Naive bayes classifier

# 라이브러리 링크

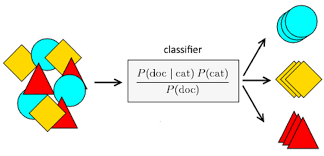
**install :**

<https://scikit-learn.org/stable/install.html>

**github :**

[https://github.com/scikit-learn/scikit-learn/blob/a9bf7f38d7d074962e8afe2d10c9e41bd8117f39/sklearn/naive\_bayes.py](https://github.com/scikit-learn/scikit-learn/blob/82df48934eba1df9a1ed3be98aaace8eada59e6e/sklearn/tree/_classes.py)

# 기초 설명

****

나이브 베이즈 분류(Naïve Bayes Classification)는 특성들 사이의 독립을 가정하는 베이즈 정리를 적용한 확률 분류기의 일종이며 이러한 나이브 베이즈 분류는 텍스트 분류에 사용됨으로써 문서를 여러 범주 (예: 스팸, 스포츠, 정치)중 하나로 판단하는 문제에 대한 대중적인 방법으로 남아있다.

# 버전 정보

* NumPy >= 1.14.6 (pip install numpy)
* Scipy >= 1.1.0 (pip install scipy)
* Joblib >= 0.11 (pip install joblib
* Threadpoolctl >= 2.0.0 (pip install threadpoolctl)

# 데이터셋 설명 및 출처

* fetch\_20newsgroups: The 20 newsgroups dataset comprises around 18000 newsgroups posts on 20 topics split in two subsets: one for training (or development) and the other one for testing (or for performance evaluation). The split between the train and test set is based upon a messages posted before and after a specific date.
* Sources : <https://github.com/scikit-learn/scikit-learn/blob/534045dc582967261c0e67e2e63e28ecdac048cd/sklearn/datasets/descr/twenty_newsgroups.rst>

# 코드 설명

* Sklearn 에서 제공하는 fetch\_20newsgroups의 train 데이터셋을 받은 후, 데이터셋 내부의 뉴스 기사에 대한 String 데이터를 DTM TF\_IDF 형태로 가공한다.
* 가공된 데이터와 각 뉴스에 대한 뉴스 카테고리 정보를 naive bayes model에 학습하여 학습된 모델을 만든뒤 이 모델을 사용하여 데이터셋 외부의 실제 자동차 뉴스 기사의 String 데이터를 입력한뒤 이 실제 뉴스 기사의 카테고리를 분류하고 그 결과를 출력한뒤 학습모델의 전체적인 정확도를 accuracy 함수를 통해 출력하는 코드.

# 검증 방법

* 코드 내부에서 sklearn 에서 제공하는 학습용 데이터셋으로 만든 학습 모델을 실제 현실의 데이터를 가져와 테스트하며.
* 또한 sklean 에서 제공하는 검증용 데이터셋을 사용하여 검증용 데이터셋의 실제 값과 학습모델의 예측값 사이의 적합도를 accuracy\_score 함수를 사용하여 평가함.