인공지능 기초

인공지능_ Day02

김새봄

2023. 06. 27

CONTENTS

NAVER Cloud x BIT BITCamp

인공지능 기초개념

인공지능 기초개념 - 퍼셉트론

퍼셉트론의 학습 원리 (손실 함수)

손실 함수의 최소점 찾는 기법
(경사 하강법)

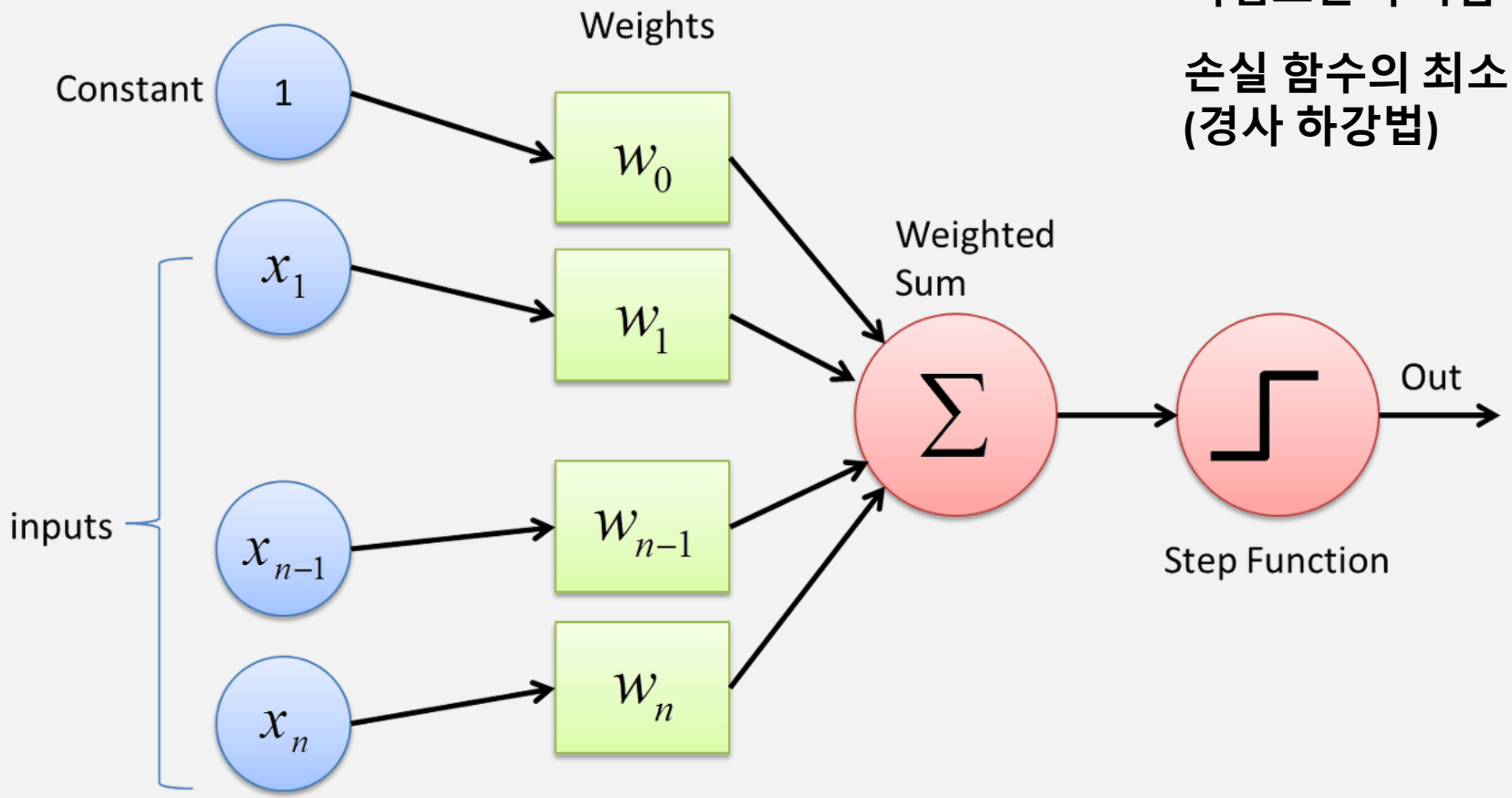


Figure 1. 인공신경망 퍼셉트론

인공지능 기초개념 - 다층 퍼셉트론

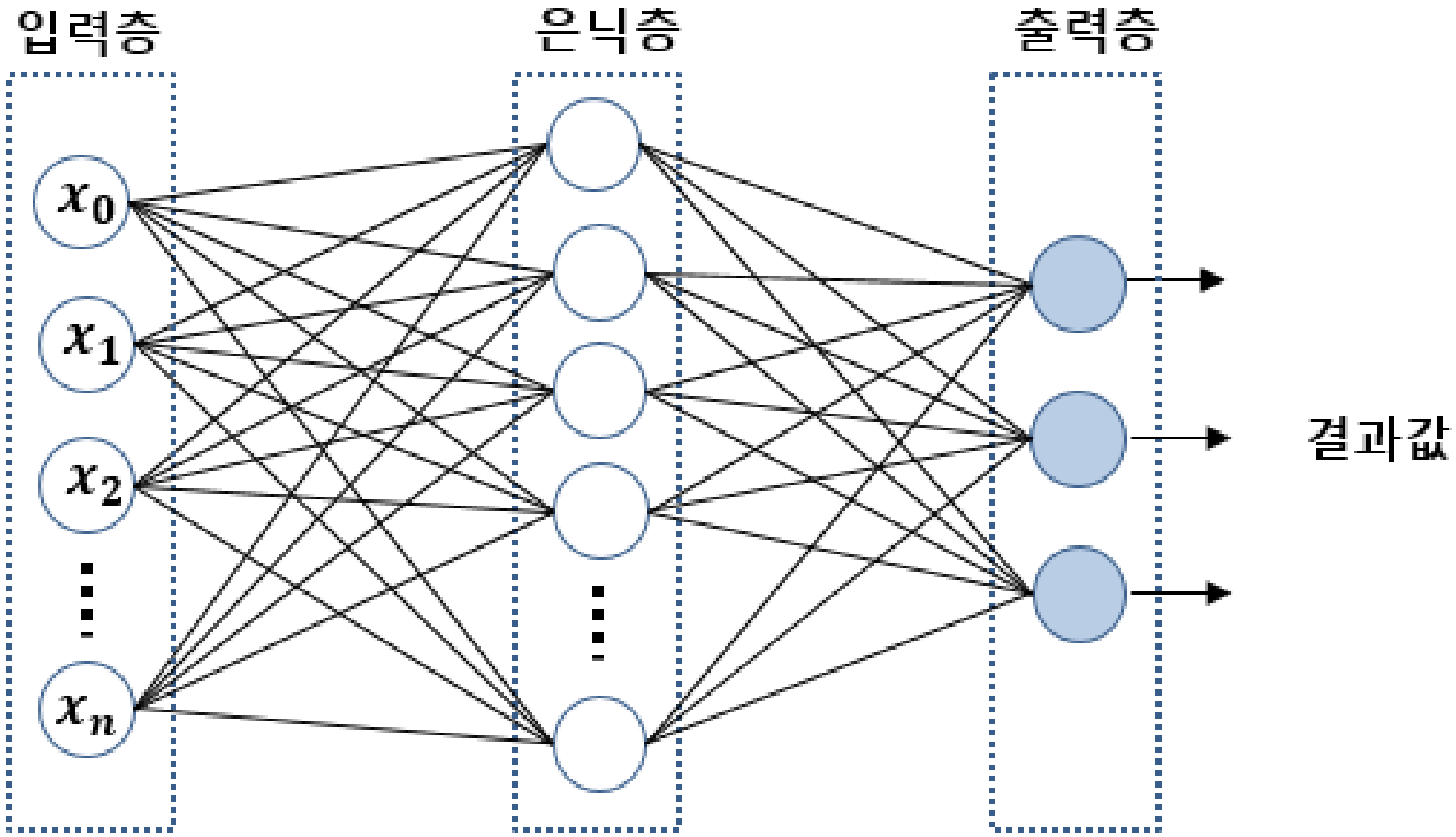
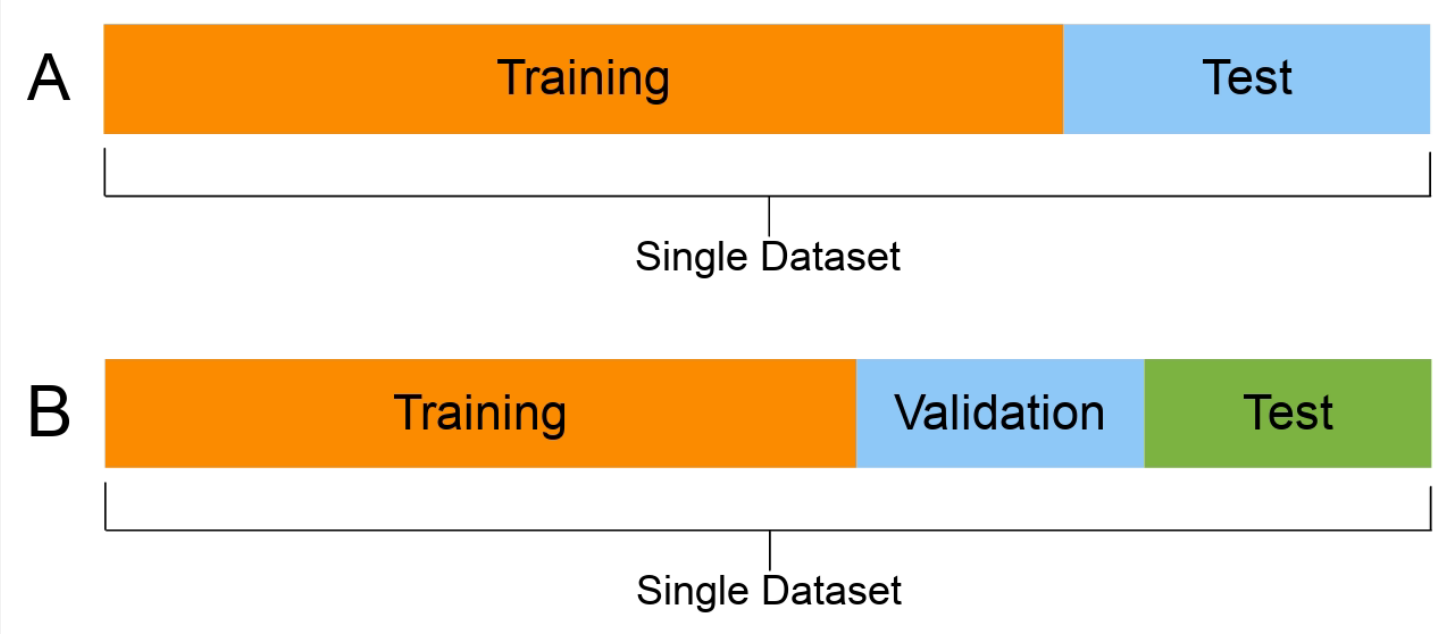


Figure 2. MLP(다층 퍼셉트론)

훈련 테스트 데이터 셋

데이터 셋 - Train, Test



출처: https://en.wikipedia.org/wiki/Training,_validation,_and_test_sets

Figure 3. Train, Test, Validation Set

Train Set

모델의 학습만을 위해서 사용
parameter나 feature 등을 수정하여
모델의 성능을 높이는 작업에 사용

Test Set

최종적으로 모델의 성능을 평가
실사용 되었을 때 모델이 얼마나 좋은 성
능을 발휘 할 수 있을지 알아보는 것

Validation Set

모델의 학습에 직접적으로 관여하지 않음
학습이 끝난 모델에 적용
최종적으로 모델을 fine tuning하는 데에 사용

회귀분석과 분류분석

회귀분석과 분류분석

- 회귀분석(Regression)

  연속된 값을 예측

  과거의 주가 데이터를 가지고 미래 주가를 예측하거나, 자동차 배기량이나 연식 등 중고차 정보를 이용하여 가격을 예측

- 분류분석(Classification)

  종류를 예측

  클래스 0 또는 1 중에서 선택하는 이진 분류(binary classification) 또는 3개 이상의 클래스 중에서 하나를 선택하는 다중 분류(multi classification)

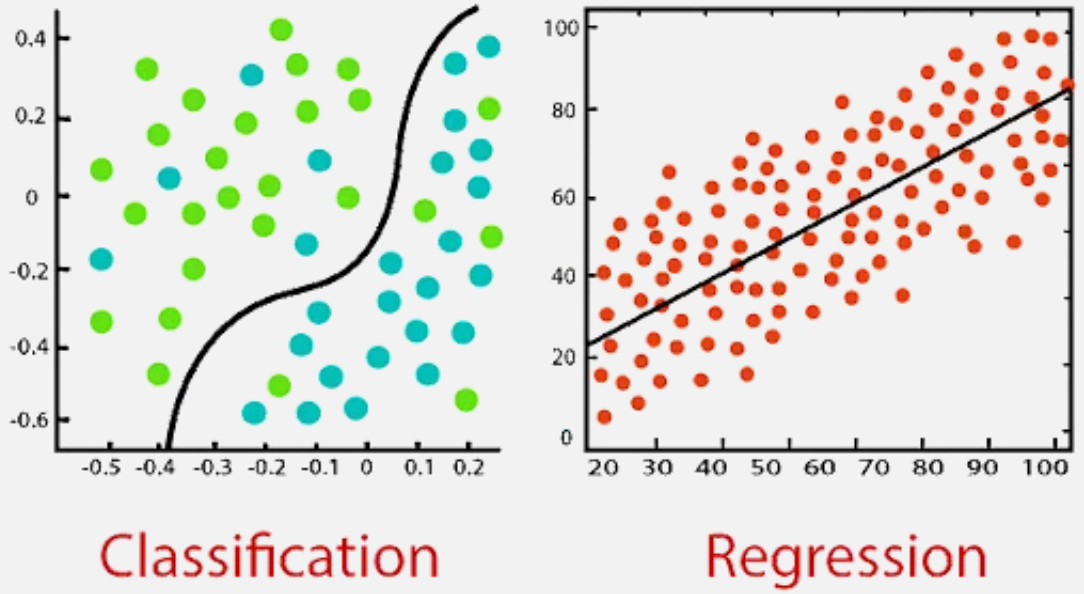
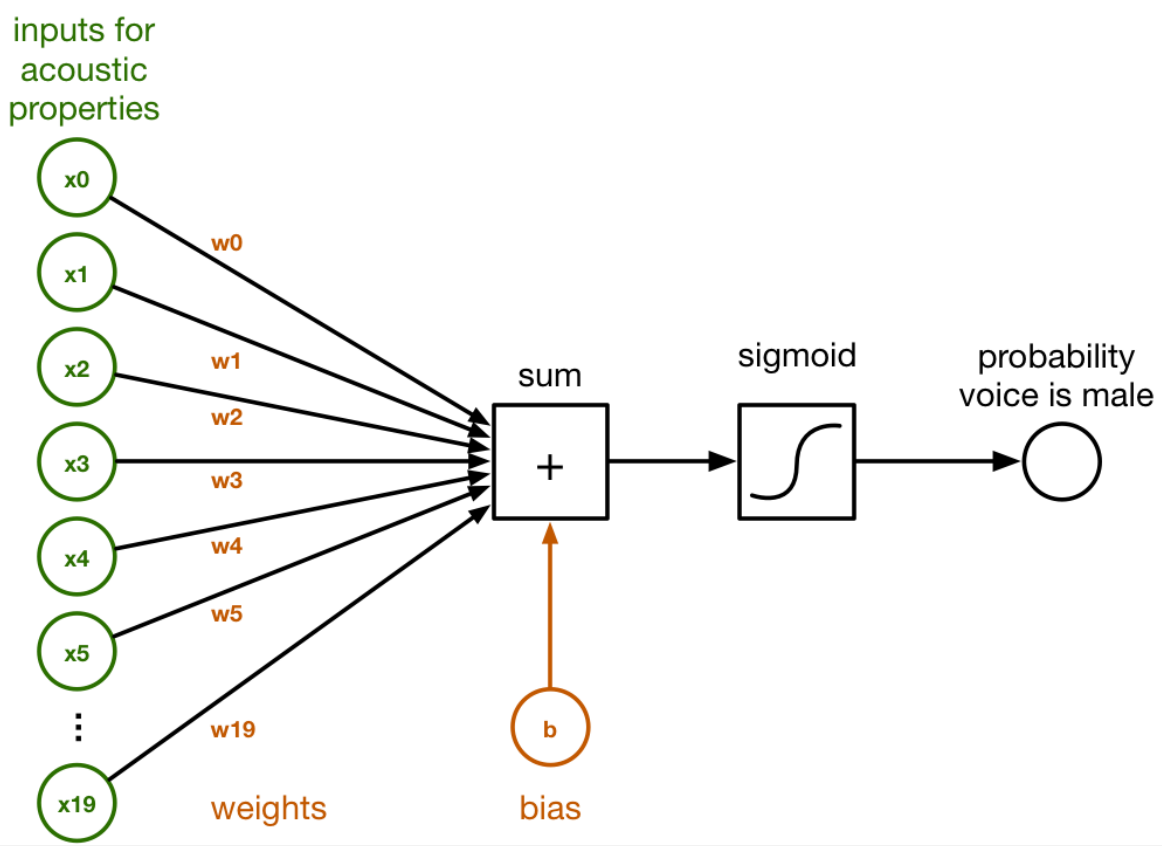


Figure 4. 분류분석과 회귀분석

이미지 출처: https://www.simplilearn.com/regression-vs-classification-in-machine-learning-article

이진분류



출처 https://wikidocs.net/41256

Figure 5. 이진분류 모델

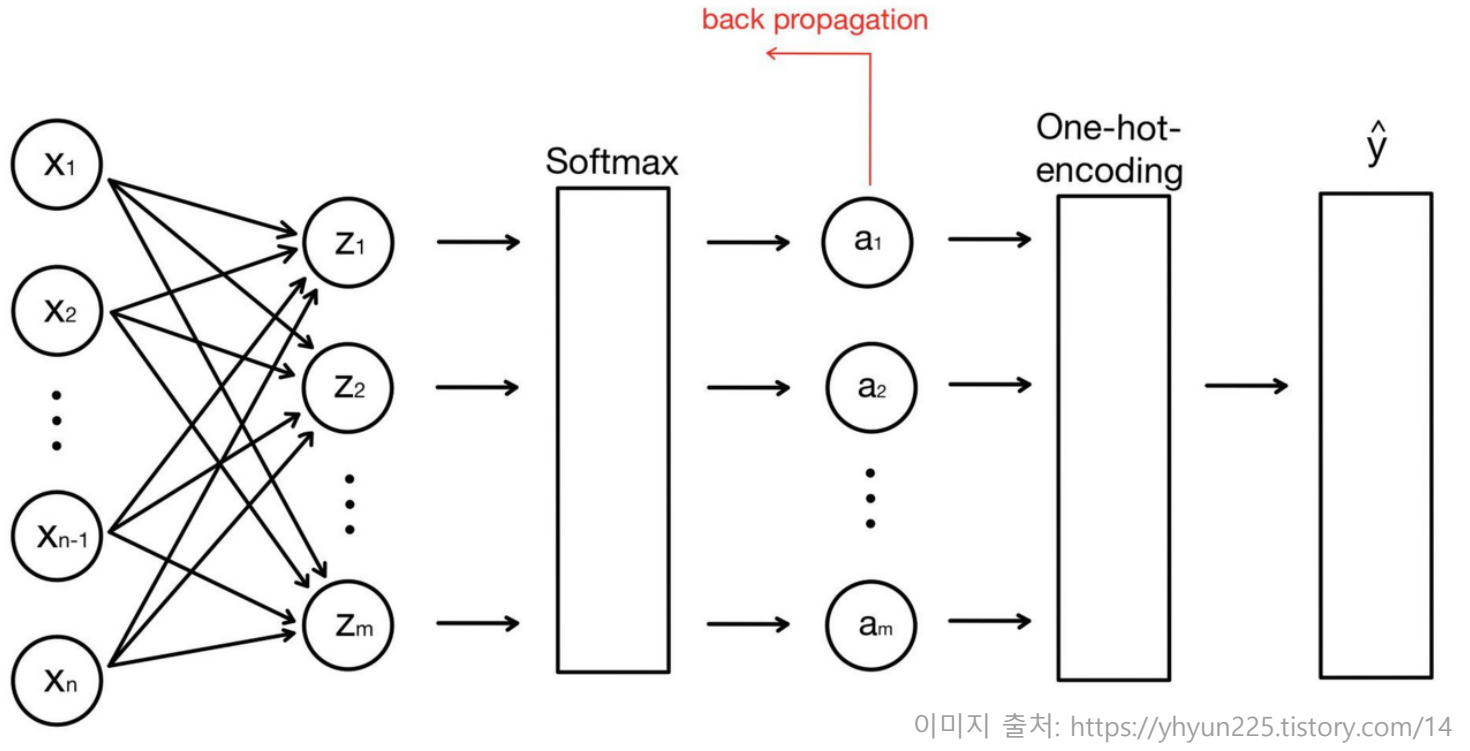
- 이진분류의 경우 참(True) 또는 거짓(False)을 판별하기 때문에 출력 값이 하나
- 출력 값을 sigmoid 함수를 이용하여 0과 1로 가공
- 로지스틱 회귀(Logistic Regression)는 이진분류모델을 분석하기 좋은 모델
  - 직선보다 적절한 곡선을 통해 분류를 함

다중분류

다중 분류 모델은 타깃의 종류가 여러 개이기 때문에 출력 값도 여러 개

다중 분류는 softmax 함수를 사용하여 0과 1사이의 값으로 가공

다중 분류 모델에서는 one-hot encoding이라는 기법을 사용



이미지 출처: https://yhyun225.tistory.com/14

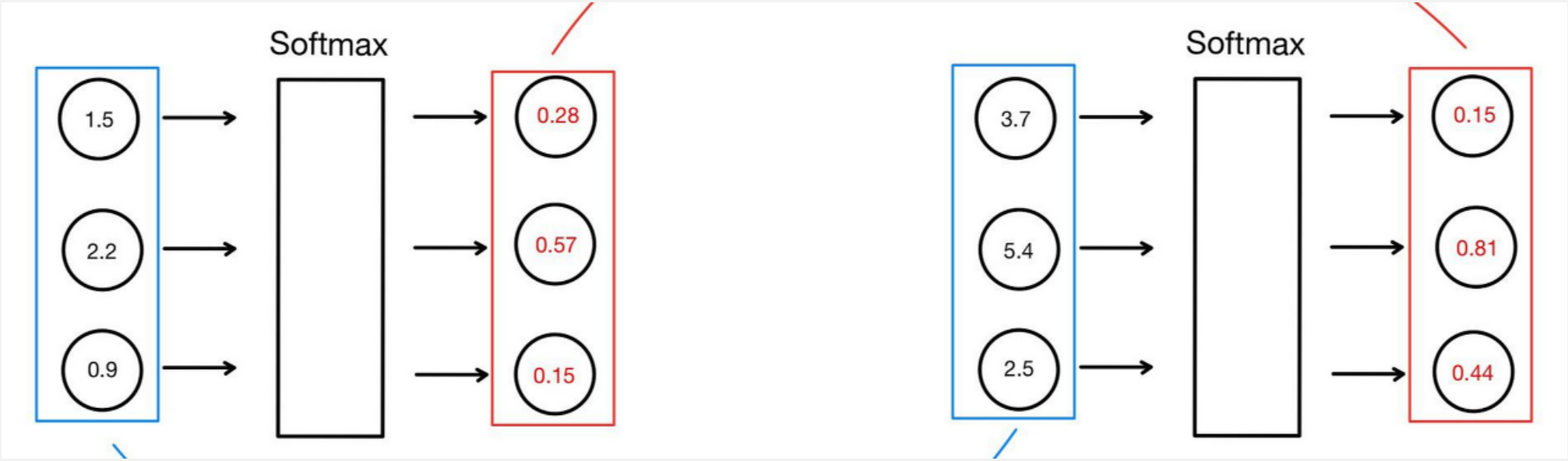
Figure 5. 다중분류 모델

다중분류

다중 분류 모델은 타깃의 종류가 여러 개이기 때문에 출력 값도 여러 개

다중 분류는 softmax 함수를 사용하여 0과 1사이의 값으로 가공

다중 분류 모델에서는 one-hot encoding이라는 기법을 사용



이미지 출처: https://yhyun225.tistory.com/14

Figure 6. 다중분류 모델의 Softmax 함수

회귀분석과 분류분석의 손실함수

손실함수(Loss Function)

1. binary_crossentropy(이항교차 엔트로피)

   - y값이 0과 1인 이진 분류기를 훈련할 때 자주 사용되는 손실 함수

   - 활성화 함수(activation) : sigmoid 사용 (출력값이 0과 1사이의 값)

손실함수(Loss Function)

2. categorical_crossentropy (범주형 교차 엔트로피)

- Y 클래스가 3개 이상일 경우, 즉 다중 분류에서 사용
- 활성화 함수(activation) : softmax 사용 (모든 벡터 요소의 값은 0과 1사이의 값이 나오고, 모든 합이 1이 됨)
- 라벨이 (0,0,1,0,0) , (0,1,0,0,0) 과 같이 one-hot encoding 된 형태로 제공될 때 사용 가능

손실함수(Loss Function)

3. sparse_categorical_crossentropy

- categorical crossentropy와 같이 다중 분류에서 사용
- one-hot encoding 된 상태일 필요 없이 정수 인코딩 된 상태에서 수행 가능
- 라벨이 (1,2,3,4) 와 같이 정수형태로 사용

손실함수(Loss Function)

4. 평균 제곱 오차 손실 (means squared error, MSE)

- 회귀 문제에서 널리 사용됨

  회귀문제에 사용될 수 있는 다른 손실 함수

  (1) 평균 절댓값 오차 (Mean absolute error, MAE)

  (2) 평균 제곱근 오차 (Root mean squared error, RMSE)

- 예측값과 실제값의 차이를 제곱하여 평균한 값 (모두 실숫값으로 계산)
- MSE가 크다는 것은 평균 사이에 차이가 크다는 뜻 / MSE가 작다는 것은 데이터와 평균사이의 차이가 작다는 뜻

데이터 셋

회귀분석 - boston 데이터셋

CRIM	per capita crime rate by town	마을 별 1 인당 범죄율
ZN	proportion of residential land zoned for lots over 25,000 sq.ft.	25,000 평방 피트 이상의 부지에 구역화 된 주거용 토지의 비율.
INDUS	proportion of non-retail business acres per town	도시 당 비 소매 사업 에이커의 비율
CHAS	Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)	Charles River 더미 변수 (유도가 강 경계면 = 1, 그렇지 않으면 0)
NOX	nitric oxides concentration (parts per 10 million)	산화 질소 농도 (1,000 만분 율)
RM	average number of rooms per dwelling	주거 당 평균 방 수
AGE	proportion of owner-occupied units built prior to 1940	1940 년 이전에 지어진 소유주 소유 유닛의 비율
DIS	weighted distances to five Boston employment centres	보스턴 고용 센터 5 곳까지의 가중 거리
RAD	index of accessibility to radial highways	방사형 고속도로 접근성 지수
TAX	full-value property-tax rate per $10,000	$ 10,000 당 전체 가치 재산 세율
PTRATIO	pupil-teacher ratio by town	도시 별 학생-교사 비율
B	1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town	1000 (Bk-0.63) ^ 2 여기서 Bk는 도시 별 흑인 비율입니다.
LSTAT	% lower status of the population	인구의 낮은 지위 %
MEDV	Median value of owner-occupied homes in $1000's	소유주가 거주하는 주택의 중간 가치 ($ 1000)

참조 : https://scikit-learn.org/stable/datasets/index.html#toy-datasets

보스턴 주택 가격 데이터 – 범죄율, 주택당 방 개수 등 14개의 변수들이 집값에 미치는 영향을 나타낸 데이터

다중분류 - iris 데이터셋



꽃잎(petal)과 꽃받침(sepal)의 사이즈를 기반으로 꽃의 종류를 예측

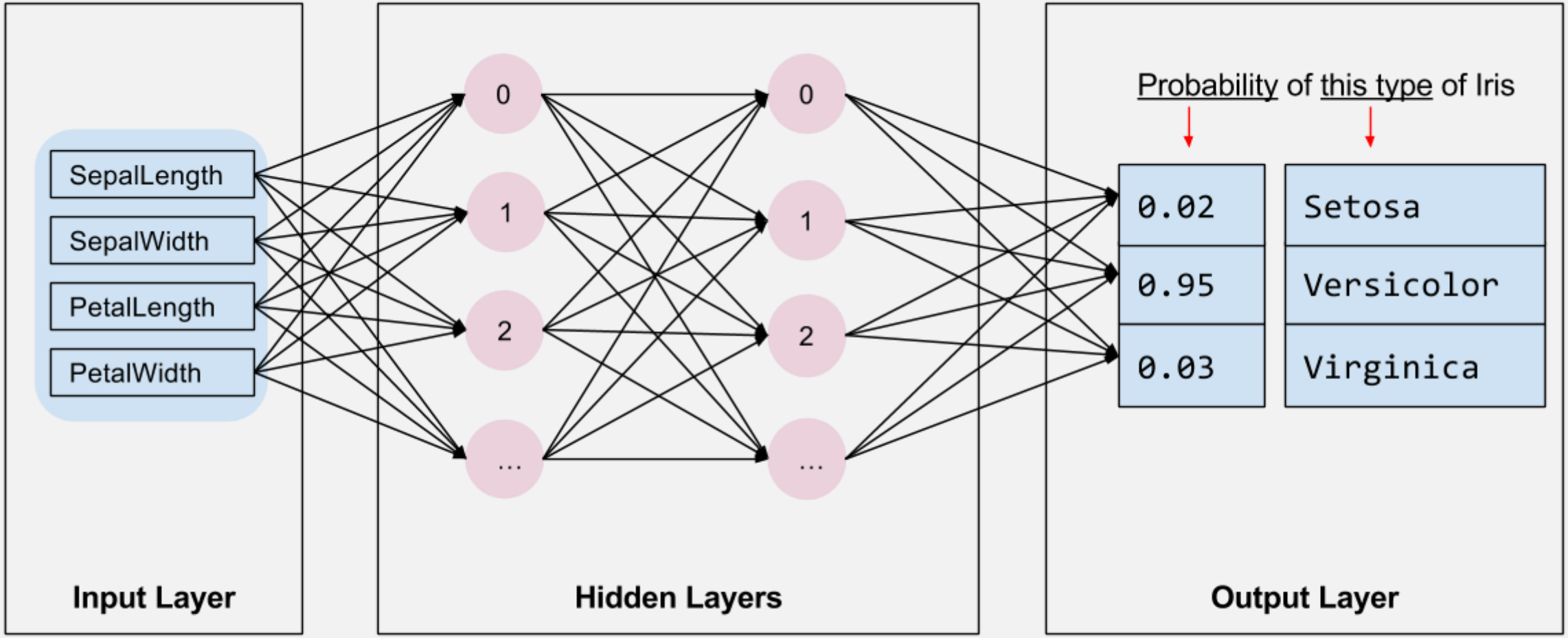
Feature(특성, 컬럼, 열)

- sepal length in cm
- sepal width in cm
- petal length in cm
- petal width in cm

Class (분류)

- Iris-Setosa
- Iris-Versicolour
- Iris-Virginica

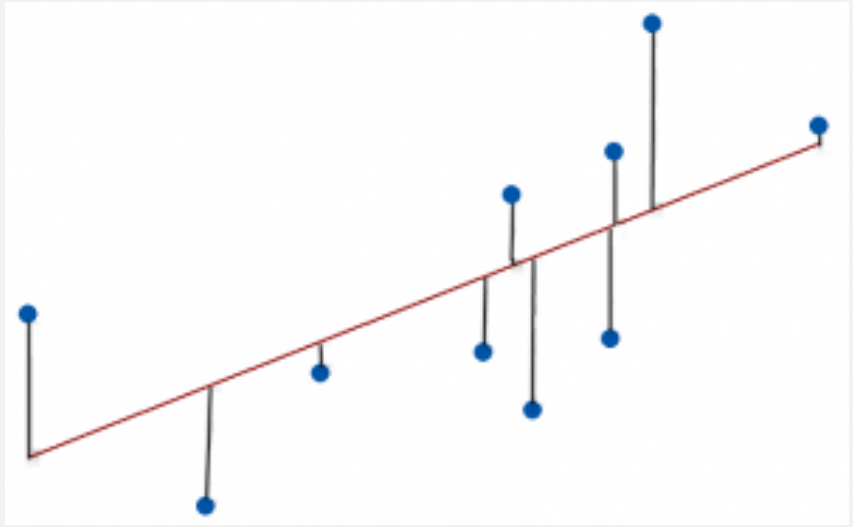
다중분류 - iris 데이터셋



실 습

1. train_test_split
2. matplotlib – scatter (산점도)
3. R2 score (결정계수)
4. validation split
5. 회귀분류
6. 이진분류
7. 다중분류

결정계수 R2 Score



출처 https://go-hard.tistory.com/125

Figure 6. 적합도 평가

- R-squared는 선형 회귀 모델에 대한 적합도 측정값
- 선형 회귀 모델을 훈련한 후, 모델이 데이터에 얼마나 적합한지 확인하는 통계 방법 중 하나
- r2 score는 0과 1사이의 값을 가지며 1에 가까울수록 선형회귀 모델이 데이터에 대하여 높은 연관성을 가지고 있다고 해석함

Day02. 인공지능 Study

1. 인공지능 개념 정리 – 머신러닝, 딥러닝
2. 퍼셉트론 (Perceptron)
3. 다층 퍼셉트론 (Multi-Layer Perceptron: MLP)
4. 옵티마이저 (Optimizer)
5. 학습률 (learning rate)
6. 경사하강법 (Gradient Descent)
7. 손실함수 (Loss Function)
8. 활성화 함수 (Activation Function) – Sigmoid, ReLU, Softmax

9. 회귀분석

10. 결정계수 R2 score

11. 분류분석

12. 원 핫 인코딩 (One Hot Encoding)

13. 난수값 (random_state)

수고하셨습니다.