2018 DBC

2018 Daegue Bigdata Camp

@Daegue University

Fri. 29, Jun ~ Wed. 4, July

Data Introduction
&
Competition evaluation

데이터 소개

- 대형 마트에서
  - 소비자의 구매 품목을 바탕으로 (ex: 빵4개, 우유 2개, 책1권)
  - 소비자의 방문 목적을 classfy하는 문제 (ex : 퇴근길에 저녁거리 구매, 일주일치 생활용품 구매 등)

- Training data 75,675개, test data 20,000개
- Input의 차원은 54차원(상품 대분류 53개 + 방문 요일)
- 12개의 Class (A, B, C, …, L)
- 데이터 예시 :

ID	Books and magazines	DCD grocery	Frozen foods	Grocery dry goods	Liquor, wine, beer	Class
23240	1	3	1	43	2	B

competition 평가

원래는 class가 12개인데 예시에선 5개만 나타냈습니다.

- 제출 파일 형식

ID	A	B	C	D	E
1	0.054	0.002	0.056	0.188	0.7
7	0.764	0.002	0.01	0.142	0.082
8	0.176	0.004	0.06	0.188	0.572
9	0.58	0.002	0	0.324	0.094
000	0.21	0	0.022	0.636	0.132
000	0.654	0	0.004	0.228	0.114
000	0.368	0	0.076	0.486	0.07
000	0.804	0	0.004	0.1	0.092
000	0.03	0	0.146	0.314	0.51
000	0.712	0	0.026	0.13	0.132
000	0.864	0.002	0.002	0.028	0.104
000	0.83	0	0.006	0.108	0.056
000	0.818	0	0.002	0.124	0.056
000	0.356	0	0.016	0.512	0.116
000	0.394	0.002	0.006	0.504	0.094
000	0.638	0.002	0.012	0.224	0.124

competition 평가

원래는 class가 12개인데 예시에선 5개만 나타냈습니다.

- 제출 파일 형식

ID	A	B	C	D	E
1	0.054	0.002	0.056	0.188	0.7
7	0.764	0.002	0.01	0.142	0.082
8	0.176	0.004	0.06	0.188	0.572
9	0.58	0.002	0	0.324	0.094
000	0.21	0	0.022	0.636	0.132
000	0.654	0	0.004	0.228	0.114
000	0.368	0	0.076	0.486	0.07
000	0.804	0	0.004	0.1	0.092
000	0.03	0	0.146	0.314	0.51
000	0.712	0	0.026	0.13	0.132
000	0.864	0.002	0.002	0.028	0.104
000	0.83	0	0.006	0.108	0.056
000	0.818	0	0.002	0.124	0.056
000	0.356	0	0.016	0.512	0.116
000	0.394	0.002	0.006	0.504	0.094
000	0.638	0.002	0.012	0.224	0.124

test data의 ID

각 class에 속할 확률

competition 평가

- 제출 파일 형식

ID	A	B	C	D	E
1	0.054	0.002	0.056	0.188	0.7
2	0.764	0.002	0.01	0.142	0.082
4	0.176	0.004	0.06	0.188	0.572
5	0.58	0.002	0	0.324	0.094
6	0.21	0	0.022	0.636	0.132
9	0.654	0	0.004	0.228	0.114
10	0.368	0	0.076	0.486	0.07
11	0.804	0	0.004	0.1	0.092
12	0.03	0	0.146	0.314	0.51
13	0.712	0	0.026	0.13	0.132
14	0.864	0.002	0.002	0.028	0.104
15	0.83	0	0.006	0.108	0.056
16	0.818	0	0.002	0.124	0.056
17	0.356	0	0.016	0.512	0.116
18	0.394	0.002	0.006	0.504	0.094
23	0.638	0.002	0.012	0.224	0.124

원래는 class가 12개인데 예시에선 5개만 나타냈습니다.

모델의 예측 결과 1번 데이터가
- A일 확률 : 5.4%
- B일 확률 : 0.02%
- C일 확률 : 5.6%
- D일 확률 : 18.8%
- E일 확률 : 70%

test data의 ID

각 class에 속할 확률

competition 평가

- **평가 기준 : Multi-class logarithm loss**

Kaggle에서 쓰는 거임.
믿어도 됨.

$$logloss = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}\log(p_{ij})$$

- N : test set의 데이터 개수(20000)
- M : outcome의 가짓수(A,B,…,L => 12)
- y_{ij} : 실제 정답 - 데이터 i가 class j에 속하면 1, 아니면 0
- p_{ij} : 모델이 예측한 데이터 i가 class j에 속할 확률
- log함수가 ∞로 발산하는 것을 막기 위해 각 예측값들은 10^{-15}와 $1-10^{-15}$ 사이로 조정됨
  (i.e. 학생들이 0,1로 파일 보내도 조교가 $10^{-15}, 1-10^{-15}$로 수정)
- 낮은 logloss = good

competition 평가

- **평가 기준 : Multi-class logarithm loss**

원래는 class가 12개인데 예시에선 5개만 나타냈습니다.

$$logloss = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}\log(p_{ij})$$

- 예시

ID	A	B	C	D	E
1	0.054	0.002	0.056	0.188	0.7

: 모델의 예측값

ID	A	B	C	D	E
1	0	0	0	0	1

: 실제 정답

$$\begin{aligned}\sum_{j=1}^{M} y_{ij}\log(p_{ij}) &= 0\times\log(0.054)\\ &+0\times\log(0.002)\\ &+0\times\log(0.056)\\ &+0\times\log(0.188)\\ &+1\times\log(0.7) \quad = \quad -0.3566\end{aligned}$$