2주차 스터디 정리

이하늘

지도학습

- 입력 데이터와 정답(레이블)을 함께 제공해 모델을 학습시키는 방식.
- 모델은 입력과 정답 간의 관계를 학습해, 새로운 입력에 대해 적절한 출력을 예측.

회귀

- **연속적인 수치형 데이터를 예측**하는 지도학습 방식입니다. 입력 변수들과 출력 변수 간의 관계를 모델링하여, 주어진 입력에 따라 **연속적인 값**(예: 집 값, 온도, 매출 등)을 예측
- **선형 회귀**: 가장 기본적인 회귀 방법으로, 입력 변수와 출력 변수 사이의 선형 관계를 가정.
- **다항 회귀**: 입력 변수와 출력 변수 사이의 비선형 관계를 모델링.
- **다중 회귀**: **여러 개의 독립 변수**(입력)와 **하나의 종속 변수**(출력) 사이의 관계를 모델링하는 회귀 방법

- 선형 회귀(Linear Regression) $y = w_1x + b$
- 다항 회귀(Polynomial Regression) $y = w_1x + w_2x^2 + w_3x^3 + ... + w_nx^n + b$
- 다중 회귀(Multiple Regression) $y = w_1x_1 + w_2x_2 + ... + w_nx_n + b$

규제

- 모델의 과적합을 방지하는 기법입니다.
- **L1 규제(Lasso)**: 모델의 가중치 합의 절대값에 패널티를 부여하여, 불필요한 가중치(특성)를 0으로 만들어 간소화.

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}\left(\hat{y}^{(i)} - y^{(i)}\right)^2 + \lambda\sum_{j=1}^{n}|w_j|$$

- **L2 규제(Ridge)**: 가중치 합의 제곱에 패널티를 부여하여, 너무 큰 가중치를 줄이고 모델이 지나치게 데이터를 따라가지 않도록 함.

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}\left(\hat{y}^{(i)} - y^{(i)}\right)^2 + \lambda\sum_{j=1}^{n}w_j^2$$

분류

- 주어진 데이터를 미리 정의된 클래스(카테고리)로 분류하는 지도학습 방식입니다. 이는 이산적인 값(예: 스팸/정상, 개/고양이)을 예측
- **로지스틱 회귀**: 이진 분류 문제에서 주로 사용되는 기법으로, 시그모이드 함수를 이용해 분류합니다.
- **SVM (서포트 벡터 머신)**: 데이터를 고차원 공간에서 분리할 수 있는 초평면을 찾아 분류
- 소프트맥스(Softmax): 다중 클래스 분류에서 각 클래스의 확률을 계산해 총합이 1이 되도록 출력하는 함수

결정트리

- 데이터를 **트리 구조로** 분할하여 예측
- 각 노드는 하나의 특징을 기준으로 데이터를 나눔
- 리프 노드는 최종 분류나 예측 값
- 직관적이고 해석이 쉬워 다양한 분류 및 회귀 문제에 사용

결정 트리 회귀 분류

- **트리 구조**:
  -데이터를 특징(feature)을 기준으로 여러 가지 질문(조건)에 따라 나누어 트리 구조를 형
  -각 내부 노드는 하나의 특징을 기준으로 데이터를 분할하고, 잎 노드는 최종 클래스 레이블을 나타냄
- **특징 선택**: 각 분할 단계에서 데이터의 불순도(impurity)를 최소화하는 특징을 선택

$$Gini(D) = 1 - \sum_{k=1}^{K} p_k^2$$

결정트리 회귀

- **트리구조:**
  - 각 내부 노드는 입력 변수의 특정 값을 기준으로 데이터를 분할
  - 잎 노드는 해당 구간의 평균 값 또는 중앙값을 예측 값으로 사용
- **분할 기준**: 각 분할 단계에서 데이터의 분산을 최소화하는 기준을 사용

결정 트리

- 트리 높이가 클 때 괴대적합 가능성 多
- 트리 높이가 작을 때 과소적합 가능성 多

- 과소/과대 적합 완화 방법
  - 높이를 제한
  - 앙상블기법: 여러 개의 모델을 결합하여 더 강력하고 일반화된 예측 성능을 얻기 위한 기법
  - Cross validation(교차검증): 데이터 세트를 여러 개의 서브 세트로 나누어 모델을 훈련하고 검증하는 과정