# 8장 나이브 베이즈 분류기

## 1. 서론

범주형 예측변수들을 갖는 데이터에 적용할 수 있는 나이브 베이즈 분류기 (Naive Bayes)

기본원리 : 예측될 각 레코드에 대해서

1. 예측변수 프로파일이 동일한 (즉, 예측변수 값들이 동일한) 다른 모든 레코드들을 찾는다.
2. 그 레코드들이 어떤 클래스에 속하고 어떤 클래스가 가장 지배적인지 결정
3. 그 클래스를 새로운 레코드에 배정

↳ 이거 대신에 관심있는 클래스에 속할 경향이 무엇인가?

⇒ 클래스확률을 임계으로서 어떤 클래스 $C_i$가 해당 레코드에 대한 가장 그럴듯한 클래스가 아니더라도, 그 레코드를 $C_i$에 속하는 것으로 분류하도록 컷오프값 조정 가능

↳ 규정하고자 원하는 관심있는 특정 클래스가 있고, 보다 많은 레코드들이 여기에 속한다고 분류할 때 유용

- 컷오프값 확률 방법

1. 레코드가 해당 클래스에 속한다고 간주되는 관심있는 클래스에 대한 컷오프 확률 설정
2. 새로운 레코드와 동일한 예측변수 프로파일을 갖는 모든 학습레코드들을 찾는다
3. 그 레코드들이 관심있는 클래스에 속하는 확률을 계산
4. 그 확률이 컷오프 확률보다 크면 새로운 레코드를 관심있는 그 클래스로 배정

- 조건부 확률

사건 B가 발생했을 때 사건 A의 확률 $P(A|B)$

→ 해당 레코드의 예측변수 값들이 $X_1, X_2, \cdots, X_p$ 값을 취할 때, 그 레코드가 $C_i$에 속할 확률.

⇒ $P(C_i|X_1, X_2, \cdots, X_p)$ 레코드 하나를 분류하기 위해서 이런 방법으로 각 클래스에 속할 확률을 계산

가장 큰 확률을 갖는 클래스로 분류 / 관심있는 클래스로 배정하기위해 컷오프값 이용

↳ 베이지안 분류기는 범주형 변수에만 작동 (수치형 변수들이 이렇게 동일한 값을 가질 확률은 희박)

↳ 범주형 예측변수들에 특히 적합한 유일한 분류나 예측 방법

↳ 수치형 변수들은 그룹화되어 범주형 변수로 변환 필요

## 2. 완전한 (정확한) 베이지안 분류기의 적용

- "가장 가능성있는 클래스에 배정"하는 방법 사용

↳ 데이터에서 가장 많은 클래스에 압도적 배정 → 관심있는 클래스 혹은 더 위험가능성이 큰 부류를 어떻게 할 것인가?

· 귀오프 확률방법의 사용

⇒ 관심있는 클래스에 더 많은 레코드를 배정하기 위해 귀오프값 사용 (오차는 좀 더 커질 수 있다)

· 완전한 (정확한) 베이즈 절차의 실용적 어려움

↳ 예측변수들이 모두 동일한 레코드 ⇒ 샘플에서 완전히 일치하는 데이터를 찾는 것과 같다

↳ 레코드의 수와 예측변수의 수가 많아지면 찾기 힘들다! (ex: 대한민국에서 '김'씨 성을 가지고 부산에 살며 자식이 셋이고 차량을 가지고 연 소득이 5천만원 이하인 40대 남성 과 같은 레코드)

⇒ 일반적으로 적은 예측변수의 수일지라도 완전 일치는 어려움

· 해결책: 나이브 베이즈

그 레코드와 정확하게 일치하는 레코드의 확률계산에 집중하지 않는다

⇒ 전체 데이터셋 사용

분류절차의 변경

① $C_i$ 클래스에 대해 각 예측변수별 조건부 확률 $P(X_j/C_i)$ ($C_i$중에 $X_j$값을 가진 레코드의 비율에 의해 추정)을 추정 ⇒ 각 예측변수들이 $C_i$클래스에 발생할 확률

② 이 확률을 서로 곱한 후 $C_i$ 클래스에 속하는 레코드의 비율을 곱함

③ ①,② 단계를 모든 클래스에 대해서 반복

④ 클래스 $C_i$에 대해 ②에서 계산된 값에 모든 클래스에 대한 값의 합으로 나눔 ⇒ 클래스 $C_i$에 대한 확률

⑤ 이 레코드를 예측변수 값들에 대해서 가장 큰 확률값을 갖는 클래스로 배정

ex) 예측변수 $X_1, X_2, \cdots, X_p$ 클래스 $C_i$에 속할 확률은

$$P(C_i|X_1,X_2,\cdots,X_p) = \frac{P(C_i)[P(X_1|C_i)P(X_2|C_i)\cdots P(X_p|C_i)]}{P(C_1)[P(X_1|C_1)\cdots P(X_p|C_1)]+\cdots+P(C_m)[P(X_1|C_m)\cdots P(X_p|C_m)]}$$

다소 계산 복잡 피벗테이블로 활용가능

· 조건부 독립의 나이브 베이즈 가정

특정 클래스 $P(X_1,X_2,\cdots,X_p|C_i)$ 내에서 예측변수 프로파일 $X_1,X_2,\cdots,X_p$를 갖는 레코드의 정확한 조건부 확률을 개별적인 조건부 확률 $P(X_1|C_i)\times P(X_2|C_i)\times\cdots\times P(X_p|C_i)$ 의 곱으로 추정 ⇒ 각 예측변수들이 독립이면 2개의 값이 같다

→ 각 레코드에 대한 정확한 경향이 아니라 어떤 클래스에 속할 경향이 높은가의 경향 문제이기 때문에 완벽한 독립이 아니어도 어느 정도 잘 들어맞는다 → 정확한 값은 기대하면 안됨

## 3. 나이브 베이즈 분류기의 장점과 단점

장점: 단순성, 계산효율성, 좋은 분류성능, 범주형 변수를 직접 다룰 수 있음

↳ 예측변수들의 독립성이 한참 떨어져도 좋은 결과를 내기도 함 / 예측변수의 수가 많을 때 두드러짐

단점

1. 좋은 결과를 얻기 위해서 매우 많은 수의 레코드 필요

2. 예측변수 범주가 학습데이터에 나타나지 않을 경우 예측변수의 해당 범주를 가지는 새로운 레코드의 확률은 0이라고 가정

↳ 드문 예측변수 값이 등장할 때 문제가 될 수 있다

3. 어떤 클래스에 속할 확률에 따라서 레코드들의 분류나 순위를 구하는 것이 목표일 때는 좋은 성능

↳ 클래스 경향도(확률)를 '추정'하는 것이 목표라면 매우 편향된 결과를 내놓는다