TOPCIT 학습법 특강: 데이터 이해와 활용 영역

Sep 23, 2024

Myoungsung You (famous77@kaist.ac.kr)

Network and System Security (NS^2) Lab

School of Electrical Engineering at KAIST

NS2 Network and System Security Laboratory KAIST

Index

2. 데이터 이해와 활용 영역 학습 방법

2. 데이터 이해와 활용 영역 학습 방법

데이터 이해와 활용 영역 특징

- TOPCIT 시험에서 두 번째로 높은 점수 배점을 가진 영역
- 문항수: 9
- 배점: 145점
- 주관식, 단답형, 서술형 (SQL문 및 ERD 작성)

2. 데이터 이해와 활용 영역 학습 방법

데이터 이해와 활용 영역 세부 영역

- 데이터와 데이터베이스의 이해
- 데이터베이스 종류의 이해
- 데이터베이스 설계 및 구축 절차
- 데이터 모델링
- 정규화와 반정규화
- 데이터베이스 물리 설계
- 데이터베이스 품질과 표준화
- 관계연산
- 관계 데이터베이스 언어
- …

2. 데이터 이해와 활용 영역 학습 방법

효율적인 학습 방법

- 개념에 대한 세부적인 이해
  - 관계형 데이터베이스 설계 / 물리 설계 (SQL)
  - 정규화
  - 인공지능 및 빅데이터
- 단순 암기
  - 데이터 품질관리
  - 관계연산
  - 동시성 제어
  - 데이터베이스 복구 / 분석 이해

3. 데이터 이해와 활용 영역 주요 개념

3. 데이터 이해와 활용 영역 주요 개념

데이터베이스(시스템)

- 데이터를 데이터베이스에 저장하고 관리해서 필요한 정보를 생성하는 시스템
  - 데이터베이스 언어: 사용자와 데이터베이스간의 인터페이스 (SQL 등)
  - 사용자: 데이터베이스 관리자, 데이터베이스 사용자, 응용 프로그래머
  - **데이터베이스 관리 시스템 (DBMS)**: 데이터베이스를 구축하고 이용하는 기능을 제공하는 시스템

3. 데이터 이해와 활용 영역 주요 개념

데이터베이스의 등장 배경

- **데이터 종속성**과 **데이터 중복성**을 제거
- **데이터 종속성** (data dependency)
  - 응용프로그램과 데이터간의 의존관계, 데이터의 구성이나 접근방식이 변경되면 응용프로그램도 이에 맞게 수정되어야 함

- **데이터 중복성** (data redundancy)
  - 한 시스템에 같은 내용의 데이터가 여러 파일에 중복되어 저장 및 관리 되는 것

3. 데이터 이해와 활용 영역 주요 개념

키 (Key)

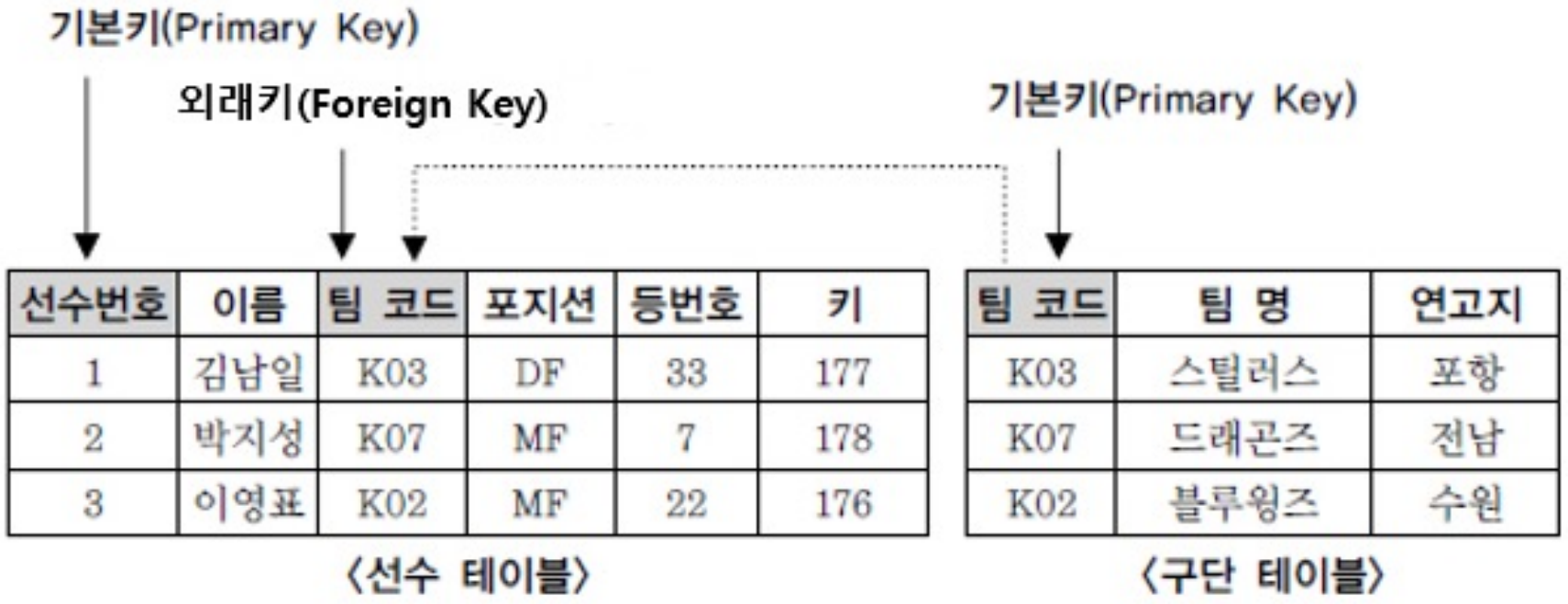
- 테이블에서 각 레코드를 식별하기 위해 사용되는 속성 또는 속성의 집합
  - **유일성**: 키가 레코드를 유일하게 구분함
  - **최소성**: 키를 구성하는 속성 중 어떤 하나라도 제거하면 유일성이 깨짐
- 후보키: 유일성과 최소성을 갖는 모든 속성들
- **기본키**: 후보키 중 레코드 식별을 위해 선정된 키
- 대체키: 후보키 중 기본키를 제외한 속성들
- 외래키: 하나의 테이블에서 다른 테이블을 참조하기 위해 사용되는 속성

학과	학번	학년	이름
컴퓨터공학	20221234	2	송민규
소프트웨어	20215845	3	김재한
전기및전자	20229513	2	유명성
컴퓨터공학	20203818	4	서민재

기본키 **대체키**

3. 데이터 이해와 활용 영역 주요 개념

- 외래키 예시



선수번호	이름	팀 코드	포지션	등번호	키
1	김남일	K03	DF	33	177
2	박지성	K07	MF	7	178
3	이영표	K02	MF	22	176

〈선수 테이블〉

팀 코드	팀 명	연고지
K03	스틸러스	포항
K07	드래곤즈	전남
K02	블루윙즈	수원

〈구단 테이블〉

3. 데이터 이해와 활용 영역 주요 개념

참조 무결성

- 외래키 값은 Null 값이거나 참조되는 테이블의 기본키와 같아야함
- 기본키가 참조되는 외래키가 존재할 경우 해당 데이터는 삭제 또는 변경될 수 없음

3. 데이터 이해와 활용 영역 주요 개념

관계 데이터베이스 언어

- SQL (Structured Query Language): RDB에 저장된 데이터를 처리하기 위한 표준 언어
- 비절차적인 언어로 데이터 처리 과정을 명시하는 것이 아니라, 단지 얻고자 하는 연산 결과만 명시

명칭	설명	예제
DDL (data definition language)	데이터 구성요소 정의 및 변경 삭제 처리 언어	CREATE, ALTER, DROP
DML (data manipulation language)	데이터 조작 언어	SELECT, INSERT, UPDATE, DELETE
DCL (data control language)	사용자별 데이터 접근 또는 권한 제어 언어	GRANT, REVOKE, COMMIT, ROLLBACK

3. 데이터 이해와 활용 영역 주요 개념

데이터 정의 언어 (DDL)

- 데이터베이스나 테이블 생성, 변경, 삭제에 사용되는 언어

명령어	설명
CREATE	데이터베이스나 테이블 생성
ALTER	데이터베이스나 테이블 수정
DROP	데이터베이스나 테이블 삭제
TRUNCATE	테이블 내의 모든 레코드 삭제, ROLLBACK 불가

3. 데이터 이해와 활용 영역 주요 개념

CREATE

- CREATE Table, CREATE Sequences, CREATE Index, CREATE View 등과 같이 객체를 생성하는 데 사용
- CONSTRANT 설정 가능
  - PRIMARY KEY, FOREIGN KEY, UNIQUE KEY, NOT NULL, CKECK

```
CREATE TABLE table_name (
    column1 datatype1 [constraint1],
    column2 datatype2 [constraint2],
    ...
    [table_constraint]
);
```

```
1  CREATE TABLE employees (
2      employee_id INT PRIMARY KEY,
3      employee_name VARCHAR(50),
4      department_id INT,
5      CONSTRAINT fk_department FOREIGN KEY (department_id)
6      REFERENCES departments(department_id)
7  );
```

3. 데이터 이해와 활용 영역 주요 개념

VIEW

- DB내 하나 이상의 테이블로부터 유도된 가상 테이블
- 기존 테이블에서 데이터를 검색할 때 조건에 따라 필요한 열만 선택하거나 조인한 결과를 가져올 때 유용함

```
CREATE VIEW view_name AS
SELECT column1, column2, ...
FROM table_name
WHERE condition;
```

```
1   CREATE VIEW vw_employee_department AS
2   SELECT e.employee_id, e.employee_name, d.department_name
3   FROM employees e
```

3. 데이터 이해와 활용 영역 주요 개념

데이터 조작 언어 (DML)

- 데이터의 검색, 등록, 삭제, 수정을 위한 언어

명령어	설명
SELECT	테이블에서 특정 데이터를 조회
INSERT	테이블에 새로운 데이터를 삽입
UPDATE	테이블에서 특정 데이터를 갱신
DELETE	테이블에서 특정 데이터를 삭제

3. 데이터 이해와 활용 영역 주요 개념

SELECT

- 하나 또는 그 이상의 테이블에서 원하는 데이터를 추출하는 조작 언어
- 조회할 데이터에 대해 제약 조건 설정 가능

```sql
SELECT * FROM employees;
SELECT employee_id, employee_name FROM employees;
SELECT * FROM employees WHERE department_id = 1;
SELECT * FROM employees ORDER BY employee_name ASC;
```

3. 데이터 이해와 활용 영역 주요 개념

JOIN

- RDB에서 두 개 이상의 테이블 간에 연결을 만드는데 사용되는 구문
- JOIN을 사용해 관련된 테이블의 데이터를 결합해 단일 결과를 생성 가능
  - INNER JOIN, OUTER JOIN, FULL JOIN, CROSS JOIN

```
1    SELECT e.employee_id, e.employee_name, d.department_name
2    FROM employees e
3    JOIN departments d ON e.department_id = d.department_id;
```

3. 데이터 이해와 활용 영역 주요 개념

INSERT

- 테이블에 새로운 데이터를 추가, 컬럼의 순서와 값의 순서가 같아야 함
- 모든 컬럼의 값을 입력할 경우 컬럼 명 생략 가능

```sql
INSERT INTO employees (employee_id, employee_name, department_id, salary)
VALUES (1, 'John Doe', 101, 50000);

INSERT INTO employees
VALUES (1, 'John Doe', 101, 50000);

INSERT INTO employees (employee_name, department_id)
VALUES ('Jane Smith', 102);
```

3. 데이터 이해와 활용 영역 주요 개념

INSERT

- SELECT 문을 사용해 새로 데이터를 입력하지 않고 INSERT 가능

```
INSERT INTO new_table (column1, column2, ...)
SELECT column1, column2, ...
FROM existing_table
WHERE condition;
```

```
1    INSERT INTO high_salary_employees (employee_id, employee_name, salary)
2    SELECT employee_id, employee_name, salary
3    FROM employees
4    WHERE salary >= 50000;
```

3. 데이터 이해와 활용 영역 주요 개념

UPDATE

- 테이블에서 하나 이상의 레코드의 내용을 갱신, WHERE 절로 특정 조건에 맞는 레코드만 업데이트 가능

```
UPDATE table_name
SET column1 = value1, column2 = value2, ...
WHERE condition;
```

```
1   UPDATE employees
2   SET employee_name = 'John'
3   WHERE employee_id = 1;
```

3. 데이터 이해와 활용 영역 주요 개념

DELETE

- 테이블에서 한 개 이상의 행을 삭제, WHERE 절을 통해 제약조건 명시가능, 명시하지 않을 경우 모든 행을 삭제

```
DELETE FROM table_name
WHERE condition;
```

```
1   DELETE FROM employees;
2
3   DELETE FROM employees
4   WHERE employee_id = 1;
```

3. 데이터 이해와 활용 영역 주요 개념

데이터 제어 언어 (DCL)

- 데이터의 무결성, 보안, 회복, 동시성 관리를 위해 사용자별 DB 접근 또는 사용 권한을 부여, 제거하는 언어

명령어	설명
GRANT	사용자에게 특정 작업에 대한 권한 부여
REVOKE	사용자에게 특정 작업에 대한 권한 삭제
COMMIT	트랜잭션의 작업을 DB에 반영
ROLLBACK	트랜잭션의 작업이 비정상 종료되었을 때 이전 상태로 복구

3. 데이터 이해와 활용 영역 주요 개념

COMMIT and ROLLBACK

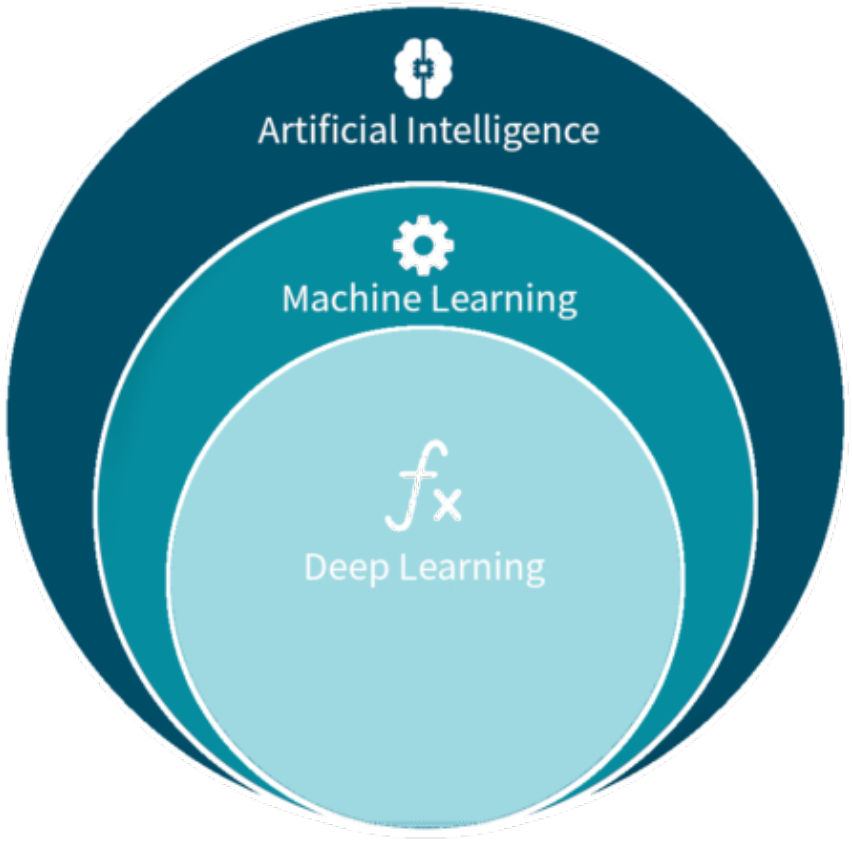
- 트랜잭션 제어에 사용되는 구문으로 트랜잭션의 결과를 DB에 반영 (commit) 할지 폐기 (rollback)할지 결정
  - 트랜잭션: DB에서 처리되는 하나의 논리적인 작업 단위

```
1   BEGIN TRANSACTION;
2
3   UPDATE employees
4   SET salary = salary * 1.1
5   WHERE department_id = 'HR';
6
7   COMMIT;
```

```
1   BEGIN TRANSACTION;
2
3   DELETE FROM employees
4   WHERE department_id = 'HR';
5
6   ROLLBACK;
```

3. 데이터 이해와 활용 영역 주요 개념

- 인공지능 / 머신러닝 / 딥러닝



ARTIFICIAL INTELLIGENCE
A technique which enables machines to mimic human behaviour

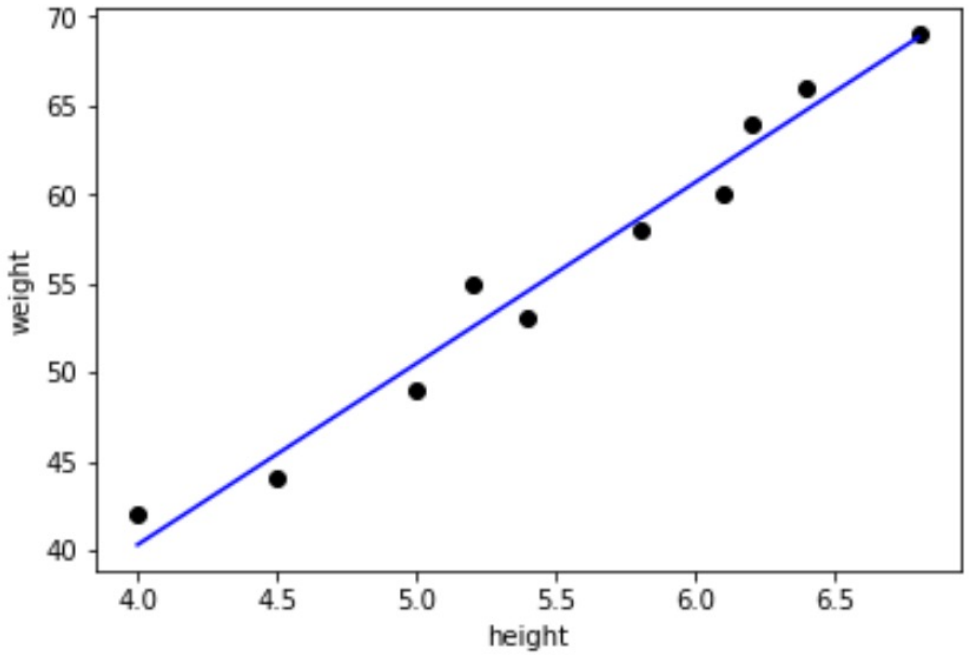
MACHINE LEARNING
Subset of AI technique which use statistical methods to enable machines to improve with experience

DEEP LEARNING
Subset of ML which make the computation of multi-layer neural network feasible

3. 데이터 이해와 활용 영역 주요 개념

머신러닝

- 통계적인 기법을 통해 데이터로부터 특정 작업을 수행할 수 있는 알고리즘을 기계가 스스로 찾아내게 하는 기술



[height에 대한 weight를 예측하는 그래프 찾기]

3. 데이터 이해와 활용 영역 주요 개념

머신러닝 모델의 학습 방법

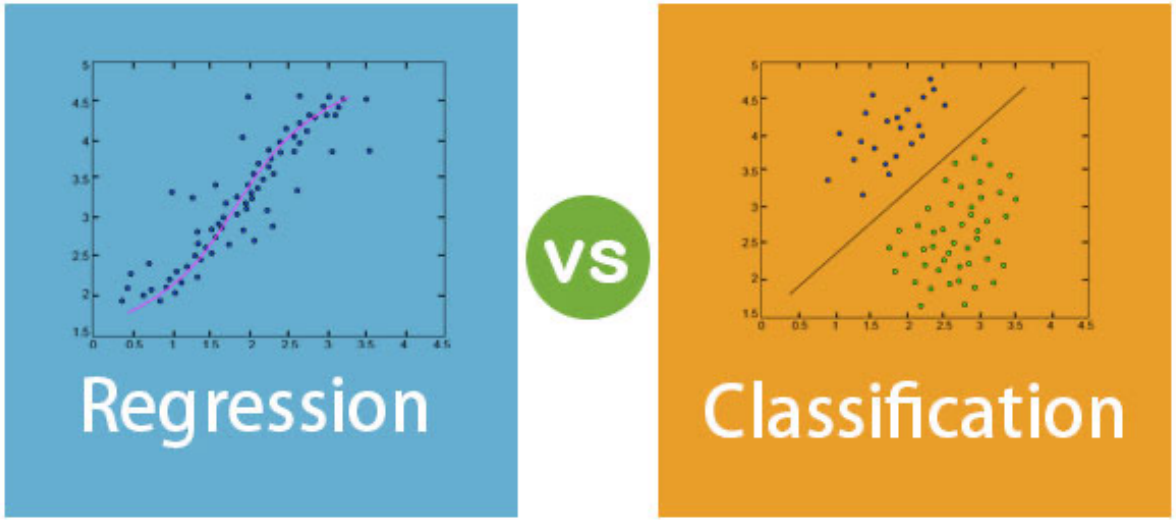
- 지도학습 (Supervised learning)
  - 입력 데이터와 정답을 함께 제공

- 비지도학습 (Unsupervised learning)
  - 입력 데이터로만 학습, 주어진 데이터를 자동으로 분석해 묶어주는 클러스터링
  - Self supervised learning, Autoencoder 등

3. 데이터 이해와 활용 영역 주요 개념

지도학습

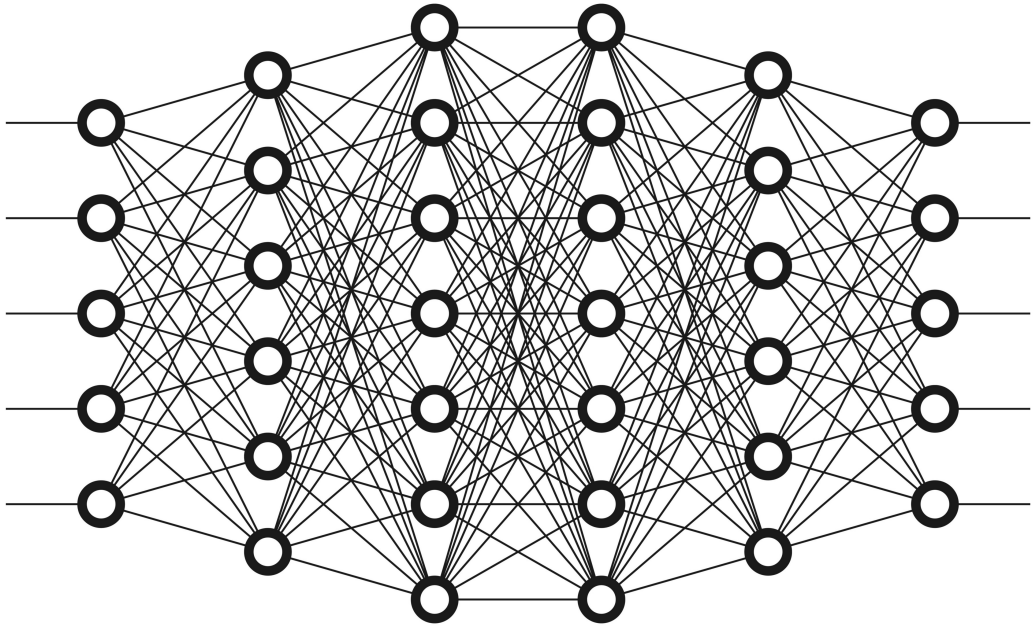
- 분류 (Classification)
  - 모델의 예측 값이 이산 값 (discrete value)으로 나오는 유형
  - 동물 사진 분류, 고혈압 유무 판단
- 회귀 (Regression)
  - 모델의 예측 값이 연속적인 값 (Continuous value)으로 나오는 유형
  - 하루 예상 매출액, 주가, 부동산 가격



3. 데이터 이해와 활용 영역 주요 개념

딥 러닝 (Deep learning)

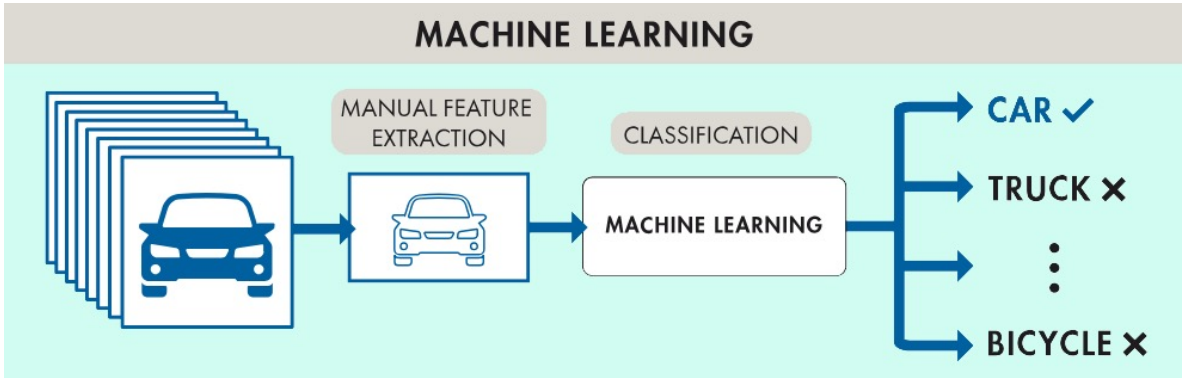
- 많은 양의 신경층 (Neural Network)에 입력된 데이터가 여러 층을 거치면서 특징이 추출되고, 자동으로 추상적인 지식을 추출해 모델을 학습시키는 방법

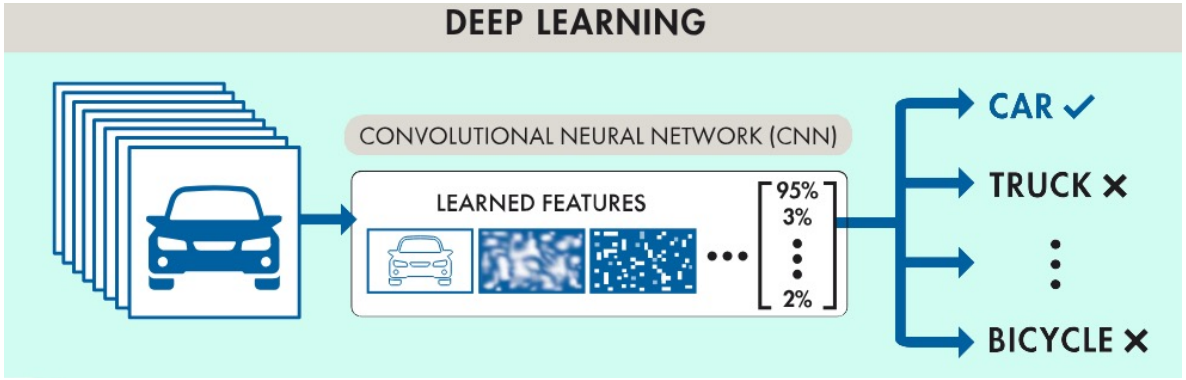


3. 데이터 이해와 활용 영역 주요 개념

딥 러닝 (Deep learning) VS 머신 러닝 (Machine learning)

- 머신러닝의 경우 학습 데이터에서 학습에 사용할 특성 (feature)을 수동으로 찾아서 사용
- 딥 러닝은 학습에 사용할 특성을 자동으로 찾아서 사용 (feature를 찾는 것까지 학습)

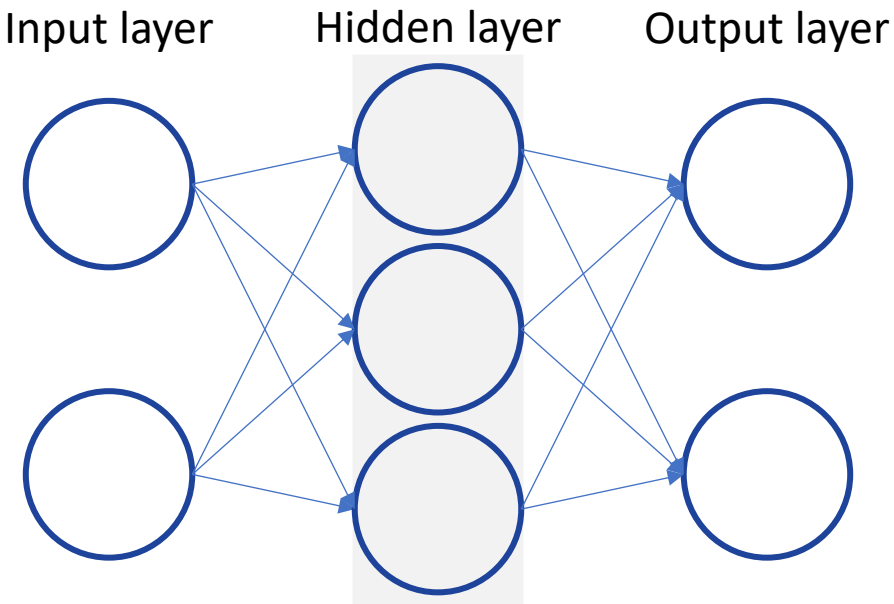




3. 데이터 이해와 활용 영역 주요 개념

인공 신경망 (Neural Network)

- 신경망은 여러 퍼셉트론을 서로 연결한 일종의 Network로, 데이터를 통해 스스로 학습하여 가중치를 찾아냄
- 신경망이 학습한다는 것은 특정 데이터에 맞게 가중치를 찾아 업데이트 하는 것을 의미

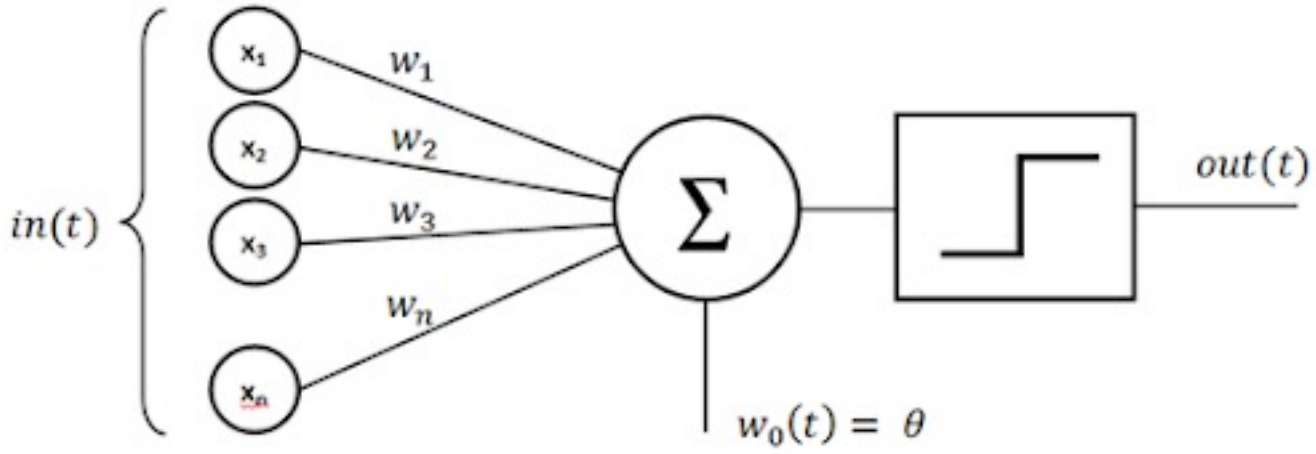


3. 데이터 이해와 활용 영역 주요 개념

Perceptron

- 퍼셉트론은 딥 러닝 네트워크를 구성하는 가장 작은 단위로 인간의 뇌 내에 있는 뉴런을 모델링 한 것
  - 다수의 퍼셉트론이 서로 연결되어 네트워크의 층 (layer)을 형성
- 입력값, 가중치, 편향, 활성화 함수로 구성됨

$$Y = f(W \cdot X + b)$$



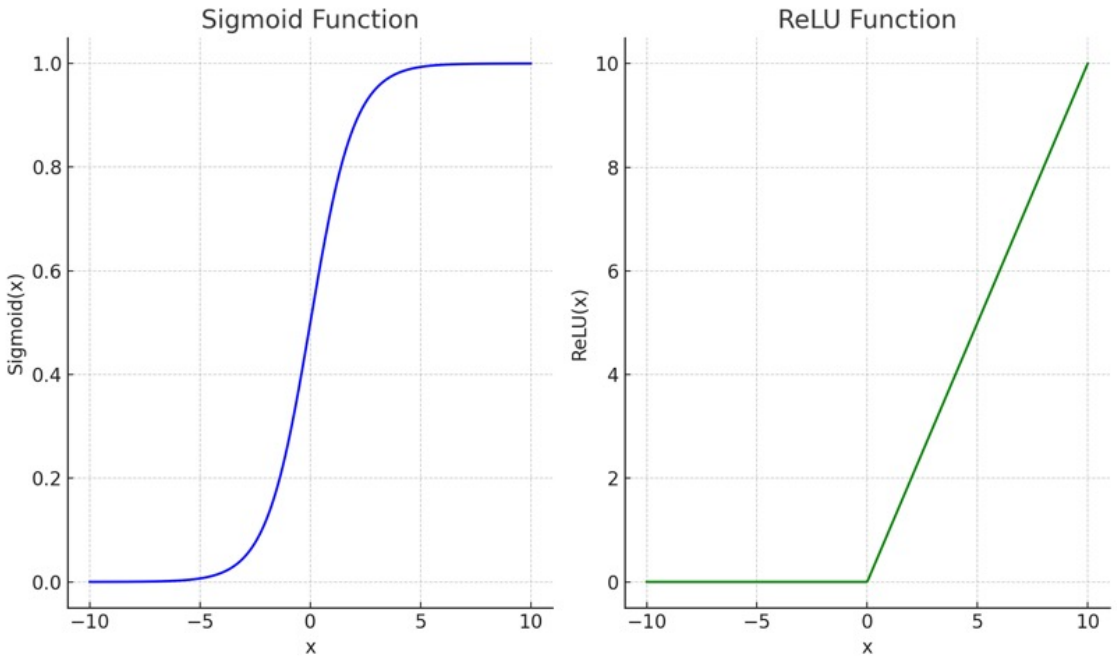
3. 데이터 이해와 활용 영역 주요 개념

Activation Function

- 활성화 함수는 퍼셉트론이 출력 값을 다음 퍼셉트론으로 전달할지를 결정할 때 사용하는 함수
- 일반적으로 비 선형 함수가 사용되어 모델이 입력 값에서 복잡한 패턴을 학습할 수 있게 함
  - Sigmoid, ReLU, Tanh 등의 함수가 사용됨

$$f(z) = \frac{1}{1 + e^{-z}}$$

$$f(z) = \max(0, z)$$

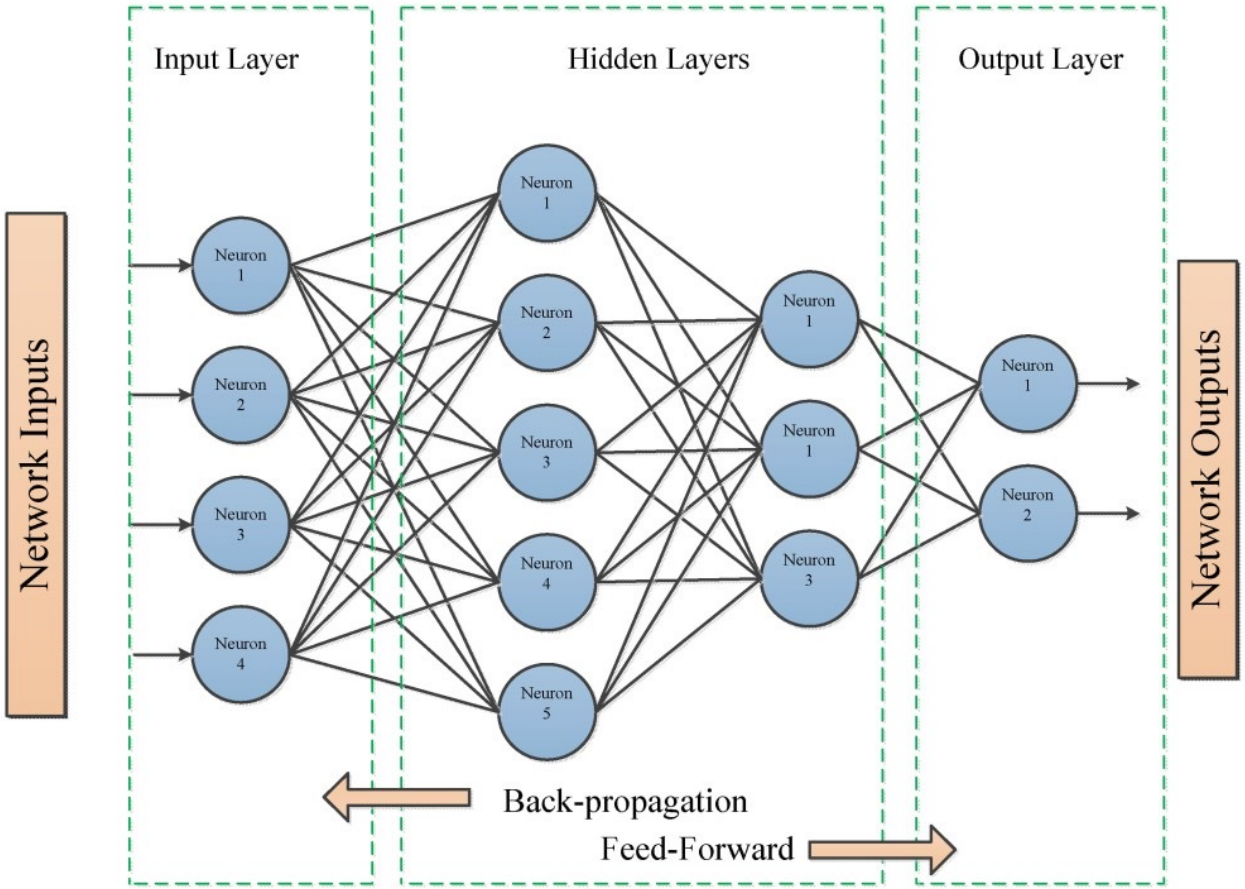


3. 데이터 이해와 활용 영역 주요 개념

Forward/Backward Propagation

- 순전파는 신경망에서 입력 데이터가 네트워크를 통과하여 출력을 생성하는 과정이다

- 역전파는 순전파의 결과를 바탕으로 네트워크의 가중치를 조정하는 과정으로, 손실 함수가 사용된다.



3. 데이터 이해와 활용 영역 주요 개념

Loss Function

- 손실 함수는 네트워크의 출력 값이 실제 정답 (ground truth)와 얼마나 다른지 측정하는 함수
- 손실 함수의 출력은 네트워크의 학습에 사용됨
- 대표적으로 Mean Squared Error 함수와 Cross Entropy 함수 등이 사용됨

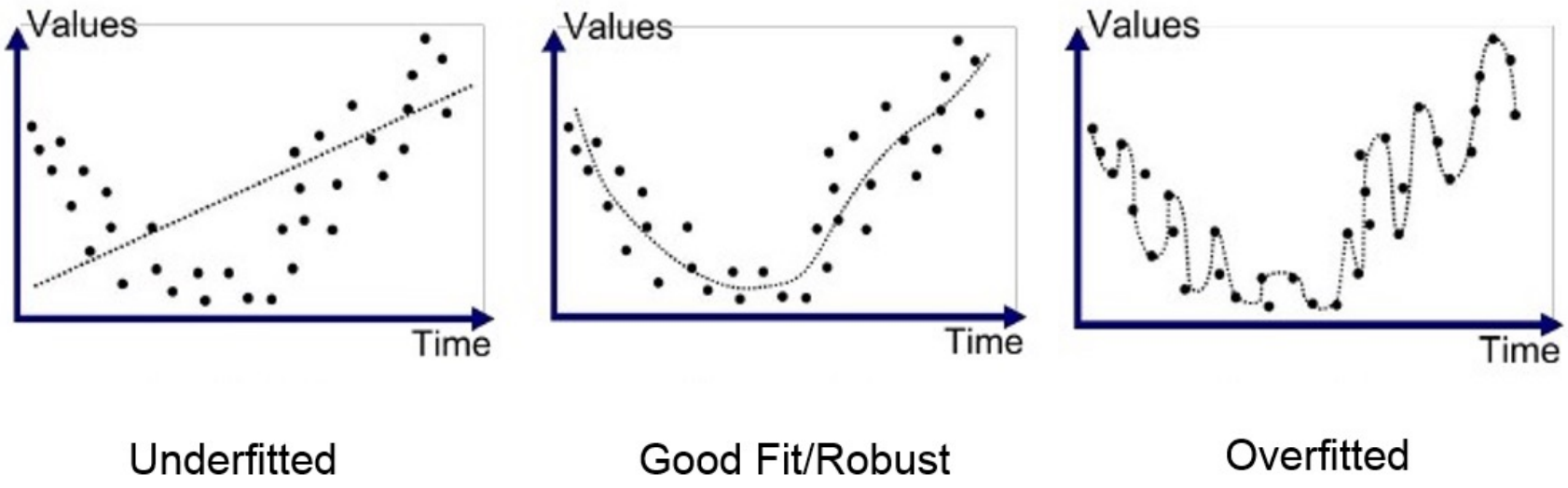
$$L(y, \hat{y}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2$$

$$L(y, \hat{y}) = -\sum_{i=1}^{n} y_i \log(\hat{y_i})$$

3. 데이터 이해와 활용 영역 주요 개념

Overfitting

- 모델이 학습 데이터를 과도하게 학습하여(암기) 다른 데이터에서 제대로 동작하지 못 하는 상태
- 모델이 일반화 되어 있지 않고 훈련 데이터에서만 제대로 동작
- Training error는 낮아지는 반면, Test error는 높아진다.



3. 데이터 이해와 활용 영역 주요 개념

Overfitting을 막기 위한 방법

- 더 많은 훈련 데이터 사용

- Data augmentation & Noise

3. 데이터 이해와 활용 영역 주요 개념

Overfitting을 막기 위한 방법

- Regularization

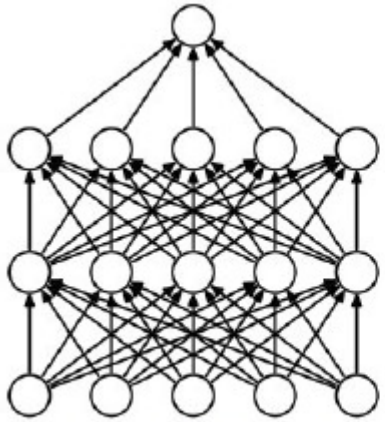
L1 regularization on least squares:

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \sum_j \left(t(\mathbf{x}_j) - \sum_i w_i h_i(\mathbf{x}_j) \right)^2 + \lambda \sum_{i=1}^{k} |w_i|$$

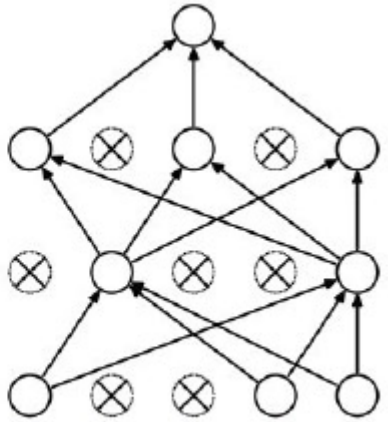
L2 regularization on least squares:

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \sum_j \left(t(\mathbf{x}_j) - \sum_i w_i h_i(\mathbf{x}_j) \right)^2 + \lambda \sum_{i=1}^{k} w_i^2$$

- Dropout & Dropconnection



(a) Standard Neural Net



(b) After applying dropout.

4. 기출문제 풀이

4. 기출문제 풀이

5

객관식
5점

검토하기

다음에 대한 원인과 근본적인 해결안을 가장 잘 제시한 의견을 고르시오.

① A회사는 고객이 회원가입을 하고 첫 주문을 했는데, 고객이 주문을 취소하는 바람에 고객의 회원가입 정보까지 제되었다.

② B회사는 고객이 전화번호를 변경할 경우, 모든 주문 내역의 고객정보를 같이 수정하여야한다. 따라서 주기적으로 잘못된 데이터에 대한 데이터 보정작업을 수행하고 있다.

○ ① 최 대리: 데이터 설계가 잘못되어 이상현상이 발생한 것 같네요. 데이터 모델에 대한 점검이 필요합니다.

○ ② 박 대리: 프로그램 개발 시 에러처리를 잘못 한 것 같네요. 원하지 않는 데이터에 대한 수정/삭제가 발생하지 않도록 에러처리 단계를 추가해야 합니다.

○ ③ 이 대리: 트랜잭션 설계가 잘못된 것 같군요. 트랜잭션을 더 작은 단위로 끊어서 주문 변경과 회원정보 변경이 별개로 이루어지도록 해야 할 필요가 있습니다.

○ ④ 김 대리: 동시성 제어가 안됐군요. 데이터에 대한 락(Lock) 관리 수준을 객체 직렬화(Serializable)로 변경해서 불필요한 데이터 삭제나 변경이 발생하지 않도록 해야 합니다.

4. 기출문제 풀이

6

ⓘ단답형
10점

☐
검토하기

다음 T_EMP 테이블의 SQL문 실행 전과 후의 값을 참고하여 ㉠, ㉡에 들어갈 정확한 DCL구문을 적으시오(각 5점)

▪ T_EMP - 실행 전

EMP_NO	SALARY
0001	1000
0002	2000
0003	3000

▪ [SQL문]

```
BEGIN TRANSACTION A
UPDATE T_EMP SET SALARY=3000 WHERE EMP_NO='0001'
㉠ TRANSACTION A

BEGIN TRANSACTION B
UPDATE T_EMP SET SALARY=5000 WHERE EMP_NO='0003'
㉡ TRANSACTION B
```

Commit

Rollback

▪ T_EMP - 실행 후

EMP_NO	SALARY
0001	3000
0002	2000
0003	3000

4. 기출문제 풀이

데이터 이해와 활용 객관식

- **다음 중 데이터베이스의 뷰(View)에 대한 설명으로 옳지 않은 것은?**

  ① 뷰는 물리적으로 구현된 테이블이다.

  ② 뷰는 다른 뷰를 기반으로 정의될 수 있다.

  ③ 뷰는 보안상 내부의 세부적인 데이터를 감출 수 있다.

  ④ 뷰는 복잡한 질의를 단순화시킬 수 있다.

- 정답: **1**, 뷰(View)는 논리적으로 정의된 가상 테이블로, 물리적으로 구현되지 않는다. 뷰는 기본 테이블로부터 유도된 테이블이며, 실제 데이터를 저장하지 않고 쿼리 정의만을 저장한다.

4. 기출문제 풀이

데이터 이해와 활용 객관식

- **다음 중 데이터베이스 설계 단계를 순서대로 나열한 것은?**

  ① 요구사항 분석 → 개념적 설계 → 논리적 설계 → 물리적 설계

  ② 요구사항 분석 → 논리적 설계 → 개념적 설계 → 물리적 설계

  ③ 개념적 설계 → 요구사항 분석 → 논리적 설계 → 물리적 설계

  ④ 물리적 설계 → 논리적 설계 → 개념적 설계 → 요구사항 분석

- 정답: **1**

4. 기출문제 풀이

데이터 이해와 활용 객관식

- **텐서플로(TensorFlow)에 대한 설명으로 옳은 것은?**

  ① 머신러닝을 위한 엔드 투 엔드 오픈소스 플랫폼이다.

  ② 관계형 데이터베이스 관리 시스템이다.

  ③ 네트워크 보안을 위한 방화벽 소프트웨어이다.

  ④ 웹 애플리케이션 개발을 위한 프레임워크이다.

- 정답: **1,** 텐서플로(TensorFlow)는 구글에서 개발한 머신러닝과 딥러닝을 위한 오픈소스 라이브러리이다.

4. 기출문제 풀이

데이터 이해와 활용 객관식

- **딥러닝에서 사용되는 활성화 함수(Activation Function)가 아닌 것은?**

  ① ReLU

  ② Sigmoid

  ③ Tanh

  ④ K-means

- 정답: **4,** ReLU(Rectified Linear Unit), Sigmoid, Tanh(Hyperbolic Tangent)는 모두 딥러닝에서 흔히 사용되는 활성화 함수입니다. K-means는 활성화 함수가 아니라 군집화 알고리즘이다.

4. 기출문제 풀이

데이터 이해와 활용 객관식

- **과적합(Overfitting)을 방지하기 위한 방법이 아닌 것은?**

  ① 데이터 증강

  ② 드롭아웃

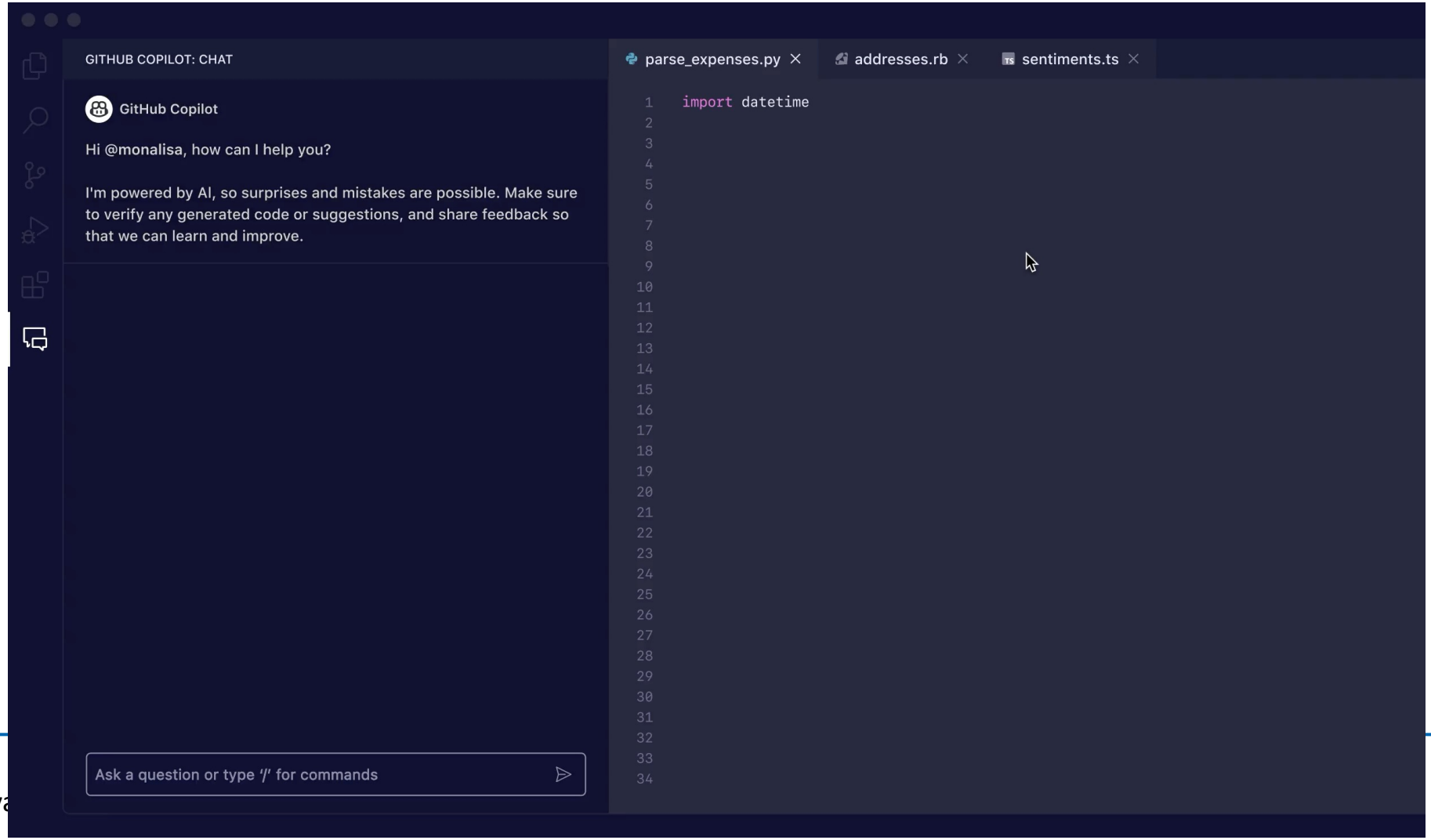
  ③ 정규화

  ④ 학습률 증가

- 정답: **4,** 데이터 증강, 드롭아웃, 정규화는 모두 과적합을 방지하기 위한 기법이다. 반면, 학습률을 증가시키는 것은 과적합 방지와 직접적인 관련이 없으며, 오히려 학습의 불안정성을 높일 수 있다.

TMI

- AI powered software development



TMI

How to Keep Your Job Safe in The Age of AI

- Continuous Learning
- Adaptability
- Domain Knowledge
- Communication Skills
- Collaborate with AI

famous77@kaist.ac.kr