**나이브 베이즈 분류 & KNN**

**나이브 베이즈 분류(Naïve Bayes Classification)**

**Theory**

나이브 베이즈 분류는 데이터에서 변수들에 대한 조건부 독립을 가정하는 알고리즘으로 클래스에 대한 사전 정보와 데이터로보터 추출된 정보를 결합하고, **베이즈 정리**를 이용하여 어떤 데이터가 특정 클래스에 속하는지를 분류하는 알고리즘이다.

나이브 베이즈 분류는 하나의 속성 값을 기준으로 다른 속성이 독립적이라 전제했을 때 해당 속성 값이 클래스 분류에 미치는 영향을 측정한다. 속성값에 대해 다른 속성이 독립적이라는 가정은 **클래스 조건 독립성(Class Conditional Independece)**이라 한다.

**베이즈 정리(Bayes Theorem)**

나이브 베이즈 알고리즘의 기본이 되는 개념으로, 두 확률 변수의 사전 확률과 사후 확률 사이의 관계를 나타내는 정리이다. 사건 A와 B가 있을 때, 사건 B가 일어난 것을 전제로 한 사건 A의 조건부 확률을 구하고자 한다. 이 때, A의 확률, B의 확률 그리고 사건 A가 일어난 것을 전제로 한 사건 B의 조건부 확률을 알면 해당 확률을 계산할 수 있다는 것이 베이즈 정리이다.

**베이즈 정리 공식(Bayes Theorem Formula)**

* P(A|B) = P(A^B) / P(B) = P(A)\*P(B|A) / P(A)\*P(B|A) + P(A\*)\*P(B|A\*)
* P(A) = 사전확률, P(B) = 관찰값, P(A|B) = 사후확룰, P(B|A) = 우도

**KNN(K-Nearest Neighbors Classifier)**

**Theory**

K-NN은 어떤 범주로 나누어져 있는 데이터 셋이 있을 때, 새로운 데이터가 추가된다면 이를 어떤 범주로 분류할 것인지를 결정할 때 사용할 수 있는 분류 알고리즘이다. K-NN은 새로운 데이터의 클래스를 해당 데이터와 **거리 상 가장 가까운 k개 데이터들의 클래스(범주)**로 결정한다.

K-NN은 최근접 이웃 간의 거리를 계산할 때 유클리디안 거리, 맨하탄 거리, 민코우스키 거리 등을 사용할 수 있으며, 대표적으로 유클리디안 거리를 사용한다.

K-NN은 특징 공간에 분포하는 데이터에 대해 K개의 가장 가까운 이웃을 살펴보고 다수결 방식으로 데이터의 레이블을 할당하는 방식으로 특징 공간에 있는 모든 데이터에 대한 정보가 필요하다. 따라서 데이터셋이 크다면, 많은 메모리 공간과 계산 시간이 필요하다는 단점이 있지만, 반면에 알고리즘이 매우 단순하고 직관적이며, 사전 학습이나 특별한 준비가 필요 없다는 특징이 있다.

**하이퍼파라미터**

* N\_neighbors : default =5
* Weights : [‘uniform’, ‘disance’] default = ‘uniform’