2025학년도 1학기

1주차 리포트

과목명	인공지능개론
제 목	1주차 리포트
담 당	이명규 교수님
제출일	2025년 3월 11일

학 과	컴퓨터공학과
학 번	202235268
성 명	송이두

[Q1] 인공지능에서 지능에 해당하는 기능은 무엇인가?
[A1] 인공지능은 인간의 지능을 모방하여 데이터를 통해 특성을 파악하는 기술이다. 인공지능의 '지능'은 인간의 인지 능력을 모방하는 다양한 기능을 의미한다. 학습 기능, 추론 기능, 문제 해결 기능, 인지 기능이 이에 해당한다. 학습 기능은 데이터에서 패턴을 학습하고 예측 또는 결정을 내리는 기능을 뜻한다. 추론 기능은 주어진 정보를 바탕으로 논리적인 결론을 도출하는 기능이다. 문제 해결 기능은 복잡한 문제를 해결하기 위해 최적의 해법을 찾는 기능이다. 인지 기능은 인간의 언어나 이미지, 영상을 분석하고 인식, 생성하는 기능이다.

[Q2] 인공지능의 종류 3가지에 대해 설명하시오. (지도 학습, 반지도 학습, 강화 학습)
[A2] 인공지능의 종류에는 지도 학습, 반지도 학습, 강화 학습이 있다. 지도 학습은 정답 레이블이 있는 데이터를 사용하여 모델을 학습시키는 방법이다. 명확한 정답이 주어진 상태에서 학습하므로 예측 정확도가 높다. 예시로는 스팸 메일 필터링과 이미지 분류가 있다. 반지도 학습은 지도 학습과 비지도 학습의 중간 형태로, 일부 데이터에는 정답 레이블이 있고, 나머지 데이터에는 없는 상태로 학습을 진행하는 방식이다. 반지도 학습을 통해 데이터 수집 비용을 절감할 수 있고, 비지도 학습보다 높은 성능을 기록할 수 있다. 반지도 학습은 음성 인식, 웹 페이지 분류 등에서 사용된다. 강화 학습은 인공지능이 환경과의 상호 작용을 통해 보상을 최대화하는 방향으로 학습하는 방법이다. 시행착오를 통해 최적의 행동 방식을 학습하며, 예시로는 자율 주행 자동차, 로봇 청소기가 있다.

[Q3] 전통적인 프로그래밍 방법과 인공지능 프로그램의 차이점은 무엇인가?
[A3] 전통적인 프로그래밍은 개발자가 직접 규칙을 정해야 한다. 반면 인공지능 프로그램은 인공지능이 데이터를 기반으로 패턴을 학습하여 스스로 규칙을 찾아낸다.

[Q4] 딥러닝과 머신러닝의 차이점은 무엇인가?
[A4] 머신러닝은 딥러닝보다 가볍고 다양한 알고리즘을 사용할 수 있지만, 학습에 중요한 특징은 사람이 직접 선택해야 한다. 반면 딥러닝은 머신러닝보다 더 많은 데이터와 연산 능력이 필요하지만, 다층 신경망을 활용하여 인공지능 모델이 데이터의 특징을 스스로 학습한다.

[Q5] Classification과 Regression의 주된 차이점은 무엇인가?
[Q5] Classification은 주어진 데이터를 특정 카테고리로 분류하며 0과 1, 개와 고양이 같은 이산적인 값을 출력한다. Regression은 연속적인 숫자 값을 예측하며 가격, 온도와 같은 연속적인 값을 출력한다.

[Q6] 머신러닝에서 차원의 저주(curse of dimensionality)란 무엇인가?
[A6] 차원의 저주란 머신러닝에서 특성 차원이 증가할수록 데이터 분석과 학습이 어려워지는 현상을 의미한다. 특성 차원이 증가할수록 각 데이터 포인트 간 거리가 멀어지며, 밀집된 학습이 어려워진다. 차원의 저주는 계산 복잡도를 증가시키고 과적합(오버피팅) 위험성을 증가시킨다. 차원의 저주를 해결하기 위해서는 차원 축소 기법을 활용해야 한다.

[Q7] Dimensionality Reduction은 왜 필요한가?
[A7] Dimensionality Reduction은 차원 축소라고도 하며 고차원 데이터를 저차원으로 변환하여 학습 성능을 향상시키는 기법이다. 차원 축소는 차원의 저주를 해결하며, 데이터 시각화를 용이하게 하며, 해석 가능성을 향상시킬 수 있다. 따라서 차원 축소는 머신러닝 모델 성능을 최적화하기 위해 필요하다.

[Q8] Ridge와 Lasso의 공통점과 차이점은 무엇인가? (Regularization, 규제 , Scaling)
[A8] Ridge와 Lasso는 모두 선형 회귀의 정규화 기법으로 과적합(오버피팅)을 방지하고 모델의 일반화 성능을 향상시키는 역할을 한다. 두 기법 모두 정규화 기법을 사용하고, 손실 함수에 패널티를 추가하며 데이터 스케일링이 필요하다는 점에서는 공통점이 있다. 하지만 Ridge는 L2 정규화를 사용하며 계수들의 제곱을 패널티로 추가한다. Lasso는 L1 정규화를 사용하며 계수들의 절대값을 패널티로 추가한다. Ridge는 모든 특성을 유지하지만 Lasso는 일부 특성을 제거한다.
[Q9] Overfitting vs. Underfitting
[A9] 과적합(Overfitting)은 훈련 데이터에 너무 잘 맞춰져서 새로운 데이터에 대한 성능이 낮

은 것을 의미한다. 모델이 너무 복잡하거나 훈련 데이터에 과하게 최적화되었을 때 발생한다. 해결하기 위해서는 모델을 단순화하거나 Ridge나 Lasso 같은 정규화 기법을 이용하며 데이터의 양을 늘리는 방법을 사용한다. 과소적합(Underfitting)은 모델이 너무 단순해서 훈련 데이터조차 제대로 학습하지 못하는 것을 의미한다. 모델이 너무 단순하거나 데이터나 연산이 부족할 때 발생한다. 해결하기 위해서는 더 복잡한 모델을 사용하고, 더 많은 데이터와 특성을 추가하는 방법이 있다.

[Q10] Feature Engineering과 Feature Selection의 차이점은 무엇인가?
[Q10] Feature Engineering은 기존 데이터를 가공하여 새로운 특징(feature)을 생성하는 것이고, Feature Selection은 기존 특징 중 중요한 것만 선택하는 것이다. Engineering은 새로운 변수가 추가되며 Feature Selection은 기존 특징의 수가 감소한다.

[Q11] 전처리(Preprocessing)의 목적과 방법은 무엇인가? (노이즈, 이상치, 결측치)
[A11] 전처리란 머신러닝 모델을 훈련하기 전에 데이터를 정리하고 변환하여 품질을 향상시키는 과정이다. 전처리는 데이터 품질을 향상하여 모델의 성능을 개선하고, 정확한 예측을 보장하기 위해 실시한다. 전처리 방법에는 노이즈 제거, 이상치 처리, 결측치 처리가 있다. 노이즈 제거는 데이터에 포함된 불필요한 정보를 제거하여 정확한 학습을 유도하는 것이다. 평균 필터와 중앙값 필터를 통해 데이터의 이상 변동을 줄이며 스무딩을 통해 이동 평균을 적용한다. 이상치 처리는 극단적인 값이 모델 학습을 방해하는 것을 방지한다. 통계적인 방법(사분위 범위, 표준편차)과 모델 기반(Isolation Forest, LOF)으로 이상치를 탐지한 후 이상치를 수정한다. 결측치 처리는 누락된 데이터를 보완하여 분석의 신뢰성을 높이는 것이다. 결측치가 적을 경우 해당 행 또는 열을 삭제하거나 평균, 중앙, 최빈값으로 결측값을 채운다. 또는 모델을 사용하여 결측값을 예측하거나 결측 여부를 학습할 수 있도록 별도 처리한다.

[Q12] EDA(Exploratory Data Analysis)란 무엇인가? 데이터의 특성 파악(분포, 상관관계)
[A12] EDA란 머신러닝 모델을 학습하기 전에 데이터를 이해하고 특성을 분석하는 과정이다. 데이터의 전반적인 특성을 파악해 데이터를 이해하고, 변수 간 관계 분석을 위해 실시한다. 데이터의 특성을 파악하기 위해서는 각 변수의 값들이 어떻게 분포하는지 분석해야 한다. 히스토그램으로 연속형 변수의 분포를 확인하고, 박스 플롯으로 이상치를 탐지한다. 커널 밀도 추정을 통해 데이터의 확률 밀도를 시각화한다. 또한 독립 변수와 종속 변수 간 관계를 분석해야 한다. 상관 행렬을 통해 변수 간 상관관계를 시각적으로 확인하고, 산점도를 통해 두 변수 간 관계를 확인한다.

[Q13] 회귀에서 절편과 기울기가 의미하는 바는 무엇인가? 딥러닝과 어떻게 연관되는가?
[A13] 회귀에서 절편은 x가 0일 때, y의 값을 의미한다. 기울기는 x가 1 증가할 때, y가 얼마나 변하는지 나타낸다. 딥러닝에서 절편은 편향을 의미한다. 회귀에서 x에 해당하는 입력값이 0일 때에도 뉴런이 활성화될 수 있도록 한다. 또한 기울기는 가중치를 의미한다. 입력이 모델의 예측값에 얼마나 영향을 미치는지를 결정한다.

[Q14] 결정 트리에서 불순도(Impurity) - 지니 계수(Gini Index)란 무엇인가?
[A14] 불순도란 한 노드에 다양한 클래스의 샘플이 섞여 있는 정도를 나타낸다. 불순도가 높을수록 혼합된 데이터가 많아 분류가 어렵고 불순도가 낮을수록 순수한 노드가 된다. 지니 계수란 결정 트리에서 불순도를 측정하는 지표 중 하나이다. 한 노드에 여러 클래스가 섞여 있을 확률을 나타내며, 값이 작을수록 순수한 노드이다. 지니 계수는 0에서 1사이의 범위를 나타낸다. 지니 계수가 0이면 모든 샘플이 동일한 클래스에 속해 있다는 의미이며 1이면 모든 샘플이 서로 다른 클래스에 속해있다는 의미이다. 결정 트리 알고리즘은 지니 계수를 사용하여 가장 불순도가 낮은 방향으로 데이터를 분할한다. 즉, 지니 계수가 가장 작은 특성을 기준으로 데이터를 분할하여 정보 이득을 최대화한다.

[Q15] 앙상블이란 무엇인가?
[A15] 앙상블은 여러 개의 머신러닝 모델을 조합하여 더 높은 성능을 얻는 기법이다. 하나의 모델이 데이터의 모든 패턴을 완벽하게 학습하기 어렵고, 과적합을 방지하기 위해 앙상블 학습이 필요하다. 앙상블 학습의 대표 알고리즘으로는 랜덤 포레스트가 있다.

[Q16] 부트 스트랩핑(bootstrapping)이란 무엇인가?
[A16] 부트 스트랩핑은 데이터에서 여러 개의 샘플을 랜덤하게 뽑아 새로운 데이터셋을 만드는 재표본 추출 기법이다. 부트 스트랩핑은 통계적 신뢰도를 향상시키며 모델의 성능을 안정화하고, 과적합을 방지한다. 랜덤 포레스트 같은 앙상블 모델에서 각 결정 트리를 훈련할 때 부트 스트랩핑 샘플을 활용한다.

[Q17] 배깅(Bagging)이란 무엇인가?
[A17] 배깅은 데이터의 분산을 줄여 모델의 안정성을 높이는 앙상블 학습 기법이다. 배깅은 여러 개의 모델을 평균화하여 분산을 줄이고 과적합을 방지한다. 불안정한 모델을 사용하더라도 배깅을 통해 안정적인 예측을 얻을 수 있으며 병렬 처리를 통해 학습 속도를 높일 수 있다. 배깅은 고차원 데이터나 노이즈가 많은 데이터, 불안정한 모델에서 효과적이다.

[Q18] 주성분 분석(PCA)이란 무엇인가?
[A18] 주성분 분석은 많은 변수들로 이루어진 데이터를 핵심적인 몇 개의 변수로 요약하는 차원 축소 기법이다. 주성분 분석을 통해 데이터의 복잡성을 줄여 분석 및 시각화를 용이하게 할 수 있으며 과적합을 방지할 수 있다. 또한 핵심 정보를 추출하여 해석력을 높일 수 있다. 다만, 원본 변수 간의 관계를 해석하기 어려울 수 있으며 비선형적인 데이터에는 적합하지 않을 수 있다.