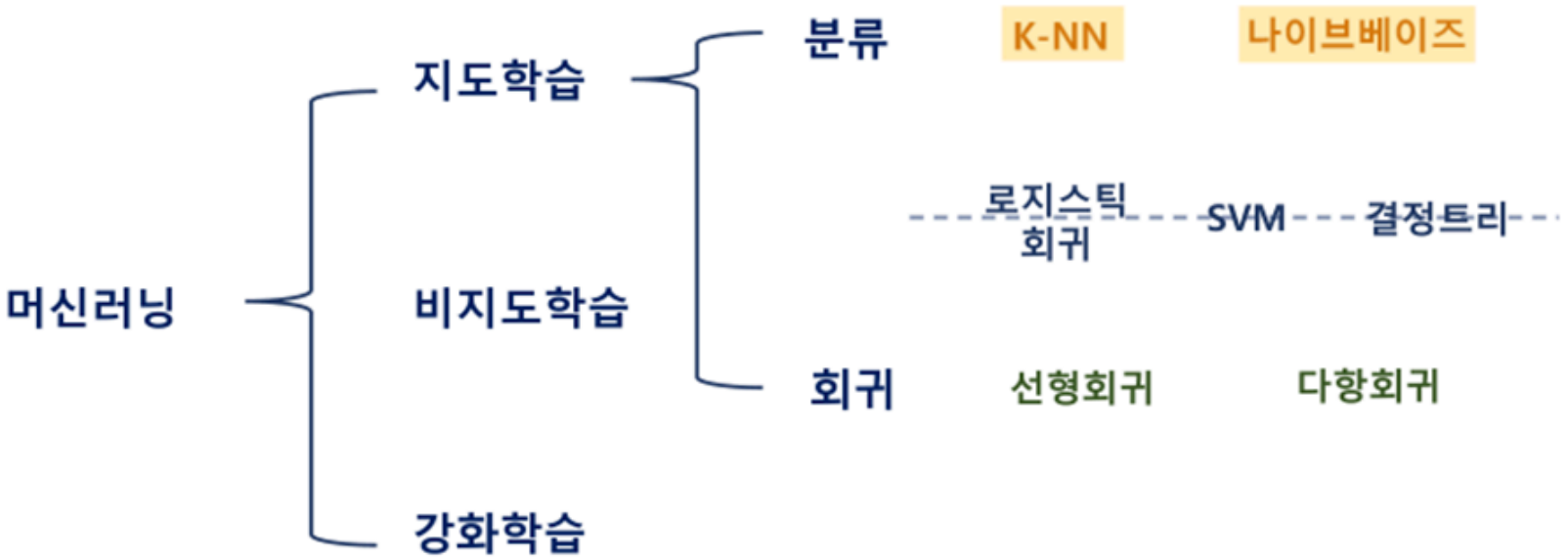
분류와 혼동행렬-EDA/범주형데이터/KNN/유클리디안거리/베이지정리/나이브베이지정리/혼동행렬/ROC

4강. 분류(Classification)와 혼동행렬

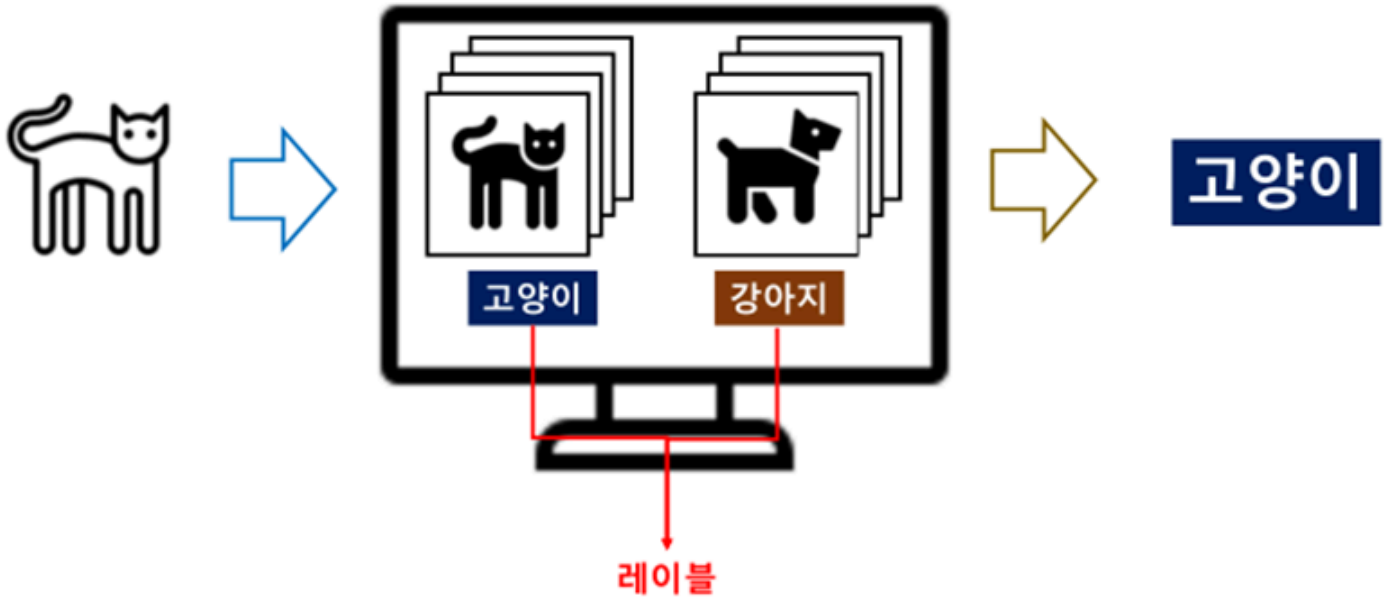
- Label Encoding과 One-Hot Encoding
- K-NN(K-Nearest Neighbor)
- 나이브베이즈 분류
- 혼동행렬(Confusion Matrix)

지도학습의 종류

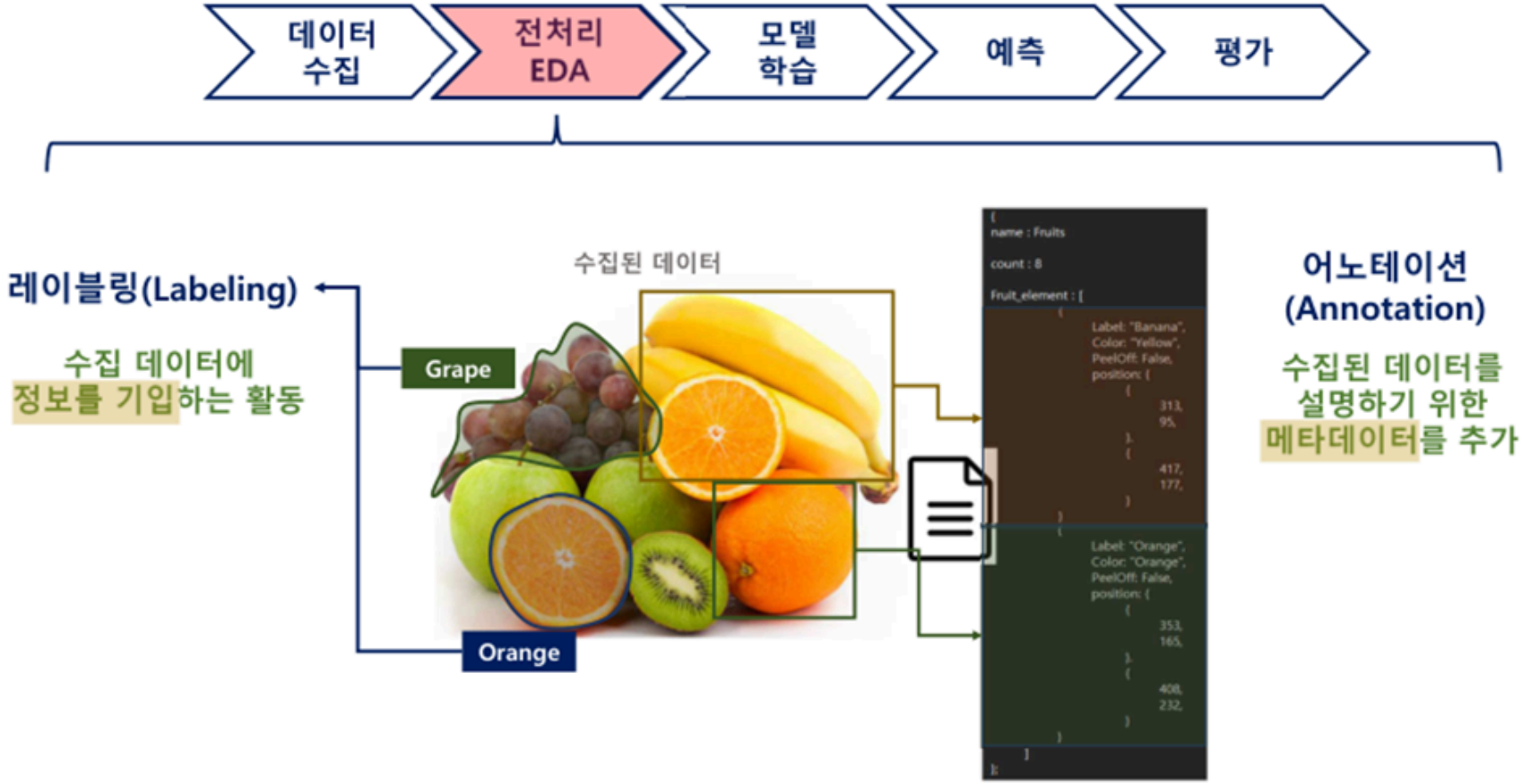

머신러닝
지도학습
비지도학습
강화학습
분류
K-NN
나이브베이즈
로지스틱 회귀
SVM
결정트리
회귀
선형회귀
다항회귀

분류(Classification)

데이터를 기반으로 패턴을 학습하여, 새로운 입력 데이터에 대한 레이블(Label)을 예측하는 모델



분류(Classification)



분류(Classification)

데이터 인코딩(Encoding)

데이터를 수치형으로 변환해주는 전처리 작업

Label Encoding

N개의 데이터를 0부터 N-1의 수치로 표현

숫자의 크기가 학습에 영향을 미침

오렌지	Label Encoding	0
바나나		1
포도		2

One-Hot Encoding

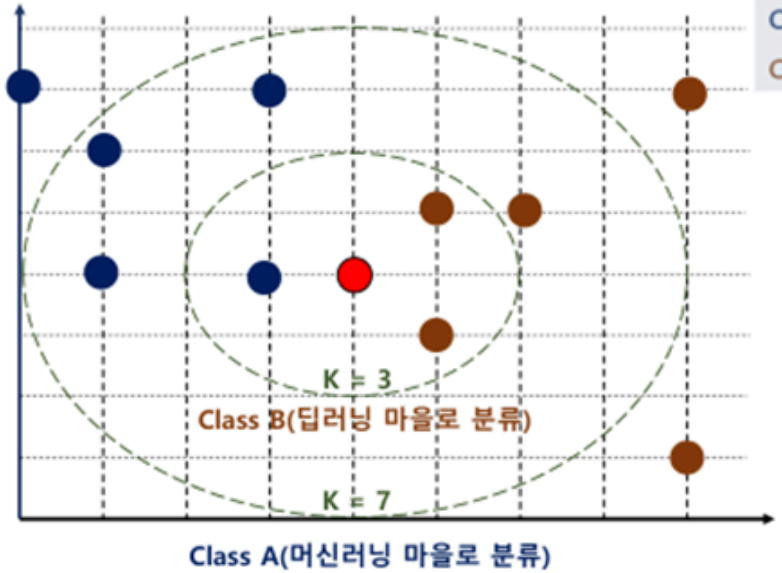
고유 값에 해당하는 컬럼에만 1을 표시하고, 나머지는 0을 표시

		오렌지	바나나	포도
오렌지	One-Hot Encoding	1	0	0
바나나		0	1	0
포도		0	0	1

K-Nearest Neighbor

데이터 수집 → 전처리 EDA → 모델 학습 → 예측 → 평가

데이터가 주어지면 거리 기반으로 이웃에 더 많은 데이터가 포함되어 있는 범주로 분류하는 방식

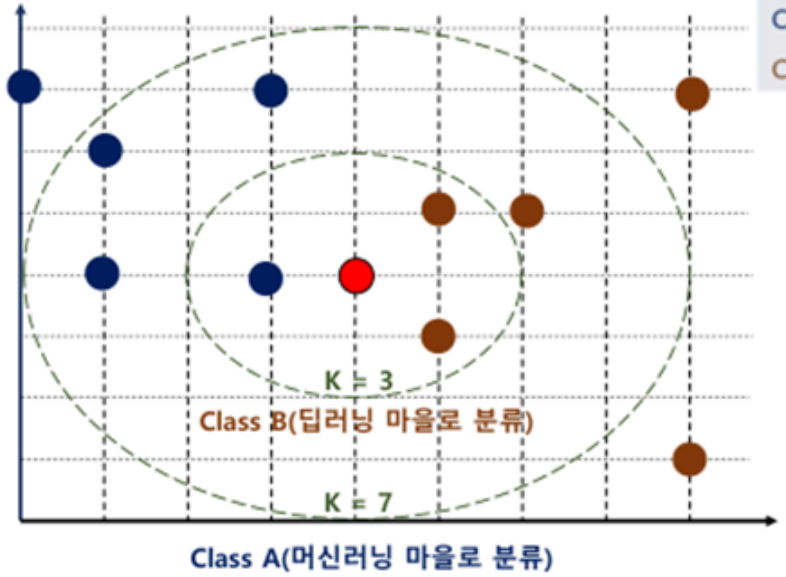


Q) 새롭게 생긴
빨간 집은 어느 마을로 분류?

K-Nearest Neighbor

데이터 수집 → 전처리 EDA → 모델 학습 → 예측 → 평가

데이터가 주어지면 거리 기반으로 이웃에 더 많은 데이터가 포함되어 있는 범주로 분류하는 방식

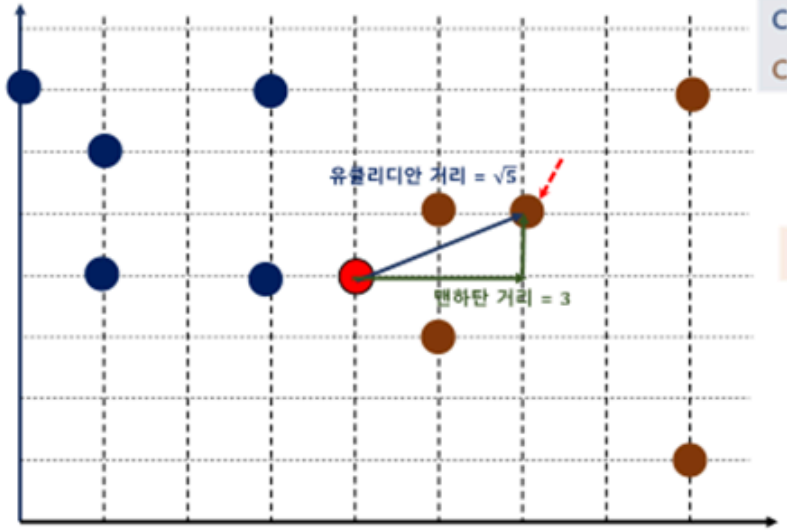


- 단순하고 효율적인 모델
- K의 값에 따라 결과가 달라짐
- 훈련이 따로 필요 없음 (Lazy Model)
- 관계를 이해하는데 제한적
- 데이터가 많아지면 분류가 느림

K-Nearest Neighbor

데이터 수집 → 전처리 EDA → 모델 학습 → 예측 → 평가

데이터가 주어지면 거리 기반으로 이웃에 더 많은 데이터가 포함되어 있는 범주로 분류하는 방식



Class A : 머신러닝 마을
Class B : 딥러닝 마을

거리 계산 방법

유클리디안 거리

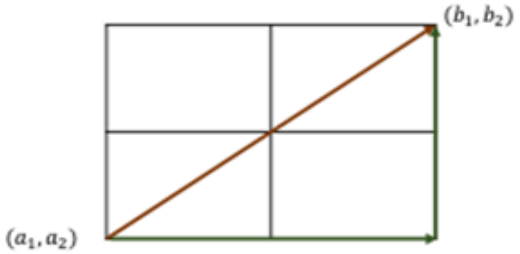
점과 점 사이의 직선거리

$$d = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2}$$

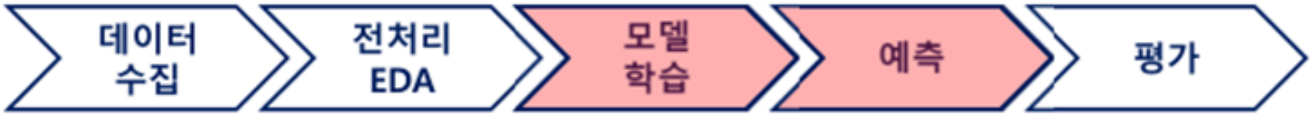
맨하탄 거리

X축 Y축을 따라간 거리

$$d = |a_1 - b_1| + |a_2 - b_2|$$



나이브 베이즈 분류



베이즈 정리란?

두 확률 변수의 사전 확률과 사후 확률 사이의 관계를 나타내는 정리

우도(likelihood) 사전확률

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

사후확률 주변 우도

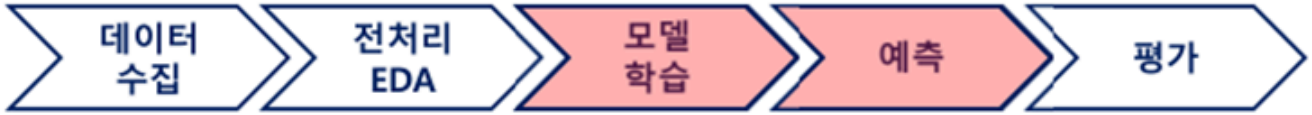
- $P(A \cap B)$: 사건 A와 B가 동시에 일어날 확률
- $P(A)$: 사건 A가 일어날 확률
- $P(B)$: 사건 B가 일어날 확률
- $P(A|B)$: 사건 B가 일어난 후 사건 A가 일어날 확률
- $P(B|A)$: 사건 A가 일어난 후 사건 B가 일어날 확률

사건 A와 B가 동시에 일어날 확률

$$P(A,B) = P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$$

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

나이브 베이즈 분류



베이즈 정리란?

두 확률 변수의 사전 확률과 사후 확률 사이의 관계를 나타내는 정리

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

A대학 입시에 응시한 남학생과 여학생의 비율이 60%와 40%이고
남학생의 합격률은 30%, 여학생의 합격률은 50%이다.
이때, A대학에 합격한 신입생 중 남학생을 고를 확률은?

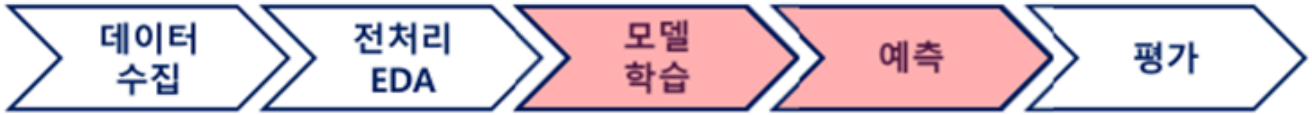
우도표	합격률	불합격률	
남학생	0.6*0.3 = 0.18 ($P(A \cap B)$)	0.6*0.7 = 0.42	0.6
여학생	0.4*0.5 = 0.20	0.4*0.5 = 0.20	0.4
	0.38 ($P(B)$)	0.62	

- $P(A \cap B)$: 사건 A와 B가 동시에 일어날 확률
  → 남학생이면서 합격한 신입생일 확률
- $P(A)$: 사건 A가 일어날 확률
  → 남학생일 확률
- $P(B)$: 사건 B가 일어날 확률
  → 신입생 합격률
- $P(A|B)$: 사건 B가 일어난 후 사건 A가 일어날 확률
  → 합격한 신입생 중 남학생일 확률
- $P(B|A)$: 사건 A가 일어난 후 사건 B가 일어날 확률
  → 남학생 중 합격생일 확률

$$P(A|B) = \frac{P(A \cap B) = 0.18}{P(B) = 0.38} = \frac{P(B|A)P(A) = (\frac{0.18}{0.6})(0.6)}{P(B) = 0.38}$$

$$= 0.47$$

나이브 베이즈 분류



나이브 베이즈 분류란?

데이터의 특징을 가지고 각 클래스(레이블)에 속할 확률을 계산하는 분류 기법

나이브

데이터 간의 특징이 상호 독립적이라는 가정 하에 확률을 계산

$$P(A,B) = P(A \cap B) = P(A|B)P(B) = P(A)P(B)$$

베이즈 이론

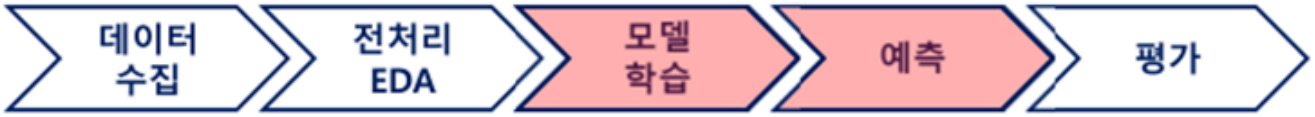
$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

데이터 특징이 여러 개

$$P(A|B,C,D) = \frac{P(B,C,D|A)P(A)}{P(B,C,D)}$$

$$= \frac{P(B|A)P(C|A)P(D|A)P(A)}{P(B)P(C)P(D)}$$

나이브 베이즈 분류



나이브 베이즈 분류란? 데이터의 특징을 가지고 각 클래스(레이블)에 속할 확률을 계산하는 분류 기법

$$P(A|B,C,D) = \frac{P(B|A)P(C|A)P(D|A)P(A)}{P(B)P(C)P(D)}$$

지금까지 받은 메일의 내용을 '광고', '전화', '대출' 을 Keyword로 하여 데이터를 수집하였을 때,
새로운 메일이 SPAM인지 아닌지를 판단하시오.

SPAM 메일이 광고라는 단어를 포함할 확률

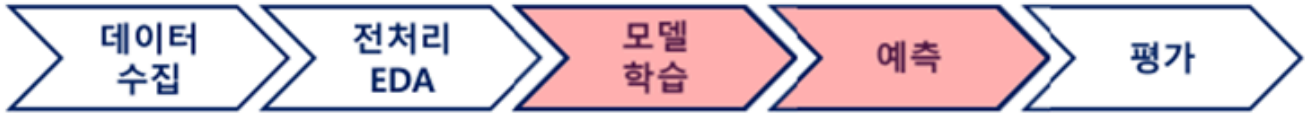
모든 메일 중 SPAM 메일의 비율

$$P(SPAM|\text{광고},\text{전화},\text{대출}) = \frac{P(\text{광고}|SPAM)\,P(\text{전화}|SPAM)\,P(\text{대출}|SPAM)\,P(SPAM)}{P(\text{광고})\,P(\text{전화})\,P(\text{대출})}$$

광고, 전화, 대출을 포함한 메일이 SPAM일 확률

모든 메일 중에 광고라는 단어를 포함한 메일의 비율

나이브 베이즈 분류



나이브 베이즈 분류란?

데이터의 특징을 가지고 각 클래스(레이블)에 속할 확률을 계산하는 분류 기법

$$P(A|B,C,D) = \frac{P(B|A)P(C|A)P(D|A)P(A)}{P(B)P(C)P(D)}$$

- 간단하고 빠르며 비교적 정확한 모델
- 큰 데이터 셋에서도 적합
- 특징 간에 독립성이 보장되어야 함
- 실생활에 적용하기 어려움

혼동행렬(Confusion Matrix)

분류 모델의 예측 결과를 평가하는 데 사용하는 표

실제

	SPAM	정상

실제

	TRUE	FALSE
TRUE	TRUE	FALSE

혼동행렬(Confusion Matrix)

분류 모델의 예측 결과를 평가하는 데 사용하는 표

지표	계산식
정밀도(Precision)	$\frac{TP}{TP+FP}$

실제

	TRUE	FALSE