◦ 분류 (Classification) 는 독립변수 값이 주어졌을 때, 그 독립변수 값과 가장 연관이 큰 종속변수 (카테고리) 를 찾아주는 알고리즘. 연관성이라는 카테고리 혹은 클래스라 미리 주어졌다는 점에서 뒤에 나올 군집화 알고리즘 알고리즘과 다르다고 볼 수 있다.

- 분류 모형의 종류

◦ 분류알고리즘을 푸는 방법은 크게 두가지로 나눌 수 있다.

판별함수 모형 (discriminant function)

↳ 카테고리에 따라 서로 다른 영역을 나누는 boundary를 찾은 다음, 주어진 데이터가 어디에 위치해 있는지를 판별함수로 예측하는 방법

◦ Perceptron / support vector machine / Neural network

확률적 모형

↳ 조건부 확률을 기반으로 각각의 카테고리를 계산한다. 조건부 확률을 계산하는 방식에 따라 두가지로 나뉜다. 조건부 확률은

◦ Linear, Quadratic discriminant Analysis / Naive Bayesian (확률적 "생성" 모형)

◦ Logistic regression / Decision tree (확률적 "판별" 모형)

확률적 모형의 정의

훈련데이터 N개와 K개의 클래스 $1, \cdots K$ 중의 하나의 값을 가진다고 가정하자.
확률적모형은 다음과 같은 입력 x에 대한 클래스를 예측한다.

(1) 입력 x가 주어졌을 때, 각각 클래스 k가 될 확률 $P(y=k \mid x)$을 모두 계산하고,

$$P_1 = P(y=1 \mid x)$$
$$\vdots$$
$$P_K = P(y=K \mid x)$$

(2) 다음으로는 가장 확률이 큰 클래스를 선택하는 방법이다.

$$y = \arg\max_k P(y=k \mid x)$$

Scikit-learn에서 조건부 확률은 생성모형은 모두 "predict_proba()"와 "predict_log_proba()"를 지원한다.

확률적 생성 모형의 정의

- 생성모형은 모든 각 클래스별로 데이터 독립변수 $P(x \mid y=k)$를 추정한 다음 베이즈 정리를 사용하여 $P(y=k \mid x)$를 계산하는 방법이다.

$$P(y=k \mid x) = \frac{P(x \mid y=k)\, P(y=k)}{P(x)}$$

- 생성모형에서는 전체 확률법칙을 이용하여 특징데이터 x의 주변확률 $P(x)$를 구할 수 있다. (모든 클래스를 생각할 수 있다)

$$P(x) = \sum_{k=1}^{K} P(x \mid y=k)\, P(y=k)$$

↳ 계산량이 많다는 단점

나이브 베이즈 (Naive bayesian)

- 조건부확률 기반 생성 모형의 정확한 계산은 클래스가 3개면 충분히 바로 적용할 수 있다는 장점이다. 나이브 베이즈는 독립적 조건부 확률로 모형을 하나 이다.

확률적 판별모형

- QDA나 나이브베이즈 방법은 베이지안 방법을 하는 조건부확률 기반 생성 모델이었다. 조건부확률 기반 생성 모델은 조건부 확률 $P(y \mid x)$를 구하기 위해 likelihood $P(x \mid y)$를 구하고 베이즈 정리를 사용해 조건부 확률을 계산한다.

- 확률적 판별모형은 조건부확률 $P(y=1 \mid x)$이 x에 대한 함수 $f(x)$로 표현될 수 있다고 가정하고 그 함수를 직접 찾아내는 방법이다.

$$P(y=k \mid x) = f(x)$$

분류모형