2.1 이론

2.1.1 분류 문제란?

분류(Classification)는 입력 데이터 $\boldsymbol{x} \in \mathbb{R}^d$ 가 주어졌을 때, 이를 미리 정의된 클래스 집합 중 하나에 할당하는 문제입니다. 여기서 입력 $\boldsymbol{x}$ 는 d 차원의 실수 벡터이며, 출력 클래스 y 는 이산적인 값입니다.

$$f : \mathbb{R}^d \to \{1, 2, \ldots, K\}$$

분류 문제는 목적에 따라 다음과 같이 나뉩니다:

- **이진 분류**: 클래스가 2개 (예: 스팸/정상)
- **다중 클래스 분류**: 클래스가 3개 이상 중 하나 (예: 손글씨 숫자 인식)
- **다중 레이블 분류**: 여러 클래스가 동시에 정답일 수 있음 (예: 영화 장르)

2.1.2 대표 분류 알고리즘 비교

로지스틱 회귀 (Logistic Regression)

선형 결합 후 시그모이드 함수로 확률을 계산하는 모델:

$$P(y = 1 \mid \boldsymbol{x}) = \frac{1}{1 + e^{-(\boldsymbol{w}^T \boldsymbol{x} + b)}}$$

결정 경계는 선형: $\boldsymbol{w}^T \boldsymbol{x} + b = 0$

K-최근접 이웃 (KNN)

새로운 입력이 들어왔을 때 가장 가까운 K개의 이웃을 기준으로 다수결 분류. 학습이 따로 없고 직관적이지만, 고차원에서는 성능이 급격히 떨어질 수 있습니다.

결정 트리 (Decision Tree)

입력을 조건문으로 분기하며 분류를 수행. 해석이 쉬우나 트리 깊이가 깊어지면 과적합 위험이 있습니다.

서포트 벡터 머신 (SVM)

두 클래스를 **가장 넓은 마진**으로 나누는 결정 경계를 학습합니다:

$\boldsymbol{w}^T \boldsymbol{x} + b = 0$

커널 기법을 이용하면 비선형 경계도 가능하며 일반화 성능이 뛰어납니다. 단점은 학습 속도와 매개변수 조정의 어려움입니다.

2.1.3 결정 경계와 해석 가능성

각 모델은 입력 공간을 구분짓는 결정 경계(Decision Boundary)를 학습합니다. 이 경계는 예측 기준이 되며, 모델의 성격을 파악하는 핵심 요소입니다.

> 해석력: 모델이 어떻게 예측을 수행하는지, 어떤 변수가 어떤 경향을 미치는지, 사람이 이해할 수 있는 방식으로 설명 가능한 정도입니다.

모델	경계 형태	해석력	특이사항
로지스틱 회귀	선형	매우 높음	가중치로 변수 영향도 해석 가능
결정 트리	계단형	높음	조건 분기 구조 설명 가능
KNN	매우 유연	낮음	해석 어려움
SVM	선형/비선형	낮음	고차원 해석 어려움

2.1.4 고려사항

특성 스케일링

거리 기반 모델(KNN, SVM)이나 내적(dot product) 기반 모델(로지스틱 회귀)은 스케일 차이에 민감합니다. 따라서 표준화를 통해 다음과 같이 조정합니다.

$$x'_j = \frac{x_j - \mu_j}{\sigma_j}$$

기호	의미	설명
x_j	원래 특성값	예: 키 = 172cm
μ_j	변수 j의 평균	해당 열 전체의 평균
σ_j	변수 j의 표준편차	분산의 제곱근
x'_j	표준화된 값	평균 0, 표준편차 1로 변환된 값

클래스 불균형 대응

정확도(Accuracy)만으로는 불균형 문제를 파악할 수 없습니다.(클래스 불균형으로 왜곡이 될 수 있습니다.)

다음과 같은 대응이 필요합니다.

- 데이터: 오버샘플링(소수 클래스의 데이터를 인위적으로 늘리기), 언더샘플링(다수 클래스 일부 제거)
- 모델: `class_weight='balanced'`
- 평가: 정밀도(Precision), 재현율(Recall), F1 점수 활용

예측 성능 vs 해석 가능성

모델	성능	해석력	적합한 사용 예
로지스틱 회귀	보통	매우 높음	영향도 분석
결정 트리	보통	높음	조건 설명 필요할 때
SVM	높음	낮음	복잡한 결정 경계
KNN	낮음	없음	단순한 실험용
신경망	매우 높음	매우 낮음	설명 불필요, 고성능 예측 필요