REPORT

지도학습을 이용한

유방암 판별 모델 설계

-

학과: 전자공학과

학번: 2016142020

이름: 박재석

**INDEX**

1. 이론

* 지도학습
* 분류
* K fold Cross Validation

1. 설계

* Cost Function
* Training Model

1. K fold Cross Validation 결과 및 검증
2. 파이썬 코드
3. 고찰
4. **이론**

이 프로젝트는 인공지능설계실습 과목의 중간고사 대체 프로젝트로써, 지도학습을 이용한 유방암 판별모델을 설계한다. 지도학습은 크게 회귀모델과 분류모델 두 모델로 나뉘며 이 프로젝트에서는 분류모델을 사용하여 설계한다. 분류모델을 설계하는데 사용되는 모델은 로지스틱 회귀를 사용하며, 비용함수로는 CEE(cross entropy error)를 사용한다. 또한 모델을 검증하는 방법에는 K fold Cross Validation(K 겹 교차검증)을 사용한다. 이제 앞서 설명한 이론 및 설계 모델에 대해서 설명하도록 하겠다.

* **지도학습**

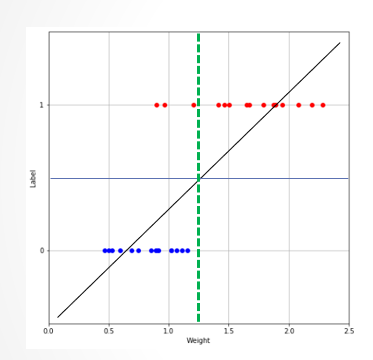
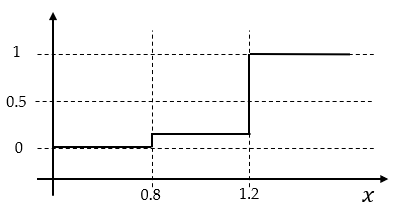
지도 학습이란? 말 그대로 모델의 input data와 output data를 둘 다 이용하여 모델을 설계하는 것을 말한다. 따라서 Training Data는 input, output data를 둘 다 포함한다. 따라서 보다 쉽게 설계된 모델이 원하는 성능에 일치하게 동작하는지 확인하는 것이 보다 쉽다.

지도학습에는 크게 분류와 회귀 두가지 모델로 나뉜다. 분류란? 어떠한 데이터에 대해 일정한 값으로 나뉘는 것을 말한다. 예를 들어서 사진에 사과가 있는지 없는지를 판별하는 모델에 사과 사진을 넣게 되면, output data로 사과가 있는지, 없는지를 분류하여 알려준다. 따라서 연속적인 값보다. 이산적인 값을 출력해 내는데 사용한다.

회귀는 분류와 반대로 데이터를 나누는 것이 아니라. 연속적인 값으로 표현하는 것이다. 예를 들어서 나이에 따른 키의 값을 추정하거나 하는 곳에 사용된다. 따라서 이번 프로젝트에는 회귀가 아닌 분류모델로 학습모델을 설계할 것이다.

* **분류**

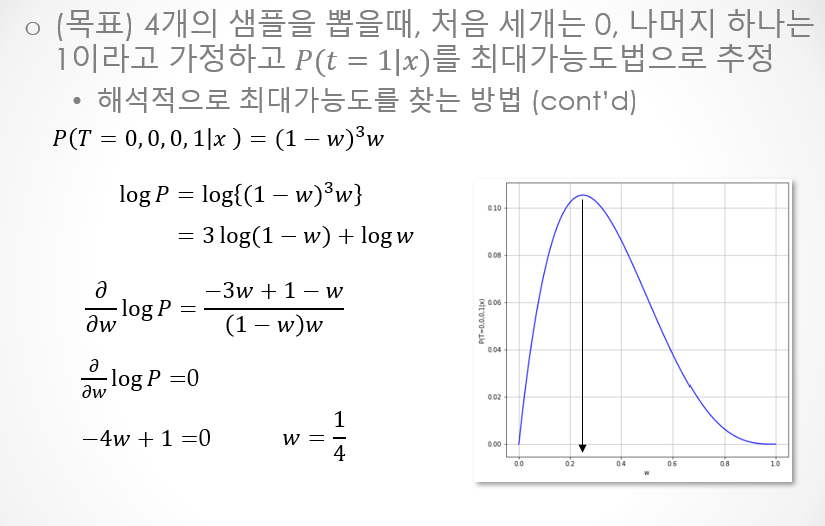
앞서 설명하였듯이 분류는 input data를 이산적인 output data로 출력하게 하는 모델이다. 따라서 Decision boundary를 정해주어야 하는데, Decision boundary는 Training Data를 이용하여 학습되어진 모델로 결정을 하는 방법이 있고, 확률개념을 도입하여 Decision boundary를 결정하는 방법이 있다. 두 방법의 차이는 아래의 그림과 같다.

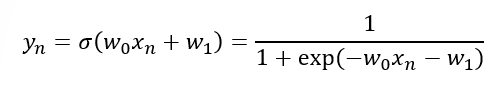
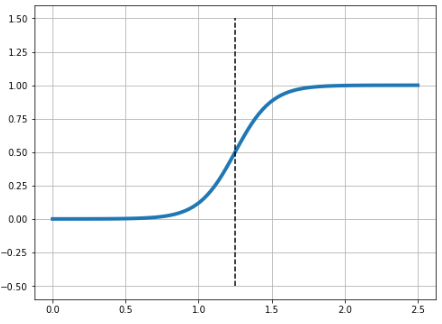
위 그림을 보면 왼쪽이 훈련된 선형 회귀 모델을 이용하여 분류를 한 예이고, 오른쪽은 확률을 이용하여 input data에 따른 output data가 1일 때의 확률을 가시화 한 것이다.

왼쪽그림을 보면 알 수 있다시피, 선형 회귀 모델은 평균제곱오차를 최소화 하기 위하여 1로 말할 수 있는 데이터의 영향을 많이 받게 된다. 따라서 Decision boundary가 1데이터 쪽으로 끌려가는 경향이 발생한다. 하지만d 확률을 이용할 경우 주어진 input data에 대해 1일 확률이 0.5이상이 되는 지점을 Decision boundary로 정하면된다. 그러므로 input data에 대해 output data가 1 또는 0일 확률에 대해서 파악을 하여야 한다.

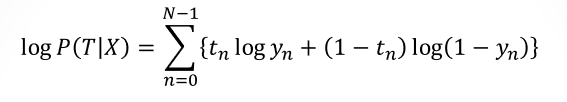
파악하는 방법에는 Maximum likelihood method이 있다 likelihood란 가능도로써 말 그대로 어떤 값이 관측되었을 때 이것이 어떤 확률 분포에서 왔을 지에 대한 확률이다. 간단히 하자면 확률의 확률이다. 따라서 우리는 likelihood가 Maximum이 되는값을 찾으면 된다. 아래는 Maximum likelihood를 찾는 예시이다.

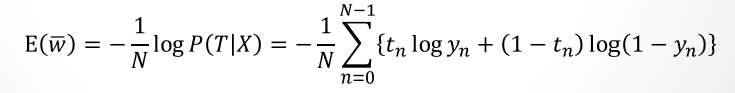


Maximum likely method를 이용하여 likelihood가 최대가 되는 w값, 즉 관측치가 나올 가능성이 제일높은 1의 확률을 알 수 있다. 하지만 실제 데이터에서는 확률이 일정하지 않다. 따라서 일정하지 않은 확률을 반영하는 모델이 필요하다. 우리는 학습모델을 로지스틱회귀 모델을 사용한다. 로지스틱 회귀모델은 아래 그림과 같다.

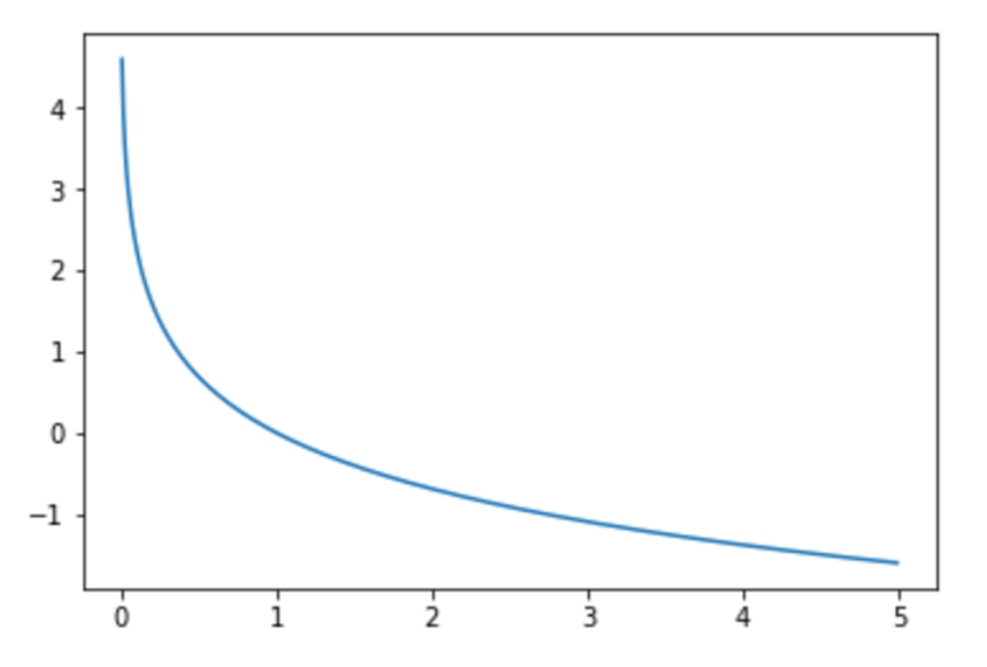
또한 우리는 비용함수를 정의하여야 한다. 비용함수란 예측한 값과 실제값까지와의 차이이다. 우리는 이 차이가 가장작은 값을 가지는 모델을 설계 하여야 한다. 하지만 분류에서는 가능도가 제일 높은 모델을 사용하면 되는데, Likelihood 함수에 로그를 씌우게 되면 예측값이 정확할 수록 0이 되게 되고, 예측값이 부정확할 수록 무한대로 발산한다.



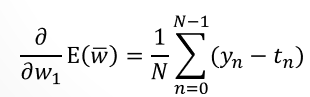
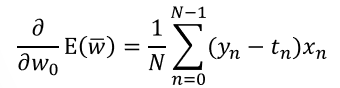


위의 첫번째 식이 log likelihood이고, 아래식이 교차 엔트로피 오차, 즉 비용함수이다.

여기서 주목해야될 점은 왜 교차 엔트로피 오차에는 마이너스가 붙는데 붙는이유는 로그함수의 특성때문이다. X가 0일 때 Y값이 0이 아닌 1이 되기 때문이다. 아래의 그림을 참조 바란다.



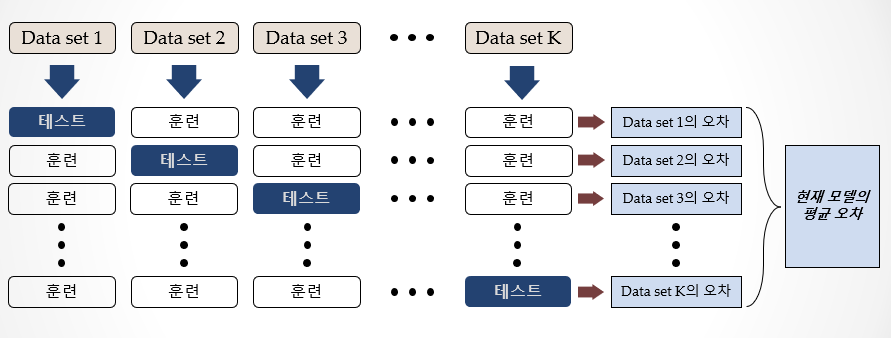
이제 교차엔트로피 오차를 이용하여 오차값이 제일 작을 때의 w값을 구해야 하는데, 경사하강법을 이용하여 구하게 된다. 따라서 교차 엔트로피 오차를 w로 편미분을 한다.



위의 식과 같이 교차엔트로피 오차의 기울기 식이 나오게 되고, 위의 식을 이용하여 경사하강법을 진행하게 되면 최적의 w값이 나오게 된다.

* **K fold Cross Validation**

모델을 설계함에 있어서 검증과정은 필히 존재하여야 한다. 따라서 우리가 설계한 모델에 대해서도 검증을 진행해야하는데, 우리는 K fold Cross Validation 방법을 이용하여 검증한다. 이 방법은 Data의 일부분을 Test data로 두고 나머지 Data를 Training data로 사용한다.



위의 방법대로 검증을 진행하게 되고 K개 만큼의 가중치벡터와 CEE(교차엔트로피오차)값을 얻을 수 있다. 이를 통해서 현재 모델의 평균 오차를 알 수 있고, 설계한 모델이 잘 설계 되었는지 알 수 있다. 그리고 오차를 구한 값 중에서 오차가 가장 적은 가중치벡터를 최종적으로 선택하게 된다.

1. **설계**

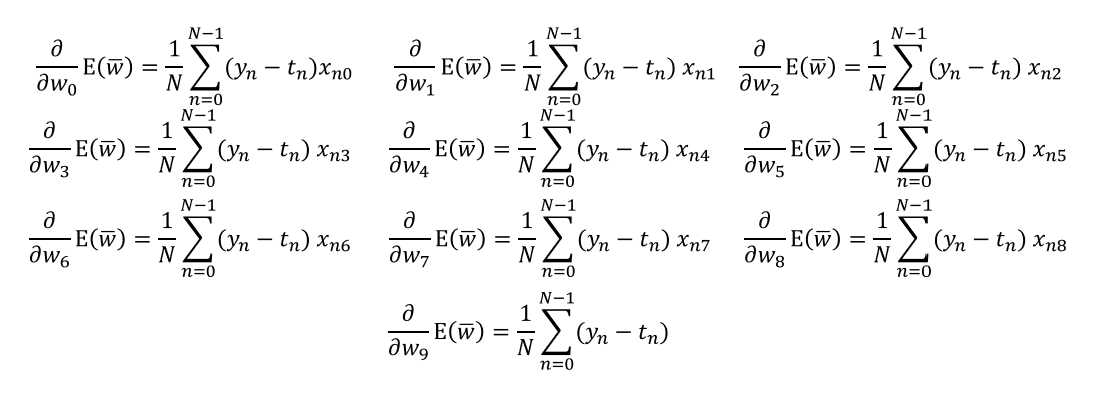
위에서 설계의 순서에 대해서 언급하였다. 처음에는 회귀, 분류중에서 선택을 하고, 선택을 한 후에는 Cost fucntion에 대해서 설계를 한다. 그다음에는, 훈련 모델을 선택하여야 한다. 그 뒤에는 검증 방법을 통하여 설계된 모델이 원하는 성능을 내는지에 대해서 확인을 한다. 따라서 Cost function, 즉 비용함수를 어떻게 설계 하였는지에 대해서 서술하겠다.

* **Cost function**

비용함수는 CEE(Cross Entropy Error)를 선택 하였다. 따라서 수식은 아래와 같다.

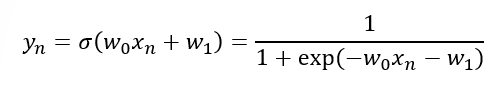
위의 교차 엔트로피 오차는 1일 확률이 y일 때의 교차엔트로피 오차이다. 따라서 위의 Cost function을 사용하기 위해서는 training data의 output을 반전 시켜주면 된다. 반전을 시켜주게 되면은 0일 확률이 y일 때의 교차 엔트로피 오차가 되게 된다.

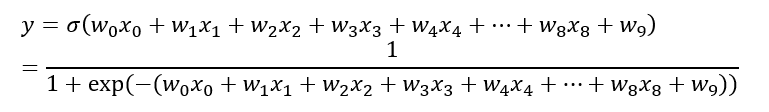
비용함수를 정의 하였으면 비용함수의 값이 제일 작을 때의 가중치값(w)를 구하여야 한다. 가중치값을 구하기 위해서는 교차엔트로피의 오차도 구해야한다. 따라서 우리는 입력값이 8개인 모델의 교차엔트로피 오차의 기울기를 구하여야한다. 식은 아래와 같다.



* **Training Model**

Cost function을 정의 하였으니 학습을 위한 Training Model을 정의해야 한다. Training 모델은 말 그대로 훈련을 하는데 사용되는 모델이다. 이 모델에 맞춰서 input data에 대한 output data가 정해지게 된다. Training Model은 로지스틱 회귀 모델을 사용하며, 수식은 Input data가 9개 이므로 아래와 같다.

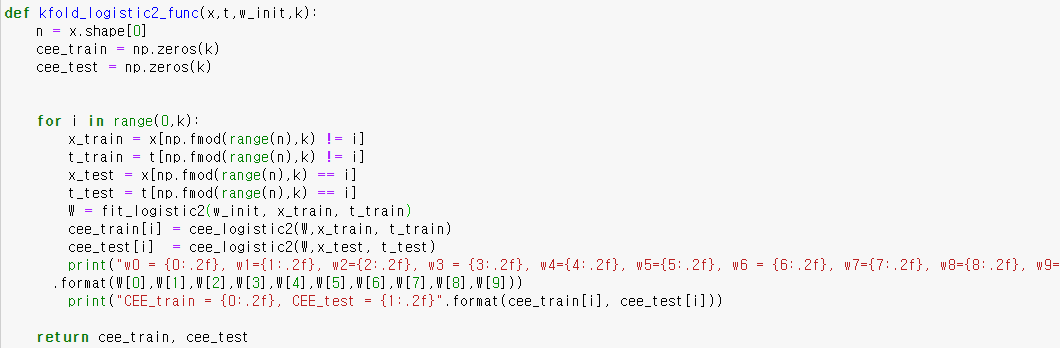


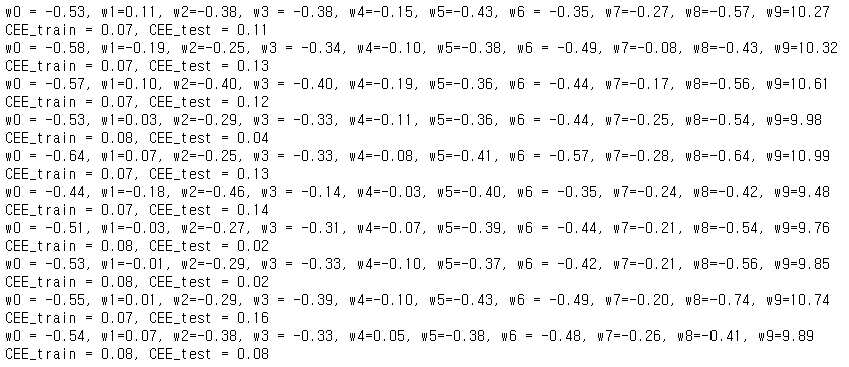


첫 번째 수식은 로지스틱 회귀 모델의 기본 수식이며, 두번째 수식이 input이 9개일 때의 로지스틱 회귀모델이 되겠다. 위와 같은 식으로 훈련데이터를 학습시켜 비용함수를 통하여 가중치 값을 estimation할 수 있게 된다.

1. **K ford Cross Validation 결과 및 검증**

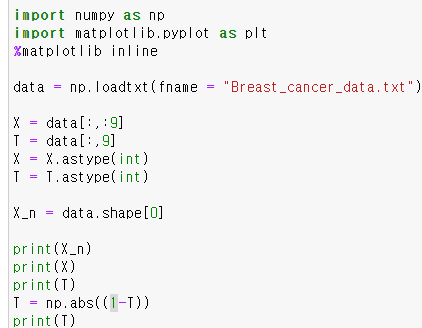
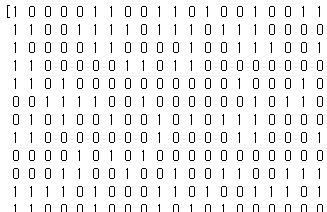
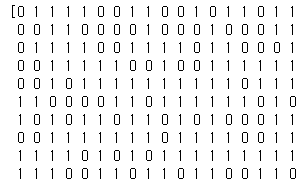
비용함수와 학습모델을 설계하였으니 검증과정을 통해 설계모델을 검증하여야 한다. 우리는 K fold Cross Validation을 통하여 검증을 진행한다.



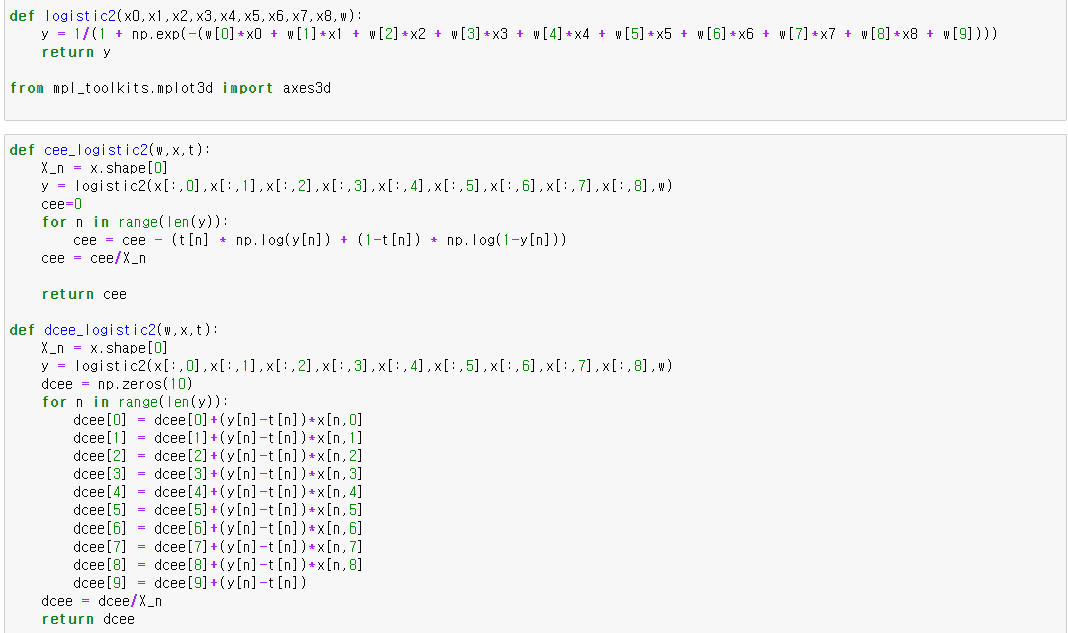


위의 함수를 이용하여 K fold Vaildation을 진행하였으며 아래는 K를 10으로 하여 10번의 학습을 진행하였다. 학습되어진 모델의 CEE를 확인하였을 때, 모델의 설계가 제대로 이루어 진 것을 확인할 수 있다. 또한 테스트 데이터의 CEE가 0.02인 2가지 가중치가 가장좋은 가중치 값으로 알 수 있다.

1. **파이썬 코드**

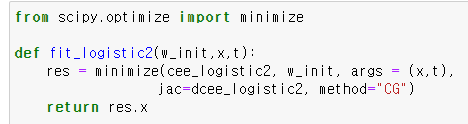
  

위의 코드는 앞서 설명했던 Training Data들의 결과값을 뒤집는 코드이다. 결과값이 0과 1이 반전된 것을 확인할 수 있다.



위의 코드들은 순서대로 학습모델은 로지스틱 회귀함수부터 교차엔트로피오차, 교차엔트로피 오차의 기울기이다. 수식은 위에서 설명을 하였으니 따로 설명하지 않겠다. 수식을 파이썬 코드로 바꾸기만 하였다.

w값은 경사하강법을 이용하여 찾아 내었는데, w값을 찾아내는 함수는 아래와 같다.



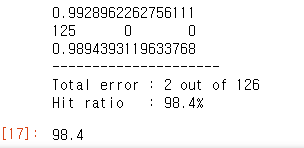
위의 코드를 보았을 때, 비용함수로 교차 엔트로피 오차를 이용하며, Jacobian은 교차엔트로피 오차의 편미분 값인 dcee\_logistic2를 넣어주었다. 또한 Optimizer로는 CG를 이용하였다.

K fold Validation 코드는 위에서 첨부하였으니 따로 첨부하지 않겠다.

1. **고찰**

설계를 완료하고 검증을 통하여 최적의 가중치값을 찾아낸 결과 w0 = -0.53, w1=-0.01, w2=-0.29, w3 = -0.33, w4=-0.10, w5=-0.37, w6 = -0.42, w7=-0.21, w8=-0.56, w9=9.85, 값을 얻을 수 있었고, 교차엔트로피 오차의 훈련값과 테스트 값은 CEE\_train = 0.08, CEE\_test = 0.02 와 같이 나오게 되었다. 그리고 설계완료된 모델과 가중치를 이용하여 모델을 테스트해본 결과 Hit ratio 즉 적중치는 126개의 테스트데이터에서 124개가 적중하는 모습을 보여줬다. 아래의 사진이 그 결과이다.





검증이후 적중치도 보다 높게 나오는 것을통하여 설계 및 학습이 보다 정확하게 이루어 졌다고 말할 수 있겠다. 적중하지못한 2개가 발생하는 이유는 내생각에는 결정경계의 근처에 있는 값들 중에서 약간의 오차에도 반응하는 데이터여서 오차가 발생한 것 같다. 이번 프로젝트를 통하여 보다 미숙하였던 이론에 대해서 다시한번 공부하는 계기가 되었다. 또한 비지도학습이 아닌 지도학습의 분류를 통해서도 보다 높은 성능의 시스템을 설계 및 응용할 수 있다는 것에 많은 놀라움이 있었다.