[머신러닝 - 분류]

데이터 전처리 후 특징 추출. 사람이 feature 선정해야 한다. feature 간 상관성이 적게끔 데이터를 구성할 것이기 때문에, feature들은 서로 독립적이라고 봐도 크게 무리는 없다. 완전히 독립적일 수는 없다. 어느 정도는 독립적이다. 상관성이 적기 때문에. 이런 가정 하에 데이터 분석을 하는 것을 생각해 보자.

Naive Bayes

데이터의 각 모든 차원의 feature가 서로 조건부 독립이라고 가정한다. 그 가정이 없으면 의미가 없는 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.

- **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 선택하는 것이다.

m-추정치

하나라도 0인 확률이 있으면, 전체적으로 아무리 다른 값이 나온다고 해도 결과값이 0이 나온다. 대소 비교를 할 수 없는 현상이 발생하는 것이다. 이 문제를 해결하기 위해 m-estimates라는 방법을 사용한다.

분자, 분모에 임의의 수를 더해 준다. 수학적으로는 그렇게 하면 안 되지만, 0을 회피해야 한다. 정확한 확률 값이 궁금한 게 아니라, 어차피 대소 비교만 하면 되기 때문에 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 있따.

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 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.

배운 점, 더 생각해볼 점

- 이전에 프로젝트할 때, 텍스트 분류에 나이브 베이지안 분류기 많이 사용한다는 것 알았었다. 다만 댓글 데이터에 나이브 베이즈 적용했을 때는 정확도가 70% 이상이 나오지 않았었다. 스팸 분류 등

더 공부해 보고, 임베딩 다시 해서 나이브 베이지안 적용해 보자.