중간 프로젝트

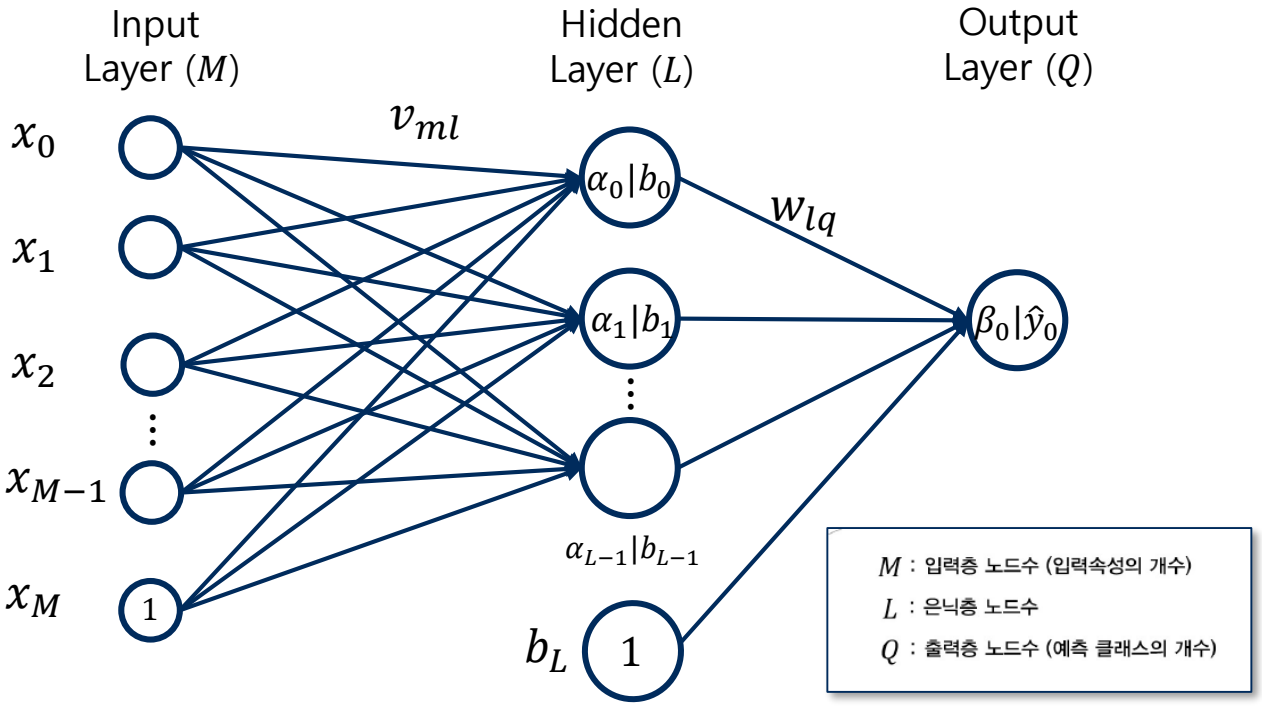
방 수 식 교수
(bang@tukorea.ac.kr)

한국공학대학교 전자공학부

2024년도 1학기
머신러닝실습 & 인공지능설계실습1

프로젝트 개요

- **본 프로젝트는 Activation Function이 Sigmoid인 Two-Layer Neural Network을 활용하여, 심근경색 데이터 Set 예측을 목표로 함.**
  - 데이터 전처리 방법에 대한 이해 (데이터 Set 로부터 특징 추출하는 방법)
  - Activation Function을 활용한 인공신경망에 대한 이해
  - 머신의 Weight를 업데이트 하는 방식에 대한 이해



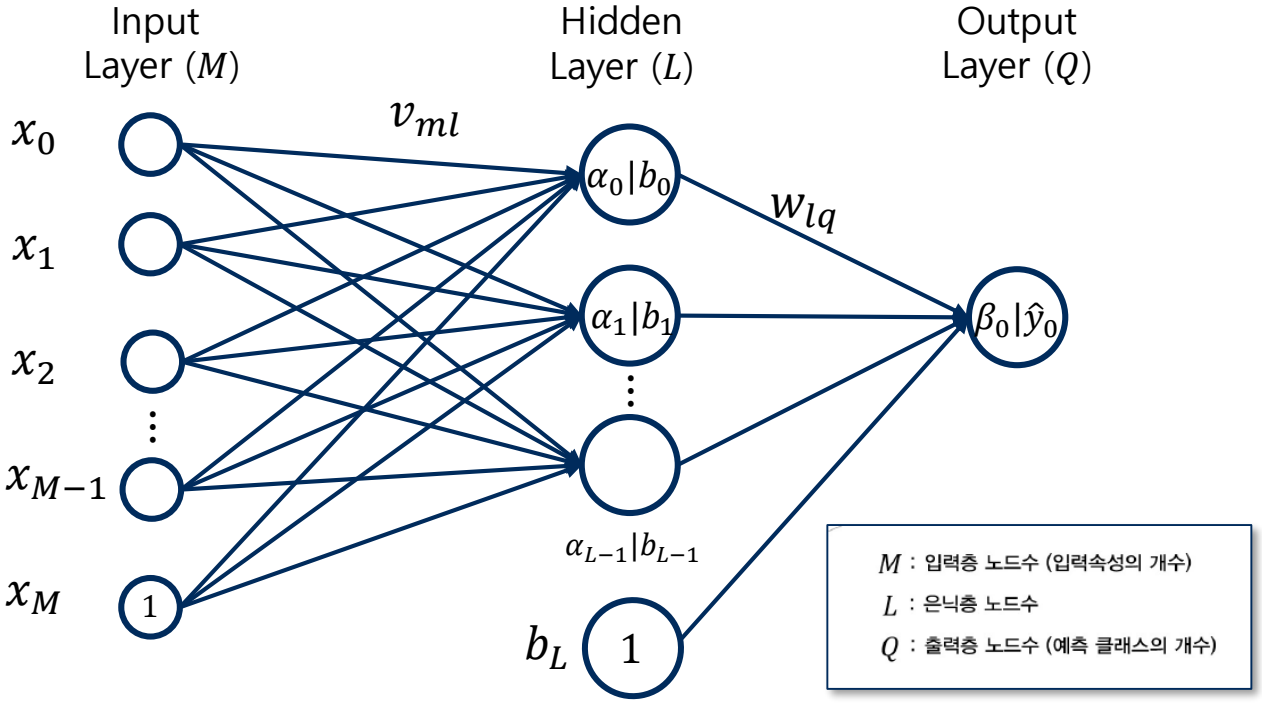
본 프로젝트에서의 심근경색 데이터 set

- 심근경색 데이터는 심장 건강 상태에 대한 대형 데이터
- 특징 수를 80개에서 15개로 축소
- Heart disease (심근경색 여부, y값) : 'yes', 'no' 로 분류, 마지막 열에 표기
- Training 데이터 : 3000개 ("yes" : 500, "no" : 2500")

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	gender	BPMeds	BMI	blood sugar levels	emic backgr	urological dis	meat intake	Heart rate	Weight	Cholesterol	Diabetes status	High blood pressure	Age	Smoking	height	heart disease
2	female	0	21.61	93	middle schoo	no	3	75	61	199	0	0	30	1	116	no
3	female	0	21.18	76	high school	no	0	70	64	331	0	0	60	0	128	no
4	male	0	26.17	71	mentary sch	no	1	90	61	355	0	0	40	1	113	no
5	female	0	17.92	73	mentary sch	no	0.2	75	49	194	0	0	40	1	100.5	no
6	female	0	24.16	NaN	high school	no	0	75	66	197	0	0	50	0	126	no
7	female	0	31.57	112	mentary sch	no	0.6	72	64	NaN	0	0	40	1	130	no
8	male	0	24.36	71	NaN	no	0.8	60	53	196	0	0	40	1	107	no
9	female	NaN	24.83	78	high school	no	0	80	64	391	0	1	60	0	126	no
10	female	0	33.19	76	mentary sch	no	0	75	60	238	0	0	40	0	118	no
11	female	0	35.99	88	mentary sch	no	3	82	80.5	174	0	1	40	1	158.5	no
12	female	0	24.53	77	high school	no	0	65	57	201	0	0	70	0	110	no
13	female	0	17.48	57	college schoo	no	0	75	53	205	0	0	40	0	109	no
14	female	0	20.51	66	college schoo	no	0	96	54.5	260	0	0	40	0	100	no
15	female	0	18.8	NaN	middle schoo	no	2	79	48	234	0	0	50	1	98	yes
16	male	0	23.03	93	high school	no	0	77	70	152	0	1	50	0	120	no
17	female	0	32.1	68	high school	no	0	72	90	306	0	1	50	0	195	no
18	female	0	34.52	72	mentary sch	no	0	80	62	212	0	0	50	0	132	no
19	female	0	24.59	NaN	NaN	no	0	55	58	265	0	0	60	0	123	no
20	male	0	28.3	70	college schoo	no	4	60	79	213	0	1	60	1	162	yes

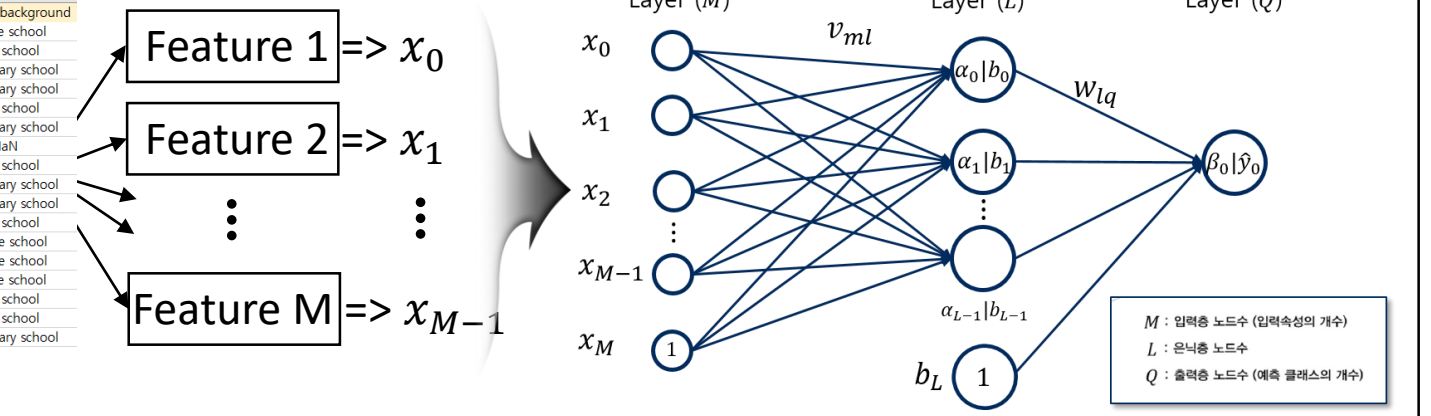
모델의 구조 (제한 조건)

- **Activation Function(활성화 함수) : Sigmoid 함수 (그 외 함수 불허)**
- **Hidden Layer 수 : 1**
- **Hidden Layer의 Node 수 : 제한 없음**
- **Input node 수 : 제한 없음**
- **Output node 수 : 1개**



특징 추출 (Feature Extraction)

	A	B	C	D	E
1	gender	BPMeds	BMI	blood sugar	academic background
2	female	0	21.61	93	middle school
3	female	0	21.18	76	high school
4	male	0	26.17	71	elementary school
5	female	0	17.92	73	elementary school
6	female	0	24.16	NaN	high school
7	female	0	31.57	112	elementary school
8	male	0	24.36	71	NaN
9	female	NaN	24.83	78	high school
10	female	0	33.19	76	elementary school
11	female	0	35.99	88	elementary school
12	female	0	24.53	77	high school
13	female	0	17.48	57	college school
14	female	0	20.51	66	college school
15	female	0	18.8	NaN	middle school
16	male	0	23.03	93	high school
17	female	0	32.1	68	high school
18	female	0	34.52	72	elementary school



Regression

- 경계값 설정
  - Output layer 개수 : 1개
  - $\hat{y} < 0.5$ ➔ 'no' 로 예측
  - $0.5 \leq \hat{y}$ ➔ 'yes'로 예측

Training data에 대한 최종 Confusion Matrix (보고서에 삽입)

실제값 \ 예측값	0 (no)	1 (yes)	
0 (no)			
1 (yes)			

프로젝트 수행 요령 (1)

- **데이터 전처리 & 특징 추출**
  - **데이터에서 주어진 특징 15개와 직접 만든 특징을 활용해 차원축소된 vector를 출력하는 알고리즘(사용자 지정함수)구현**

    ex) 직접 만든 특징 : 혈당수치 x 심박수

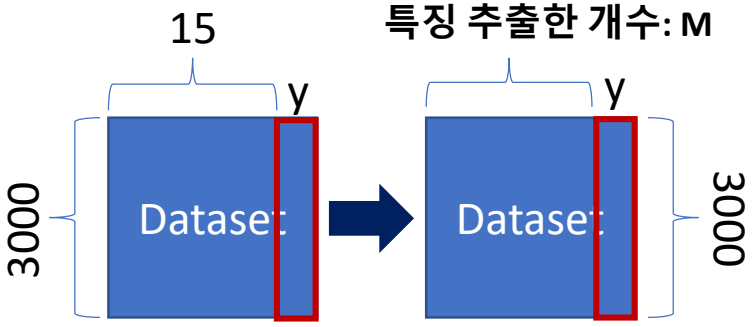
- **수업시간에 구현한 Raw level code 활용 (주간 실습에서 허용한 라이브러리* 외 사용금지)**

  * Math / Numpy / Matplotlib / Data Load 및 save를 위한 Pandas(제한적 허용)
- Weight 초기값, Learning rate, Hidden Node 수 조정을 통하여 신경망 모델 최적화

프로젝트 수행 요령 (2)

- **데이터 전처리 & 특징 추출**

```python
def select_features(directory):
    ''' 데이터 불러오기 '''
    path = directory + "heart_disease.csv"
    df = pd.read_csv(path)
    dataset = np.array(df)

    # 데이터 랜덤으로 섞기
    np.random.shuffle(dataset)

    ''' 데이터 전처리 '''
    # 데이터 형식 float으로 변환 필수
    코드 작성
    코드 작성

    ''' 특징 추출 '''
    y_data = dataset[:,-1]   # y값

    ######## 특징 1 ########
    # 주석으로 특징 설명
    특징 추출 코드 작성
    특징 추출 코드 작성
    feature_1 = ~~~~~

    ######## 특징 2 ########
    # 주석으로 특징 설명
    특징 추출 코드 작성
    특징 추출 코드 작성
    feature_2 = ~~~~~

    # 특징 합치기
    selectd_features = np.column_stack((feature_1, feature_2, ..., feature_))

    # 특징 데이터 마지막 열에 y값 추가
    features = np.column_stack((selected_features, y_data))

    return features
```



Input data size :
[3000,16]

Return 값인
output data size :
[3000,M+1]

프로젝트 제출물

- **최종 제출 파일**: 학번.zip (예시: 2020146999.zip)
  - ➢ 아래 5개의 파일을 하나의 파일로 압축하여 제출

    (제출요령을 지키지 않아 발생하는 오류는 본인 책임)

연번	파일명	파일 형식
(1)	w_hidden.csv	csv 파일
(2)	w_output.csv	csv 파일
(3)	학번_feature.py	py 파일
(4)	학번_code.pdf	pdf 파일
(5)	학번_report.pdf	pdf 파일

❖ (5) 보고서의 경우, 분반 별 공지 확인

제출물 (1), (2)

** L은 Hidden Layer Node 수*

- **Weight Matrix (csv 파일로 제출)**

① **Weight Matrix for Hidden Layer 입력 : L by M+1 (파일명: w_hidden.csv)**

② **Weight Matrix for Output Layer 입력 : 1 by L+1 (파일명 : w_output.csv)**

```
w_hidden = pd.read_csv('디렉토리/w_hidden.csv', header=None) #L by M+1 weight matrix 불러오기
w_output = pd.read_csv('디렉토리/w_output.csv', header=None) #1 by L+1 weight matrix 불러오기
```

※ 위의 코드로 본인이 제출하는 Weight Matrix가 잘 read 되는지 반드시 확인

※ 디렉토리는 Test하는 PC의 디렉토리로 입력할 예정

제출물 (3)

- **전처리 및 특징 추출 알고리즘이 포함된 1개의 python code 파일**
  - ❖ 전처리 및 특징 추출 알고리즘 부분만 발췌
  - ❖ 반드시 파일명 및 함수명에 대한 양식을 지킬 것

(파일명: 학번_feature.py)

```python
import pandas as pd
import numpy as np

def select_features(directory):
    ''' 데이터 불러오기 '''
    path = directory + "heart_disease.csv"
    df = pd.read_csv(path)
    dataset = np.array(df)

    # 데이터 랜덤으로 섞기
    np.random.shuffle(dataset)

    ''' 데이터 전처리 '''
    # 데이터 형식 float으로 변환 필수
    코드 작성
    코드 작성

    ''' 특징 추출 '''
    y_data = dataset[:,-1]   # y값

    ######## 특징 1 ########
    # 주석으로 특징 설명
    특징 추출 코드 작성
    특징 추출 코드 작성
    feature_1 = ~~~~~

    ######## 특징 2 ########
    # 주석으로 특징 설명
    특징 추출 코드 작성
    특징 추출 코드 작성
    feature_2 = ~~~~~

    # 특징 합치기
    selectd_features = np.column_stack((feature_1, feature_2, ..., feature_))

    # 특징 데이터 마지막 열에 y값 추가
    features = np.column_stack((selected_features, y_data))

    return features
```

제출물 (4), (5)

- **본인이 프로젝트 수행에 사용한 전체 python code(주석 포함)**
  (파일명: 학번_code.pdf)
  - ❖ hwp ➔ pdf 파일로 옮겨서 제출
  - ❖ E-class 내 표절 검사 시스템을 활용하여 code 표절률 30% 이상일 시, 0점 처리

- **보고서 (파일명: 학번_report.pdf)**
  - ❖ ppt ➔ pdf 파일로 옮겨서 제출

평가 기준 (100점 만점)

◆ (10점) 알고리즘 code 구현 여부

- 가이드라인을 지키지 않아 오류 발생 시, 0점 처리

◆ (60점) 정확도 평가

- A: Input Node수(사용한 특징 수), B: Hidden Node 수, C: 최종 Test set 정확도 [%]
- 정확도 평가점수 = **0.8*C – (2*A + B)** *=> 60점 초과시, 60점으로 부여*
- 가이드라인을 지키지 않아 오류 발생 시, 정확도 30%로 가정 => 15점 처리

◆ (25점) 발표자료(보고서) 평가

- (해결능력) 전처리,특징 추출 및 모델이 적절하게 설계되었는가? A(10), B(8), C(6)
- (구성) 분석 및 결과 그림이 적절하게 표현되었는가? A(10), B(8), C(6)
- (분량) 표지 포함 10 page = 5점 (초과/미만 페이지 당 –1점)

◆ (+@) 발표 평가

- 수업시간에 발표(약 10분 분량)를 원하는 학생에 대해
- 발표 수준에 따라 5~20점 부여