나이브 베이즈 학습방법

최호식

경기대학교 응용통계학과

Oct, 2019

학습내용

1. 단순 베이즈 분류기(naive Bayes classifier)
2. 다범주(multi-category) 분류

기호

- 훈련자료 $\{(x_i, y_i)\}_{i=1}^{n}$, $x_i \in \mathbb{R}^p, y_i \in \mathcal{K}$
- 회귀: $\mathcal{K} = \mathbb{R}$
- 분류: $\mathcal{K} = \{-1, +1\}$

Preliminaries I

- 조건부 확률(conditional probability)

- 두 확률변수 X, Y에 대해,

$$\Pr(Y|X) \equiv \frac{\Pr(X,Y)}{\Pr(X)} \tag{1}$$

$$\Pr(Y|X)\Pr(X) = \Pr(X,Y)$$

- Bayes' rule(베이즈 법칙)

- 두 확률변수 X, Y에 대해,

$$\begin{aligned} \Pr(Y|X) &= \Pr(X,Y)/\Pr(X) && (2) \\ &= \Pr(X|Y)\Pr(Y)/\Pr(X) && (3) \end{aligned}$$

이 성립함.

Preliminaries II

- 조건부 독립(conditional independence)

- 세 확률변수 X_1, X_2, Y에 대해

$$\Pr(X_1, X_2|Y) = \Pr(X_1|Y)\Pr(X_2|Y) \quad (4)$$
$$= \prod_{j=1}^{2} \Pr(X_j|Y) \quad (5)$$

이 성립할 때 조건부 독립이라 함.

- 기호는 $X_1 \perp X_2 | Y$라 표시.
- 보다 직관적인 정의는 $\Pr(X_1|X_2, Y) = \Pr(X_1|Y)$

Preliminaries III

- 사전확률(prior), likelihood, 사후확률(posterior)

$$\underbrace{\Pr(Y|X)}_{posterior} = \underbrace{\Pr(X|Y)}_{likelihood}\underbrace{\Pr(Y)}_{prior}/\Pr(X) \tag{6}$$

- 예) 만약, $Y = 1$ 또는 0의 값을 가지는 binary 변수일때,

$$\underbrace{\frac{\Pr(Y=1|X)}{\Pr(Y=0|X)}}_{\text{posterior ratio}} = \underbrace{\frac{\Pr(X|Y=1)}{\Pr(X|Y=0)}}_{\text{likelihood ratio}} \times \underbrace{\frac{\Pr(Y=1)}{\Pr(Y=0)}}_{\text{prior ratio}} \tag{7}$$

$$\log\frac{\Pr(Y=1|X)}{\Pr(Y=0|X)} = \log\frac{\Pr(X|Y=1)}{\Pr(X|Y=0)} + \log\frac{\Pr(Y=1)}{\Pr(Y=0)} \tag{8}$$

단순 베이즈 분류 I

편의상 X의 차원은 2로 고정.

- 입력변수의 값이 $x = (x_1, x_2)$로 주어졌을 때 $Y = k$일 사후확률

$$\begin{aligned} & \Pr(Y = k | X_1 = x_1, X_2 = x_2) \\ \propto\ & \Pr(X_1 = x_1, X_2 = x_2 | Y = k) \Pr(Y = k) \end{aligned}$$

단순 베이즈 가정(naive Bayes assumption): 조건부 독립

$$\Pr(X_1 = x_1, X_2 = x_2 | Y = k) = \prod_{j=1}^{2} \Pr(X_j = x_j | Y = k)$$

하에서 사후확률은

$$\Pr(Y = k | X_1 = x_1, X_2 = x_2) \propto \Pr(Y = k) \prod_{j=1}^{2} \Pr(X_j = x_j | Y = k)$$

로 표현

단순 베이즈 분류 II

- 훈련자료를 이용하여 모든 추정값

$$\hat{\Pr}(Y = k)$$

와

$$\hat{\Pr}(X_j = x_j | Y = k)$$

을 얻은 후 주어진 시험자료 $z = (z_1, z_2)$에 대하여 Y를

$$\arg\max_{k \in \mathcal{K}} \left(\hat{\Pr}(Y = k) \prod_{j=1}^{2} \hat{\Pr}(X_j = z_j | Y = k) \right)$$

로 예측
- 입력변수가 연속형인 경우에는 흔히 구간을 나눠서 범주형으로 변환 또는 정규분포 활용
- 특정 변수에서의 확률 추정값이 0이면 다른 변수에서의 확률 추정값이 큰 값을 가지더라도 그 곱은 항상 0이 되는 문제 발생

단순 베이즈 분류 III

- 라플라스 수정(Laplace correction)

  (예) 세 클래스의 자료의 수가 0, 990, 10

  - 라플라스 수정: 1, 991,11
  - 확률추정값:

$$
\begin{aligned}
& 0, 0.990, 0.010 \\
\rightarrow \quad & \frac{1}{1003} = 0.001, \ \frac{991}{1003} = 0.988, \ \frac{11}{1003} = 0.011
\end{aligned}
$$

  - 수정을 하지 않았을 때와 큰 차이가 없고 추정값이 정확히 0이 되어 생기는 문제가 발생하지 않음

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

참고 I

- 베이즈룰 유투브

  - `https://www.youtube.com/watch?v=TfeaZ_26iQk&list=PLpl-gQkQivXiBmGyzLrUjzsblmQsLtkzJ&index=6`

- Y(HIV감염여부)와 세 검사방법(X_1,X_2,X_3)에 따른 양성판정유무($+,-$) 자료

  - 어떤 사람에 대한 세 검사방법의 결과가 $(+,+,+)$와 같을 때, 이 사람이 HIV에 감염되었을 확률을 Naive Bayes 분류 방법에 의해 구하여 보자.

	y	x_1	x_2	x_3
1	HIV	$+$	$+$	$-$
2	HIV	$+$	$-$	$+$
3	HIV	$+$	$-$	$+$
4	Normal	$-$	$+$	$+$
5	Normal	$+$	$-$	$-$
6	Normal	$+$	$-$	$-$
7	Normal	$-$	$-$	$-$

참고 II

	y	x_1	x_2	x_3
1	HIV	$+$	$+$	$-$
2	HIV	$+$	$-$	$+$
3	HIV	$+$	$-$	$+$
4	Normal	$-$	$+$	$+$
5	Normal	$+$	$-$	$-$
6	Normal	$+$	$-$	$-$
7	Normal	$-$	$-$	$-$

(8)번식에서 naive bayes 가정 적용

$$\begin{aligned}\log\frac{\Pr(Y=1|X)}{\Pr(Y=0|X)} &= \log\frac{\Pr(X|Y=1)}{\Pr(X|Y=0)}+\log\frac{\Pr(Y=1)}{\Pr(Y=0)}\\ &= \sum_{j=1}^{3}\log\frac{\Pr(X_j=x_j|Y=1)}{\Pr(X_j=x_j|Y=0)}+\log\frac{\Pr(Y=1)}{\Pr(Y=0)}\end{aligned}$$

그러므로 $X=(x_1,x_2,x_3)=(+,+,+)$인 경우

$$\begin{aligned}\log\frac{\Pr(Y=1|X=(+,+,+))}{\Pr(Y=0|X=(+,+,+))} &= \log\frac{3/3}{2/4}+\log\frac{1/3}{1/4}+\log\frac{2/3}{1/4}+\log\frac{3/7}{4/7}\\ &= \log 2+\log\frac{4}{3}+\log\frac{8}{3}+\log\frac{3}{4}=1.674>0\end{aligned}$$

Multi-category classification I

- 이진 분류(binary classification): $\mathcal{K} = \{-1, +1\}$
- 다범주 분류(multi-category classification): $\mathcal{K} = \{1, 2, \cdots, K\}$

- (예. $K = 3$) 다범주에 대한 소프트맥스(softmax) 분류방법

$$\log \frac{\Pr(Y = 1|X)}{\Pr(Y = 3|X)} = f_1(X) \tag{9}$$

$$\log \frac{\Pr(Y = 2|X)}{\Pr(Y = 3|X)} = f_2(X) \tag{10}$$

$$\log \frac{\Pr(Y = 3|X)}{\Pr(Y = 3|X)} = f_3(X). \tag{11}$$

Multi-category classification II

- 편의상 '$|X$' 삭제, 실질적으로 $\mathbf{f}_3(\mathbf{X}) = \mathbf{0}$으로 '**3**'이 기준범주임.

$$
\begin{aligned}
\Pr(Y = 1) &= e^{f_1} \Pr(Y = 3) & (12)\\
\Pr(Y = 2) &= e^{f_2} \Pr(Y = 3) & (13)\\
\Pr(Y = 3) &= e^{f_3} \Pr(Y = 3) & (14)
\end{aligned}
$$

그러면,

$$
\begin{aligned}
1 &= \Pr(Y = 1) + \Pr(Y = 2) + \Pr(Y = 3)\\
&= \Pr(Y = 3)(e_1^f + e_2^f + e_3^f)
\end{aligned}
$$

$$
\therefore \Pr(Y = k|X) = \frac{e^{f_k(X)}}{e^{f_1(X)} + e^{f_2(X)} + e^{f_3(X)}}, k = 1, 2, 3
$$

실습 I

```
library(e1071)
data(iris)
# 모형적합
model = naiveBayes(Species~., data=iris)
# 종분류결과1
pred1 <- predict(model, newdata=iris, type="class")
table(pred1, iris$Species)

# 종분류결과2
pred2 <- predict(model, newdata=iris, type="raw")
haty <- apply(pred2, 1, which.max)
table(iris$Species, haty)
```