데이터마이닝(DataMining)

Chapter 5.1. 단순베이즈분류

- 단순베이즈분류는 문서분류(spam 또는 legitimate, sports 또는 politics 등), 의료진단 등에 많이 사용
- 사후 확률이 큰 집단으로 새로운 데이터를 분류
- 조건부 독립의 가정이 비현실적인 측면이 있으나 계산이 간편하여 널리 이용
- 단순베이즈분류(naive Bayes classification) 모형은, 베이즈 정리에 기반한 방법으로, 사후 확 률(일종의 조건부 결합확률)의 계산 시 조건부 독립을 가정하여 계산을 단순화한 방법
- 적절한 전처리 과정을 거친 단순베이즈분류는 서포트벡터머신을 포함한 보다 발전된 기법과도 경쟁 가능

단순베이즈분류

- 단순베이즈분류기는 연속형 또는 이산형에 관계없이 임의 크기의 예측변수를 다룰 수 있음
- 데이터 $x = (x_1, x_2, \cdots, x_p)$으로 주어질 때, 이 데이터가 k집단으로부터 나왔을 사후확률은, 베이즈 정리로부터

$$P(Y = k|X = x) = \frac{P(Y = k)P(X = x|Y = k)}{P(X = x)}, \qquad j = 1,2, \cdots, K$$

단순베이즈분류

- 일반적인 베이즈분류에서는 위의 사후확률이 가장 큰 집단으로 개체에 대한 분류를 수행
- 단순베이즈분류는 위의 사후확률의 계산을 좀 더 편하게 할 수 있도록 예측변수들간의 독립을 가정

$$P\left(X_1 = x_1, \cdots X_p = x_p \middle| Y = k\right) = \prod_{j=1}^{p} P(X_j = x_j | Y = k)$$

$$P\left(Y = k \middle| X_1 = x_1, \cdots X_p = x_p\right) = \frac{P(Y = k)\prod_{j=1}^{p} P\left(X_j = x_j \middle| Y = k\right)}{\prod_{j=1}^{p} P\left(X_j = x_j\right)}$$

을 이용하여 사후확률의 분자를 계산하고, 그 결과를 이용하여 분류를 수행

- 이 방법은 계산을 크게 단순화 시켜주며, 예측변수의 수가 많은 경우에도 적용이 편리

단순베이즈분류

- 훈련자료를 이용하여 모든 j와 k에 대하여 추정값 $\hat{P}(Y=k)$와 $\hat{P}\left(X_j = x_j \middle| Y=k\right)$)을 얻은 후 주어진 검증자료 $z=(z_1, z_2, \cdots, z_p)$

$$\arg\max_k \left(\hat{P}(Y=k) \prod_{j=1}^{p} \hat{P}\left(X_j = z_j \middle| Y=k\right) \right)$$

로 예측하며 입력변수가 연속형인 경우에는 흔히 구간을 나눠서 범주형으로 변환

- 이 방법은 계산을 크게 단순화 시켜주며, 예측변수의 수가 많은 경우에도 적용이 편리

- 일반화 가법모형과의 관계

$$\log\left(\frac{P\left(Y=1\middle|X_1=x_1,X_2=x_2,\cdots X_p=x_p\right)}{P\left(Y=0\middle|X_1=x_1,X_2=x_2,\cdots X_p=x_p\right)}\right)$$

$$=\log\left(\frac{P(Y=1)}{P(Y=0)}\right)+\sum_{j=1}^{p}\left(\log\left(\frac{P\left(X_j=x_j\middle|Y=1\right)}{P\left(X_j=x_j\middle|Y=0\right)}\right)\right)$$

$$=\alpha+\sum_{j=1}^{p}f_j\left(x_j\right)$$

단순베이즈분류

- 5개의 학습문서

문서번호	주요단어	문서분류
1	fun, couple, love, love	comedy
2	fast, furious, shoot	action
3	couple, fly, fast, fun, fun	comedy
4	furious, shoot, shoot, fun	action
5	fly, fast, shoot, love	action

- 입력문서가 {fast, furious, fun}을 주요단어로 가질 때, 사후확률
- $P(comedy|x) = P(comedy) \cdot P(fast|comedy) \cdot P(furious|comedy) \cdot P(fun|comedy) =$

  $\frac{2}{5} \cdot \frac{1}{9} \cdot \frac{0}{9} \cdot \frac{3}{9} = 0$

단순베이즈분류

- $P(action|x) = P(action) \cdot P(fast|action) \cdot P(furious|action) \cdot P(fun|action) = \frac{3}{5} \cdot \frac{2}{11} \cdot \frac{2}{11} \cdot \frac{1}{11} = 0.0018$
- 따라서 입력문서는 사후확률이 보다 큰 action으로 분류

- 단순베이즈분류에서 낮은-빈도 문제(low-frequency problem)에 주의할 필요
- comedy 문서에서는 furious 단어의 빈도가 0이므로, furious 단어를 포함하는 새로운 자료에 대한 사후확률은 항상 0
- 이러한 문제점을 해결하기 위해 모든 속성값-군집 조합에 대한 빈도에 작은 수를 더해 주어 계산 을 수행

단순베이즈분류

- 여러 개의 연속형 예측변수를 가지는 경우
- 총 8명에 대해 키, 몸무게, 발 크기를 측정한 훈 련자료

성별	키(feet)	몸무게(lbs)	발 크기(inches)
남성	6	180	12
남성	5.92	190	11
남성	5.58	170	12
남성	5.92	165	10
여성	5	100	6
여성	5.5	150	8
여성	5.42	130	7
여성	5.75	150	9

단순베이즈분류

- 세 변수가 모두 독립이며, 정규분포를 따른다고 가정
- 모집단의 평균과 분산

성별	키		몸무게		발 크기	
	평균	분산	평균	분산	평균	분산
남성	5.855	3.5033e-02	176.26	1.2292e+02	11.25	9.1667e-01
여성	5.4175	9.7225e-02	132.5	5.5833e+02	7.5	1.6667e+00

- 남성과 여성그룹에 속할 사전확률을 $P(\text{남성}) = 0.5, P(\text{여성}) = 0.5$
- 확률은 큰 모집단에서 의 빈도에 기초하거나 훈련자료에 기초하여 주어질 수 있음

단순베이즈분류

- 새로운 자료가 남자인지 여자인지를 분류

성별	키	몸무게	발 크기
표본(x)	6	130	8

- 주어진 자료에 대한 사후확률
- $P(\text{남성}|x) = P(\text{남성}) \cdot P(\text{키}|\text{남성}) \cdot P(\text{몸무게}|\text{남성}) \cdot P(\text{발크기}|\text{남성}) \approx 6.1984 \cdot 10^{-9}$
- $P(\text{여성}|x) = P(\text{여성}) \cdot P(\text{키}|\text{여성}) \cdot P(\text{몸무게}|\text{여성}) \cdot P(\text{발크기}|\text{여성}) \approx 5.3778 \cdot 10^{-9}$
- 주어진 자료는 사후확률이 보다 큰 여성으로 예측
- $P(\text{키}|\text{남성}) = \phi(6; 5.885,\ 3.5033e-02) \approx 1.5789$

단순베이즈분류

- 단순베이즈분석을 위해 iris 자료를 사용
- R 패키지 {e1071}의 naiveBayes() 함수를 이용하여 단순베이즈 분류를 수행

```
> m <- naiveBayes(Species ~ ., data = iris)
> m
 Naive Bayes Classifier for Discrete Predictors

 Call:
 naiveBayes.default(x = X, y = Y, laplace = laplace)
                                  # Laplacian(add-1) smoothing
 A-priori probabilities:
 Y
    setosa versicolor virginica
0.3333333  0.3333333 0.3333333
```

단순베이즈분류

```
Conditional probabilities:
             Sepal.Length
Y             [,1]      [,2]
  setosa     5.006 0.3524897    # mean(Sepal.Length[Species==“setosa”])
  versicolor 5.936 0.5161711
  virginica  6.588 0.6358796

             Sepal.Width
Y             [,1]      [,2]
  setosa     3.428 0.3790644
  versicolor 2.770 0.3137983
  virginica  2.974 0.3224966

             Petal.Length
Y             [,1]      [,2]
  setosa     1.462 0.1736640
  versicolor 4.260 0.4699110
  virginica  5.552 0.5518947

             Petal.Width
Y             [,1]      [,2]
  setosa     0.246 0.1053856
  versicolor 1.326 0.1977527
  virginica  2.026 0.2746501
```

단순베이즈분류

- predict() 함수를 이용하여 예측을 실시하고, 그 결과를 정오분류표로 나타냄

```
> table(predict(m, iris), iris[,5])
           setosa versicolor virginica
setosa         50          0         0
versicolor      0         47         3
virginica       0          3        47
```

단순베이즈분류

- 패키지 {klaR}을 이용하여 단순베이즈분류를 수행
- spam 자료는 4601개의 이메일(관측치)에서 등장하는 단어의 종류와 관련된 58개 변수로 구성
- 48개 변수(A.1~A.48)는 총 단어 수 대비 해당 단어의 출현비율을 나타냄
- 6개 변수(A.49~A.54)는 총 문자 수 대비 특정 문자의 출현비율을 나타냄
- 3개 변수 (A.55~A.57)는 연속되는 대문자 철자의 평균길이, 최대길이, 대문자의 총수를 나타냄
- 마지막 변수(A.58)는 스팸메일의 여부(1:spam, 0:non-spam)를 나타냄

```
> data(spam, package="ElemStatLearn")
> library(klaR)
```

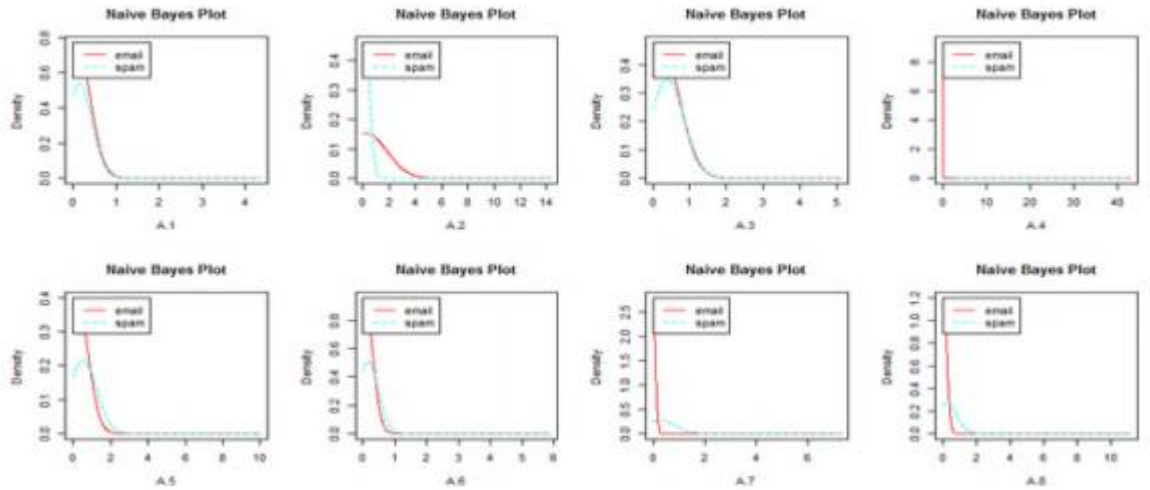
단순베이즈분류

- 전체 자료의 2/3를 훈련용 자료로 하여 NaiveBayes() 함수를 통해 단순베이즈분류를 수행

```
> train.ind <- sample(1:nrow(spam), ceiling(nrow(spam)*2/3),
                      replace=FALSE)
> nb.res <- NaiveBayes(spam ~ ., data=spam[train.ind,])
```

```
> # 결과 보여주기
> opar <- par(mfrow=c(2,4))
> plot(nb.res)
Hit <Return> to see next plot:
```

단순베이즈분류



- 결과(그림)는 57개의 예측변수별 분포를 문서의 종류별(spam, non-spam)로 그린 것
- 새로운 자료가 주어질 때, 사후확률은 사전확률과 위 확률들의 곱을 통해 구할 수 있음

단순베이즈분류

```
> par(opar)
```

- 분석에 제외된 검증용 자료를 이용하여 모형의 정확도를 구하기

```
> nb.pred <- predict(nb.res, spam[-train.ind,])
> confusion.mat <- table(nb.pred$class, spam[-train.ind,"spam"])
> confusion.mat
       email spam
email    517   33
spam     422  561
```

```
> sum(diag(confusion.mat))/sum(confusion.mat)
 [1] 0.7031963
```

단순베이즈분류

- 단순베이즈분류는 결측값을 포함하는 자료를 다음과 같이 처리
  - 훈련단계: 속성값-군집 조합에 대한 빈도 계산 시 결측값을 포함하는 케이스가 제외됨
  - 분류단계: 결측인 속성이 계산과정에서 생략

- 결측값을 포함하는 자료에 대해 단순베이즈분류를 수행
- 단순베이즈분류에서는 결측값에 대한 처리가 매우 유연하게 이루어짐
- 모형구축에서는 결측값을 포함하는 케이스를 제외하며, 분류과정에서는 결측 속성에 대한 확률만 계산에서 제외되므로 수행과정에 문제가 없음

단순베이즈분류

- HouseVote{mlbench} 자료는 미국의 하원의원 435명(민주당:267명, 공화당:168명)의 16개 주요법안에 대한 찬성여부를 조사한 자료
- R 패키지 {e1071}를 이용하여 단순베이즈분류를 수행

```
> install.packages("e1071")
> library (e1071)
> install.packages("mlbench")
> data (HouseVotes84, package="mlbench")
> head(HouseVotes84)
```

	Class	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15	V16
1	republican	n	y	n	y	y	y	n	n	n	y	<NA>	y	y	y	n	y
2	republican	n	y	n	y	y	y	n	n	n	n	n	y	y	y	n	<NA>
3	democrat	<NA>	y	y	<NA>	y	y	n	n	n	n	y	n	y	y	n	n
4	democrat	n	y	y	n	<NA>	y	n	n	n	n	y	n	y	n	n	y
5	democrat	y	y	y	n	y	y	n	n	n	n	y	<NA>	y	y	y	y
6	democrat	n	y	y	n	y	y	n	n	n	n	n	n	y	y	y	y

단순베이즈분류

```
> summary(HouseVotes84)
        Class         V1          V2          V3          V4          V5
democrat  :267   n   :236   n   :192   n   :171   n   :247   n   :208
Republican:168   y   :187   y   :195   y   :253   y   :177   y   :212
                 NA's: 12   NA's: 48   NA's: 11   NA's: 11   NA's: 15
   V6          V7          V8          V9          V10         V11         V12
n   :152   n   :182   n   :178   n   :206   n   :212   n   :264   n   :233
y   :272   y   :239   y   :242   y   :207   y   :216   y   :150   y   :171
NA's: 11   NA's: 14   NA's: 15   NA's: 22   NA's:  7   NA's: 21   NA's: 31
  V13         V14         V15         V16
n   :201   n   :170   n   :233   n   : 62
y   :209   y   :248   y   :174   y   :269
NA's: 25   NA's: 17   NA's: 28   NA's:104
```

단순베이즈분류

```
> model <- naiveBayes(Class ~ ., data = HouseVotes84)
> pred <- predict(model, HouseVotes84[,-1])
> tab <- table(pred, HouseVotes84$Class)
> tab
 pred        democrat republican
   democrat       238         13
   republican      29        155
```

```
> table(HouseVotes84$Class)
 democrat    republican
      267           168
```

```
> sum(tab[row(tab)==col(tab)])/sum(tab)
 [1] 0.9034483
```

단순베이즈분류

- 단순 베이즈 가정은 고차원 확률 추정 문제를 반복적인 일차원 확률 추정 문제로 단순화
- 비현실적인 가정에도 불구하고 단순 베이즈 분류법은 매우 복잡한 문제에서도 효율적인 경우를 흔히 볼 수 있음