Mean Average Precision @ 10

Submissions are evaluated according to the Mean Average Precision @ 10 (MAP@10):

$$MAP@10 = \frac{1}{|U|}\sum_{u=1}^{|U|}\frac{1}{min(m,10)}\sum_{k=1}^{min(n,10)}P(k)$$

where |U| is the number of users, P(k) is the precision at cutoff k, n is the number of predicted coupons, and m is the number of purchased coupons for the given user. If m = 0, the precision is defined to be 0.

Mean Average Precision = Average Precision 의 평균

Average Precision

Example) 유저1의 Average Precision @5

	Predicted Coupons				
유저1	쿠폰A	쿠폰B	쿠폰C	쿠폰D	쿠폰E

$$\frac{1/1 + 2/3 + 3/4}{3} = 0.806$$

빨간색쿠폰 : 유저가 실제로 구입한 쿠폰

Average Precision

Example) 유저1의 Average Precision @5

	Predicted Coupons Set #1				
유저1	쿠폰A	쿠폰B	쿠폰C	쿠폰D	쿠폰E

$$\text{유저1의 AP@5} = \frac{1/1 + 2/3 + 3/4}{3} = 0.806$$

	Predicted Coupons Set #2				
유저1	쿠폰A	쿠폰D	쿠폰C	쿠폰B	쿠폰E

$$\text{유저1의 AP@5} = \frac{1/1 + 2/2 + 3/3}{3} = 1.000$$

Mean Average Precision

- MAP는 각 유저의 AP의 평균

유저1의 AP@10 = 0.667
유저2의 AP@10 = 0.5
유저3의 AP@10 = 0.24
....
유저22873의 AP@10 = 0.54

MAP@10 = (0.667 + 0.5 + 0.24 + ... + 0.54) / 22873

※ 만약 유저가 Test기간중 아무 쿠폰도 구입을 안했다면?
-> 그 유저의 AP는 무조건 0

따라서, MAP의 최대값은,

$$\frac{\text{Test기간 중 쿠폰을 하나라도 구입한 유저수}}{\text{전체 유저수}}$$

Evaluation Metric – MAP@10

- 결론
  - 예측 순서도 중요함
  - 무조건 유저당 쿠폰 10개씩 꽉꽉 채워서 제출하는게 유리함

자유실습

- 최종 Submission까지 해봅시다!
- further_modified_cos_sim.r 의 Weight를 잘 조정하면 고득점도 가능!

Appendix

- Public Score & Private Score
  - Leader Board에 올라가는 score는 public score (test set의 30%만 사용하여 평가)
  - competition 기간이 끝나고 실제 순위 집계시에는 test set을 100% 사용하여 평가한 private score로 순위가 매겨진다

Appendix

- Local validation?