김서연

(경희대 산업수학 2020-1)

선형 회귀와 ANN을 이용한 COVID-19 모델 예측

음식의 지방 섭취량과 비만률이 코로나 발병률에 영향을 끼칠까?

분석 순서 : 데이터 수집 > 데이터 탐색과 준비 > 데이터로 모델 훈련 > 모델 성능 평가 > 모델 성능 개선

1) 데이터 수집

독립변수

- ✓ Animal fats(동물성 지방)
- ✓ Animal Products(동물성 식품)
- ✓ Vegetal products(식물성 식품)
- ✓ Seafood(해산물)
- ✓ Spices(양념 및 향신료)
- ✓ Obesity(비만률)

종속변수

- ✓ Confirmed(코로나 발병률)

2) 데이터 탐색

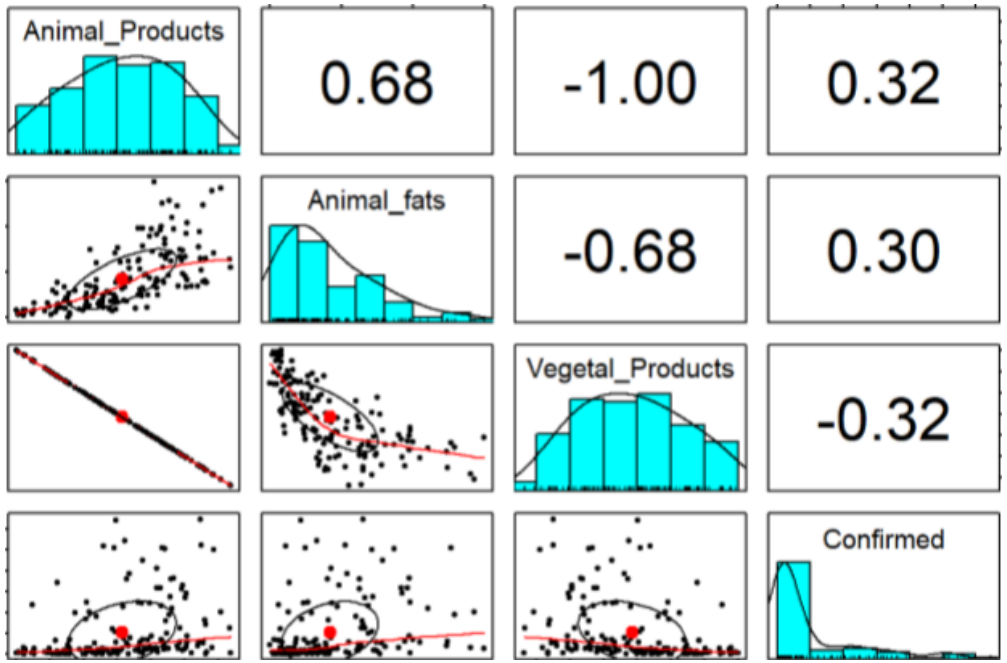
상관행렬

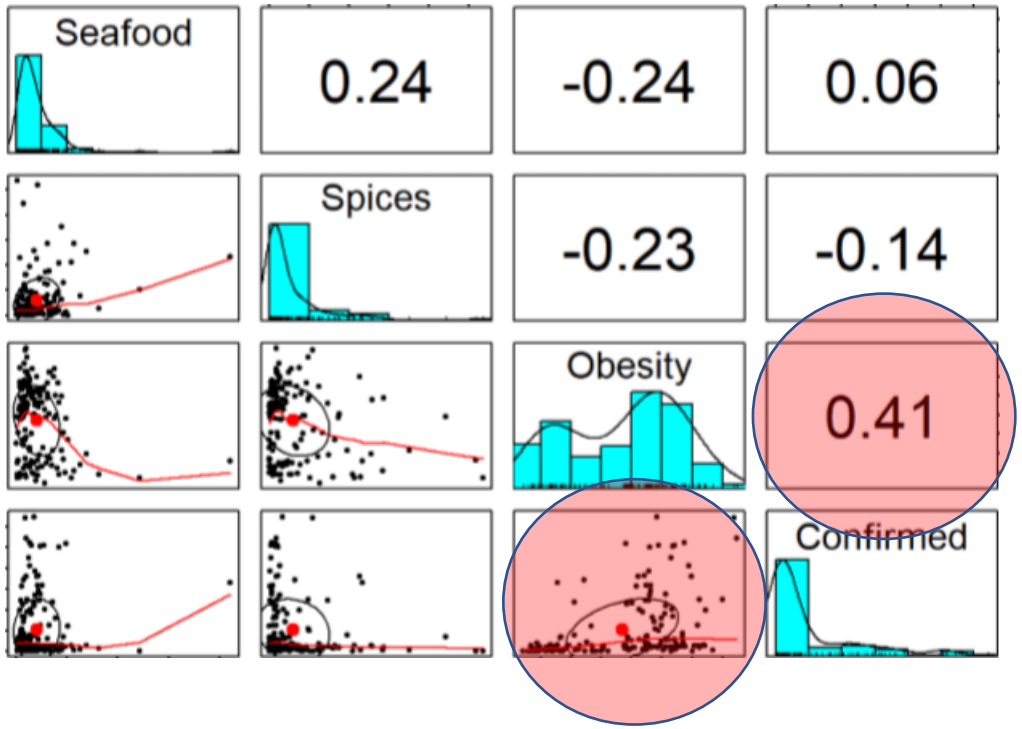
```
##                  Animal_Products Animal_fats Vegetal_Products      Seafood
## Animal_Products      1.000000000   0.6837163     -0.999999855  0.004194152
## Animal_fats          0.683716299   1.0000000     -0.683703920 -0.102092750
## Vegetal_Products    -0.999999855  -0.6837039      1.000000000 -0.004245142
## Seafood              0.004194152  -0.1020928     -0.004245142  1.000000000
## Spices              -0.187724220  -0.2069006      0.187638449  0.238484775
## Obesity              0.475672511   0.4307724     -0.475660492 -0.240079954
## Confirmed            0.316090708   0.2970380     -0.316124801  0.056206350
##                         Spices     Obesity   Confirmed
## Animal_Products   -0.1877242   0.4756725  0.31609071
## Animal_fats       -0.2069006   0.4307724  0.29703798
## Vegetal_Products   0.1876384  -0.4756605 -0.31612480
## Seafood            0.2384848  -0.2400800  0.05620635
## Spices             1.0000000  -0.2312019 -0.13708207
## Obesity           -0.2312019   1.0000000  0.41250423
## Confirmed         -0.1370821   0.4125042  1.00000000
```

2) 데이터 탐색과 준비

산포도 행렬

수치 특징 간의 관계를 시각화





상관관계 타원형

늘어질수록 강한 상관관계, 원에 가까울수록 약한 상관관계 의미

양의 상관관계
Obesity > Animal_Products > Animal_fats > Seafood

음의 상관관계
Vegetal_Products > Spices

3) 데이터로 모델 훈련

```
Call:
lm(formula = Confirmed ~ Animal_Products + Animal_fats + Vegetal_Products +
    Seafood + Spices + Obesity, data = corona_new)

Coefficients:
     (Intercept)    Animal_Products       Animal_fats  Vegetal_Products
       120.634935        -2.412844         0.004699        -2.414004
         Seafood            Spices           Obesity
        0.027316        -0.023159         0.005932
```

베타 계수(Coefficient)

각 특징에 대해 하나가 증가하고 다른 값은 모두 고정됐다고 가정할 때 코로나 발병률 추정치의 증가량

- ✓ **Animal Products** 가 증가할 때, 코로나 발병률 약 **2.4%** 감소
- ✓ **Vegetal Products** 가 증가할 때, 코로나 발병률 약 **2.4%** 감소
- ✓ **Seafood** 가 증가할 때, 코로나 발병률 약 **0.027%** 증가
- ✓ **Spices** 가 증가할 때, 코로나 발병률 약 **0.023%** 감소

4) 모델 성능 평가

```
Call:
lm(formula = Confirmed ~ Animal_Products + Animal_fats + Vegetal_Products +
    Seafood + Spices + Obesity, data = corona_new)

Residuals:
     Min       1Q   Median       3Q      Max
-0.21533 -0.08824 -0.02389  0.03411  0.49508

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     120.634935 125.394844   0.962   0.3375
Animal_Products  -2.412844   2.507811  -0.962   0.3375
Animal_fats       0.004699   0.004482   1.049   0.2961
Vegetal_Products -2.414004   2.507876  -0.963   0.3373
Seafood           0.027316   0.012655   2.158   0.0325 *
Spices           -0.023159   0.024954  -0.928   0.3548
Obesity           0.005932   0.001388   4.274 3.36e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

잔차(Residuals)

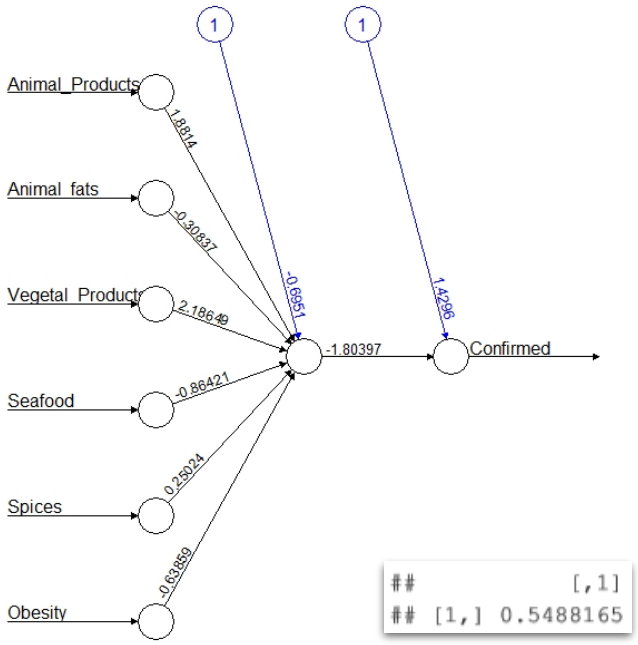
- ✓ 최대 코로나 확진률을 약 **0.495%** 높게 예측
- ✓ 최소 코로나 확진률을 약 **0.215%** 낮게 예측

※ 코로나에 걸렸는데 걸리지 않았다고 할 확률인 **0.215%**가 중요

추정계수 (Coefficients)

- ✓ P-값은 추정된 계수가 실제 관계가 없는 0일 확률 추정치로, 낮을 수록 결과가 우연히 발생할 확률이 낮음을 의미
- ✓ 통계적으로 유의한 변수: **Seafood, Obesity**

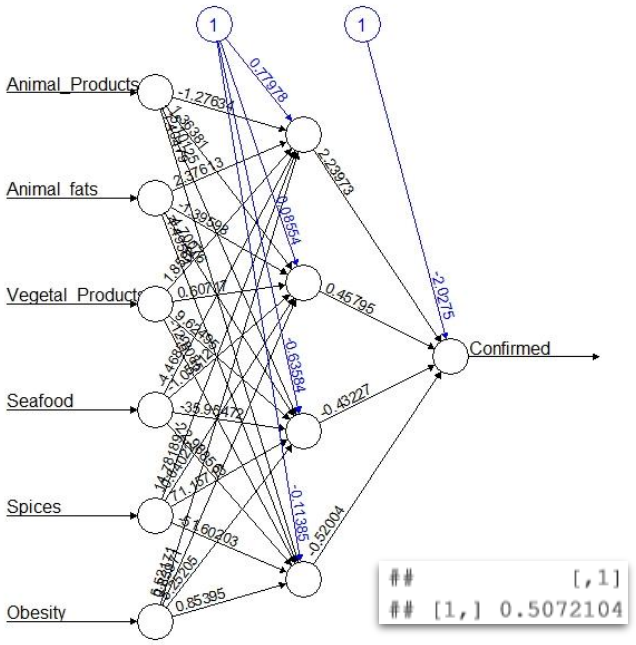
3) 데이터로 모델 훈련
4) 모델 성능평가



Error: 2.541065 Steps: 61

은닉 노드가 1개인 신경망 객체

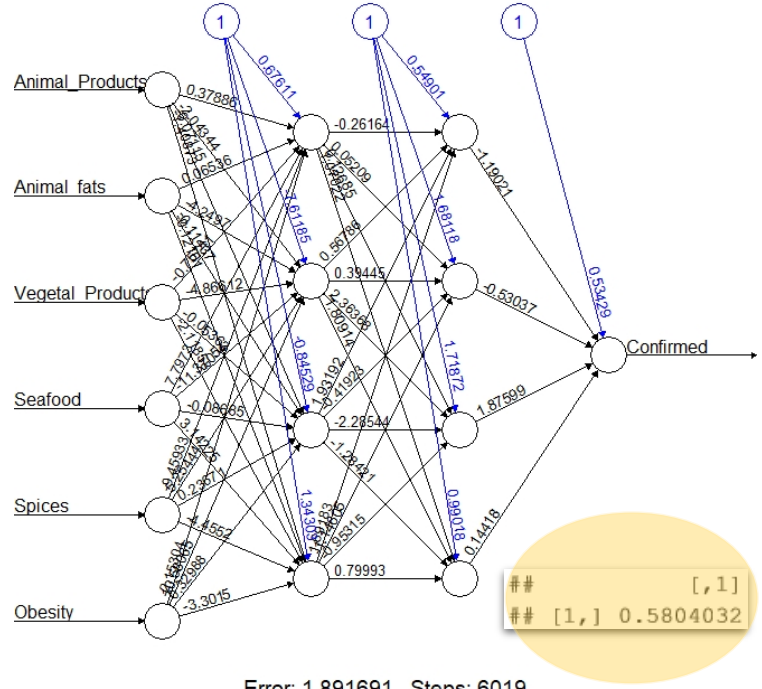
5-1) 모델 성능 개선



Error: 1.919702 Steps: 776

은닉 노드의 개수를 4개로 늘린 신경망 객체

5-2) 모델 성능 개선



Error: 1.891691 Steps: 6019

각각 4개의 노드를 가진 2 계층 네트워크를 형성하고, 소프트플러스 활성함수를 사용한 신경망 객체