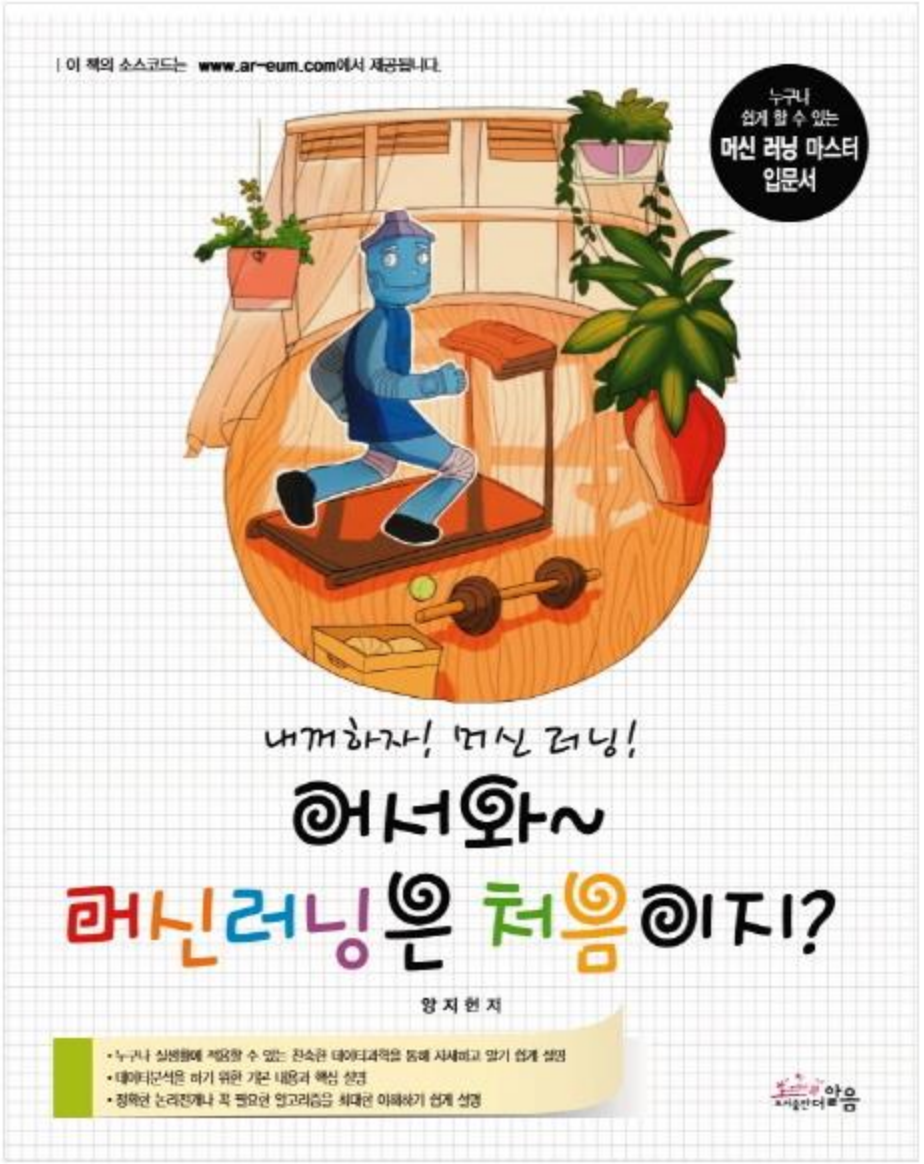
[어서와~ 머신러닝은 처음이지?]

8장. Naïve Bayes

- 장형석
- 국민대학교 빅데이터경영MBA과정 교수
- 숙명여자대학교 빅데이터센터 연구소장
- chjang1204@nate.com
- 010-3302-5543



- **저자 : 양지헌**
  - 숭실대학교 물리학과 졸업
  - 국민대학교 빅데이터경영MBA과정 석사졸업
  - 국민대학교 데이터사이언스 박사과정
  - 전) VTW 컨설턴트
  - 전) 글로벌텔레콤 IOT 분석팀장
  - 데이터 분석 전문가(ADP) 자격 보유
- **출판사 : 더알음**
- **출간일 : 2016년 12월 21일**
- **ISBN : 9791195484737**
- **http://www.ar-eum.com**

1. 생활에서 만나는 문제

1-1) 고객의 속성 ⇔ 영화 취향

K 대학의 대학원생 강승리 양은 C 사의 영화마케팅 관련부서에 인턴으로 들어갔다. 부장은 새로 들어온 신입사원들의 역량을 평가하기 위하여 각각 자유주제로 발표를 하도록 하였다. 승리는 무엇을 발표할지 고민을 하다가 영화를 보러 오는 관객들의 속성들에 대해서 탐구해보기로 마음을 먹었다. 우선은 매표소 앞에서 설문지로 관객들의 속성과 함께 특별히 좋아하는 장르에 대하여 간단한 질의조사를 하였다. 고객의 속성은 영화의 취향과 관련이 있을 것 같은 것 몇 가지를 추려보았는데 그 중에서 나이,성별,직업,결혼,이성친구로 하기로 최종 결론을 내리고 제대로 된 설문지만 추려서 다음과 같은 40 건의 데이터 셋을 만들었다. (설문조사를 해 본 사람은 알겠지만 그렇게 쉬운 작업은 아니다.) 직업은 너무 세부적이지 않게 큰 카테고리로 나누었다.

1. 생활에서 만나는 문제

1-2) 데이터셋

	A	B	C	D	E	F
1	나이	성별	직업	결혼여부	이성친구	장르
2	20대	여	디자이너	NO	YES	로맨틱
3	40대	남	홍보/마케팅	NO	NO	공포
4	10대	여	홍보/마케팅	NO	YES	로맨틱
5	20대	여	디자이너	YES	NO	로맨틱
6	10대	남	학생	NO	NO	액션
7	40대	남	자영업	NO	NO	공포
8	10대	남	학생	NO	NO	액션
9	30대	남	IT	NO	YES	SF
10	30대	남	언론	YES	NO	스릴러
11	40대	남	자영업	NO	NO	공포
12	10대	남	학생	NO	NO	액션
13	20대	여	홍보/마케팅	YES	NO	로맨틱
14	30대	여	IT	YES	NO	SF
15	30대	남	언론	YES	NO	스릴러
16	10대	여	학생	NO	YES	로맨틱

Feature's

Label

1. 생활에서 만나는 문제

2-1) 키워드 ⇔ 스팸 여부

모바일 쇼핑몰의 마케팅 부서에서 일하는 미혜는 사내메일로도 spam 메일이 많아서 짜증이 난다. 그래서 그룹웨어 운영팀의 개발직원에게 의뢰해서 특정한 키워드가 들어가 있는 메일은 아예 spam 편지함으로 넣기로 했다. 아마도 주로 '광고','회원가입'등의 단어가 spam 메일에 포함되어 있을 것인데 이 단어들로 정말 spam 필터링이 잘 되는지를 미리 테스트할 수 있는 모델이 없을까? 이 단어가 포함되어 있으면 spam 일 확률이 얼마가 나올 것이다라는 정보도 알 수가 있었으면 좋겠다. 미혜는 자신의 20 개정도의 메일을 분석하여 자신이 뽑은 몇 가지 키워드가 메일문서 안에 포함이 되어 있는지를 보여주는 행렬을 다음과 같이 엑셀로 만들었다.

1. 생활에서 만나는 문제

2-2) 데이터셋

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	문서번호	무료	빅데이터	상담	쇼핑	컨설팅	광고	방송	대학교	수강	회원	서류	모집	메일종류
2	1	1		1			1							spam
3	2		1			1								ham
4	3					1			1	1				ham
5	4		1									1		ham
6	5	1					1	1						spam
7	6													spam
8	7		1							1		1		ham
9	8					1			1					ham
10	9		1									1		ham
11	10				1						1		1	spam
12	11								1	1				ham
13	12		1			1				1				ham
14	13					1						1		ham
15	14			1	1						1			spam
16	15				1			1					1	spam
17	16		1						1			1		ham
18	17			1				1					1	spam
19	18		1							1		1		ham
20	19					1			1					ham

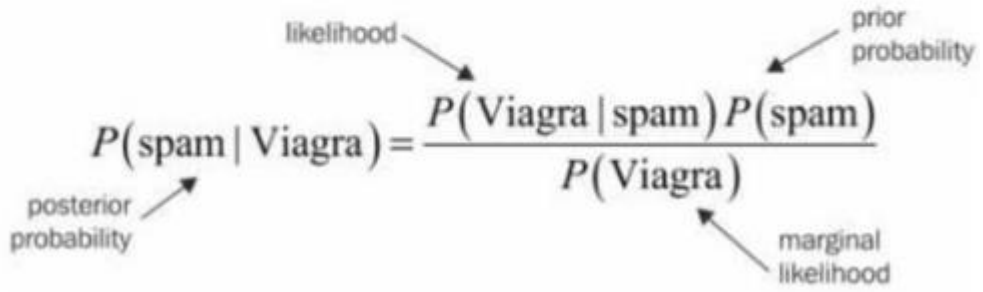
Feature's

Label

1. 생활에서 만나는 문제

2) 나이브 베이즈 알고리즘

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{P(A \cap B)}{P(B)}$$



p(B) : 나이가 20 대이고 성별이 여자이고 직업이 IT 이고 미혼일 확률(사전확률)

p(A) : '공포'영화를 선택할 확률

p(B) : '비아그라'라는 단어가 포함될 확률(사전확률)

p(A) : mail 이 'spam'일 확률

2. 코딩과 구체적인 구현

1-1) 나이브 베이즈 구현 – 영화 장르

학습

```
> movie <- read.csv("movie.csv" , header = T)
> library(e1071)
> nm <- naiveBayes(movie[1:5] , movie$장르 , laplace = 0 )
> head(movie)
  나이 성별         직업 결혼여부 이성친구   장르
1 20대  여     디자이너       NO      YES 로맨틱
2 40대  남 홍보/마케팅       NO       NO   공포
3 10대  여 홍보/마케팅       NO      YES 로맨틱
4 20대  여     디자이너      YES       NO 로맨틱
5 10대  남         학생       NO       NO   액션
6 40대  남       자영업       NO       NO   공포
```

예측

```
> result <- predict(nm , movie[1:5])
> sum(movie$장르 != result)
[1] 1
> result
 [1] 로맨틱 공포   로맨틱 로맨틱 액션   공포   액션   SF     스릴러 공포   액션   로맨틱 SF     스릴러 로맨틱
[16] 코미디 로맨틱 스릴러 코미디 액션   로맨틱 코미디 공포   공포   SF     스릴러 로맨틱 코미디 SF     액션
[31] 무협   스릴러 무협   SF     무협   공포   무협   무협   로맨틱
Levels: SF 공포 로맨틱 무협 스릴러 액션 코미디
```

1개만 틀림

2. 코딩과 구체적인 구현

1-2) 나이브 베이즈 구현 – 스팸 유무

학습

```
> mail <- read.csv("spam.csv" , header = T)
> mail[is.na(mail)] <- 0
> nm2 <- naiveBayes(mail[2:13] , mail$메일종류 , laplace = 0)
> head(mail)
  문서번호 무료 빅데이터 상담 쇼핑 컨설팅 광고 방송 대학교 수강 회원 서류 모집 메일종류
1        1    1        0    1    0      0    1    0      0    0    0    0    0     spam
2        2    0        1    0    0      1    0    0      0    0    0    0    0      ham
3        3    0        0    0    0      1    0    0      1    1    0    0    0      ham
4        4    0        1    0    0      0    0    0      0    0    0    1    0      ham
5        5    1        0    0    0      0    1    1      0    0    0    0    0     spam
6        6    0        0    0    0      0    0    0      0    0    0    0    0     spam
```

예측

```
> result2 <- predict(nm2 , mail[2:13])
> sum(mail$메일종류 != result2)
[1] 1
> result2
 [1] spam ham  ham  ham  spam ham  ham  ham  ham  spam ham  ham  ham  spam spam ham  spam ham  ham  spam
Levels: ham spam
```

1개만 틀림

2. 코딩과 구체적인 구현

2-1) 예측 모형 – 영화 장르

```
> nm

Naive Bayes Classifier for Discrete Predictors

Call:
naiveBayes.default(x = movie[1:5], y = movie$장르, laplace = 0)

A-priori probabilities:
movie$장르
       SF      공포    로맨틱      무협    스릴러      액션    코미디
0.1282051 0.1282051 0.2307692 0.1282051 0.1282051 0.1282051 0.1282051

Conditional probabilities:
          나이
movie$장르      10대      10대      20대      30대      40대
    SF     0.0000000 0.0000000 0.0000000 1.0000000 0.0000000
    공포   0.0000000 0.0000000 0.0000000 0.0000000 1.0000000
    로맨틱 0.4444444 0.0000000 0.5555556 0.0000000 0.0000000
    무협   0.2000000 0.0000000 0.2000000 0.2000000 0.4000000
    스릴러 0.0000000 0.0000000 0.0000000 1.0000000 0.0000000
    액션   0.2000000 0.8000000 0.0000000 0.0000000 0.0000000
    코미디 0.0000000 0.0000000 0.4000000 0.4000000 0.2000000
```

2. 코딩과 구체적인 구현

2-2) 예측 모형 – 스팸 유무

```
> nm2

Naive Bayes Classifier for Discrete Predictors

Call:
naiveBayes.default(x = mail[2:13], y = mail$메일종류, laplace = 0)

A-priori probabilities:
mail$메일종류
 ham spam
 0.6  0.4

Conditional probabilities:
                무료
mail$메일종류   [,1]      [,2]
          ham   0.000 0.0000000
          spam 0.375 0.5175492

              빅데이터
mail$메일종류       [,1]      [,2]
          ham   0.5833333 0.5149287
          spam 0.0000000 0.0000000
```

3. 나머지 문제

1) 스팸메일의 핵심 키워드 추출

```
> library(KoNLP)
> txt <- readLines('spam.txt')
> place <- sapply(txt , extractNoun , USE.NAMES = F)
> useSejongDic()
Backup was just finished!
87007 words were added to dic_user.txt.
> c <- unlist(place)
> place <- Filter(function(x) {nchar(x) >=2} , c)
> res <- str_replace_all(place , "[^[:alpha:]]" , "")
> res <- res[res != ""]
```

```
> res
 [1] "안녕"     "고객"     "전세계"   "서비스"   "HP"       "지속"     "노력"
 [8] "일환"     "저희"     "제품"     "구매"     "하시"     "고객"     "서비스"
[15] "향상"     "주기"     "프로그램" "도입"     "하게"     "며칠"     "저희"
[22] "HP"       "체험"     "고객"     "의견"     "제공해주십사" "메일"     "메일"
[29] "시간"     "간단"     "설문조사" "참여"     "감사"     "클릭"     "설문조사"
[36] "연결"     "설문조사" "시작"     "여러분"   "의견"     "저희"     "고객"
[43] "만족"     "향상"     "단기적"   "장기"     "도움"     "저희"     "여러분"
[50] "답변"     "감사"
```

```
> wordcount <- table(res)
> wordcount2 <- sort(table(res) , decreasing=T)
> wordcount2
res
      고객       저희   설문조사         HP       감사       메일   서비스   여러분
         4          4          3          2          2          2          2          2
      의견       향상       간단       구매       노력     단기적       답변       도움
         2          2          1          1          1          1          1          1
      도입       만족       며칠       시간       시작       안녕       연결       일환
         1          1          1          1          1          1          1          1
      장기     전세계 제공해주십사       제품       주기       지속       참여       체험
         1          1          1          1          1          1          1          1
      클릭 프로그램       하게       하시
         1          1          1          1
```

3. 나머지 문제

2) 시각화 : wordcloud

감사합니다.