출처가 명시되지 않은 모든 자료(이미지 등)는 조성현 강사님 블로그 및 강의 자료 기반.

<< 머신러닝 - 분류 >>

[Naive Bayes]

데이터의 각 모든 차원의 feature가 서로 조건부 독립이라고 가정한다. 이후 분류하고자 하는 target과 feature 간 조건부 확률을 계산해서 데이터를 분류한다. *조건부 독립 가정*이 없으면 그 의미도 사라지는 알고리즘이다.

> 데이터를 전처리한 후, feature 간 독립이 되도록 데이터 feature를 구성했다고 `Naive`하게 생각한다. 물론 현실 세계에서 feature 간 상관성이 아예 없을 수는 없다. 나이브 베이즈 알고리즘의 취약점이기도 하지만, 일단 이렇게 나이브한 가정을 하지 않으면 나이브 베이즈 알고리즘으로 분류를 하는 의미가 없다.

이러한 가정을 바탕으로, 통계학의 베이즈 정리(베이즈 정리의 자세한 내용은 여기를 참고하자.)를 활용해 각 데이터를 분류하는 기법이 머신러닝의 **나이브 베이지안 분류** 기법이다.

1. 분류 원리

- **베이즈 정리**

$$P(A_1|B) = \frac{P(B \cap A_1)}{P(B)}$$

$$= \frac{P(B|A_1)P(A_1)}{P(B)}$$

$$= \frac{P(B|A_1)P(A_1)}{P(B|A_1)P(A_1) + P(B|A_2)P(A_2)}$$

출처 : 위키피디아

나이브 베이즈 분류기의 핵심은, 예측할 데이터의 feature가 학습 데이터셋에서 얼마나 나타났는지 확률을 바탕으로, **더 높은 확률을 갖는 타겟값을 선택**하는 것이다.

아래와 같은 학습 데이터가 주어진 상태에서, 새로운 시험 데이터가 들어왔을 때의 target 값을 예측하는 상황을 가정해 보자.

학습 데이터

Home owner	Marital status	Annual income	Defaulted borrower
Yes	Single	125K	No
No	Married	100K	No
No	Single	70K	No
Yes	Married	120K	No
No	Divorced	95K	Yes
No	Married	60K	No
Yes	Divorced	220K	No
No	Single	85K	Yes
No	Married	75K	No
No	Single	90K	Yes

시험 데이터

Home owner	Marital status	Annual income	Defaulted borrower
No	Married	120K	?

X Y (target)

베이즈 정리를 이용해, 시험 데이터의 feature들을 이용해 target이 Yes일 확률과 No일 확률을 계산한다. 그리고 더 큰 확률을 갖는 target을 해당 데이터의 target 값으로 예측하는 것이다.

자세한 계산 과정은 아래 과정과 같다.

NO	YES
• Home과 Marital feature는 범목형 (categorical)이므로 학습데이터를 이용하 아래와 같이 쉽게 추정할 수 있다. $P(Home = No\|No) = \frac{4}{7}$ $P(Marital = Married\|No) = \frac{4}{7}$ • Annual feature와 같이 연속적인 실숫값은 정규분포를 이용해서 아래 pdf를 계산한다. $P(Annual = 120K\|No) \rightarrow \frac{1}{\sqrt{2\pi}\sigma} exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$ $\mu_N = \frac{125+100+70+\cdots+75}{7} = 110$ $\sigma_N^2 = \frac{(125-110)^2+\cdots+(75-110)^2}{7-1} = 2975$ $\sigma_N = \sqrt{\sigma_N^2} = 54.54$	$P(Annual = 120K\|No) \rightarrow \frac{1}{\sqrt{2\pi}*54.54} exp\left[-\frac{(120-110)^2}{2*2975}\right] = 0.007193$ • 이 값은 pdf이므로 확률을 의미하지는 않는다. 정확하는 ϵ이 작은 상수일 때 Annual이 120K ~ 120K+ϵ 사이에 있을 확률을 구해야 한다. ϵ은 상수로 곱해지는 값이므로 (연적구할 때 말면같은) $P(X\|Y)$의 근사치를 구하는 데 사용될 수 있다. • 이 재료들을 이용해서 $P(X\|Y = No)$의 확률을 계산한다. $P(Home = No, Marital = Married, Annual = 120K \mid Defaulted = No) =$ $= P(Home = No\|No) * P(Marital = Married\|No) * P(Annual = 120K\|No)$ $= \frac{4}{7} \times \frac{4}{7} \times 0.007193 = 0.002349$ • Bayes 정리를 이용해서 시험데이터 (X)의 target (Y)이 "No"일 확률을 계산한다. $P(Y = No\|X) = \frac{P(X\|Y = No) \cdot P(No)}{P(X)} = \frac{0.002349 \times \frac{7}{10}}{P(X)} = \frac{0.001644}{P(X)}$	• Home과 Marital feature는 범목형 (categorical)이므로 학습데이터를 이용하 아래와 같이 쉽게 추정할 수 있다. $P(Home = No\|Yes) = \frac{3}{3}$ $P(Marital = Married\|Yes) = \frac{0}{3}$ • Annual feature와 같이 연속적인 실숫값은 정규분포를 이용해서 아래 pdf를 계산한다. $P(Annual = 120K\|Yes) = \frac{1}{\sqrt{2\pi}\sigma} exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$ $\mu_Y = \frac{95+85+90}{3} = 90$ $\sigma_Y^2 = \frac{(95-90)^2+(85-90)^2+(90-90)^2}{3-1} = 25$ $\sigma_Y = \sqrt{\sigma_Y^2} = 5$	$P(Annual = 120K\|Yes) = \frac{1}{\sqrt{2\pi}*5.54} exp\left[-\frac{(120-110)^2}{2*2975}\right] = 1.2 \times 10^{-9}$ • 이 재료들을 이용해서 $P(X\|Y = Yes)$의 확률을 계산한다. $P(Home = No, Marital = Married, Annual = 120K \mid Defaulted = Yes) =$ $= P(Home = No\|Yes) * P(Marital = Married\|Yes) * P(Annual = 120K\|Yes)$ $= \frac{3}{3} \times \frac{0}{3} \times 1.2 \times 10^{-9} = 0$ • Bayes 정리를 이용해서 시험데이터 (X)의 target (Y)이 "Yes"일 확률을 계산한다. $P(Y = Yes\|X) = \frac{P(X\|Y = Yes) \cdot P(Yes)}{P(X)} = \frac{0 \times \frac{3}{10}}{P(X)} = \frac{0}{P(X)}$ $P(Y = No\|X) > P(Y = Yes\|X)$ → 시험 데이터의 target을 "No"로 분류한다.

2. 평활화

조건부 확률을 기반으로 타겟을 예측하는 알고리즘 특성 상, 각 feature 중 하나라도 등장 확률이 0인 것이 있으면, 해당 타겟 벡터가 등장할 것이라 예측되는 확률 역시 0이 되어 버리는 문제가 발생한다.

이 문제를 해결하기 위해 우도(likelihood) 값이 0이 되지 않도록 스무딩하는 방법을 사용할 수 있다.

m-추정치

분자, 분모에 임의의 수를 더해 준다. 수학적으로는 그렇게 하면 안 되지만, 정확한 확률 값이 궁금한 게 아니라, 대소 비교만 하면 되기 때문에 확률이 0이 되는 것을 회피하는 것이 더 중요하다. 따라서 특정 수를 더해서 스무딩하는 방법을 사용한다.

알파 값에 따라서 스무딩 방식이 달라진다. 1을 더하면 Laplace 방식, 1보다 작은 값을 더하면 Lidstone 방식이라고 하는데, Scikit-learn의 나이브베이즈 분류기에서는 이 방식을 선택할 수 있다.

3 . 모델 학습

Scikit-learn 라이브러리에서는 세 가지 종류의 나이브베이즈 분류기를 제공한다.

- GaussianNB : 정규분포
- BernoulliNB : 베르누이분포
- MultinomialNB : 다항분포

feature가 전부 실수인 경우 GaussianNB 분류기를, 범주형인 경우에는 BernoulliNB, MultinomialNB 분류기를 사용한다. 그 중에서도 전자는 분류해야 할 클래스의 수가 2개일 때(이진 분류), 후자는 분류해야 할 클래스의 수가 3개 이상일 때 사용한다.

GaussianNB 분류기는 정규분포를 바탕으로 하기 때문에, 확률값이 0이 나올 수가 없다. 따라서 평활화 보정을 해줄 필요가 없다.

실습 1. Iris Dataset

4개의 feature가 전부 실수형이기 때문에, GaussianNB 분류기를 사용한다.

실습 2. Income Dataset

범주형 변수(다중 분류)와 실수형 변수가 섞여 있기 때문에, MultinomialNB 분류기와 GaussianNB 분류기를 모두 사용한다. 두 개의 분류기를 모두 사용하여 확률값을 계산한 뒤, 곱한다. 조건부 독립을 가정하기 때문에, 곱하는 게 가능하다는 것에 유의하자.

배운 점, 더 생각해볼 점

- 이전에 프로젝트할 때, 댓글 데이터에 나이브 베이즈 적용했을 때는 정확도가 70% 이상이 나오지 않았었다. 무슨 문제였을까? 문서 분류/텍스트 분류에 나이브 베이즈 분류기 많이 사용한다는데..