Supervised Learning : Supervised Learning이란, 학습 데이터를 줘서 모델을 학습 시킨 다음 예측 결과를 출력하는 것이다. supervised learning에는 선형 회귀, 로지스틱 회귀 등이 있다.

선형 회귀와 로지스틱 회귀를 구분하는 대표적인 방법은 모델의 예측 결과가 무한하게 많은 것인가, 2개 중 하나로 예측 하는 것인가로 나뉘게 된다.

선형 회귀의 대표적인 예는 집의 평수를 판단 기준으로 집값을 예측하는 것이다. 모델이 예측하는 집값은 0원에서 몇 억까지 셀 수 없이 무수히 많다.

로지스틱 회귀의 대표적인 예는 종양의 크기를 판단 기준으로 악성인지 양성인지 예측하는 것이다. 예측 범위는 0 ~ 1이다.

Unsupervised Learning : Unsupervised Learning을 쉽게 이해하고자 하면 유튜브 알고리즘을 생각하면 된다. 모델이 스스로 유사한 데이터를 찾아서 그룹화를 시키는 것이다. 이것을 전문 용어로 클러스터라고 한다.

Supervised Learning : linear regression

집의 크기로 집 가격을 예측하는 모델을 만든다고 생각을 해보자. 그러면 우리는 어떤 모델을 써야 할까? 정답은 선형회귀 모델이다. 예측하는 집값의 범위는 0원 ~ 몇 십억으로 매우 광범위하기 때문이다.

이 선형 회귀를 공부하기 앞서 표기법에 대해 알고 있어야 한다.

* 모델을 학습하는 데이터(training data) : x
* 예측된 데이터(출력 변수) : y

그러면 이제 Data Table을 그려보자.

|  |  |  |
| --- | --- | --- |
| Data Table | | |
| m(Data number) | (Size) | (Price) |
| 1 | 2107 | 1343 |
| 2 | 2342 | 2433 |
| 3 | 3432 | 45332 |
| 4 | 3424 | 453345 |
| 5 | 4324 | 53453 |
| 6 | 4532 | 453423 |

이렇게 Data Table에는 6개의 데이터가 존재한다. 그러면 3번째 학습 데이터를 어떻게 표기할까?

* Training data의 총 개수 : m
* I번째 입력 데이터 :
* I 번째 출력 데이터 :

그러면 위의 Data Table은 m=6, 3432이라는 것을 알 수 있다.

하지만 우리는 집값을 예측하는 모델을 학습시킬 때 집의 크기만을 고려하지 않는다. 집의 크기, 위치, 방의 개수, 집 연식 등등 여러가지 조건을 고려해서 집값을 예측하게 된다. 이렇게 Training data가 많아지게 되면 어떻게 표기를 해야 할까?

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Data Table | | | | | |
| m | (size) | (room) | (age) | (floor) | y(price) |
| 1 | 2107 | 3 | 23 | 1 | 1343 |
| 2 | 2342 | 4 | 34 | 2 | 2433 |
| 3 | 3432 | 4 | 43 | 3 | 45332 |
| 4 | 3424 | 2 | 23 | 3 | 453345 |
| 5 | 4324 | 4 | 34 | 2 | 53453 |
| 6 | 4532 | 4 | 12 | 1 | 453423 |

* 학습 데이터 열의 개수 : j
* 학습 데이터 열의 위치 :

이렇게 4개의 학습 데이터를 이용해서 예측한 집값을 어떻게 표기할까?

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Data Table | | | | | | |
| m | (size) | (room) | (age) | (floor) | y(price) |  |
| 1 | 2107 | 3 | 23 | 1 | 1343 | 1000 |
| 2 | 2342 | 4 | 34 | 2 | 2433 | 2400 |
| 3 | 3432 | 4 | 43 | 3 | 45332 | 45000 |
| 4 | 3424 | 2 | 23 | 3 | 453345 | 440000 |
| 5 | 4324 | 4 | 34 | 2 | 53453 | 50000 |
| 6 | 4532 | 4 | 12 | 1 | 453423 | 450000 |

이제 표기법을 정리해보자.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Data Table | | | | | | |
| m | (size) | (room) | (age) | (floor) | y(price) |  |
| 1 | 2107 | 3 | 23 | 1 | 1343 | 1000 |
| 2 | 2342 | 4 | 34 | 2 | 2433 | 2400 |
| 3 | 3432 | 4 | 43 | 3 | 45332 | 45000 |
| 4 | 3424 | 2 | 23 | 3 | 453345 | 440000 |
| 5 | 4324 | 4 | 34 | 2 | 53453 | 50000 |
| 6 | 4532 | 4 | 12 | 1 | 453423 | 450000 |

이런 Data Table이 있다면 우리는 다음과 같이 생각할 수 있다.

Training data의 개수(m)는 6개이고, 우리가 모델에게 학습시킬 데이터 종류는 4개가 있다. 그리고 실제 집값(y) 데이터가 있고, 우리가 예측한 집값()이 있다.

그러면 room열의 3번째 데이터는 어떻게 표시할까? 이렇게 표시한다.

이제 표기법을 알았으니 간단한 Data Table을 이용해서 선형회귀 모델에 대해 알아보자.

|  |  |  |  |
| --- | --- | --- | --- |
| Data Table | | | |
| m(Data number) | (Size) | (Price) |  |
| 1 | 2107 | 1343 | 1300 |
| 2 | 2342 | 2433 | 2400 |
| … | … | … | … |
| 47 | 4532 | 453423 | 45000 |

x-y 그래프에 가로축은 x(Size), 세로축은 y(Price)로 그리면 다음과 같은 그래프가 나오게 된다.

텍스트, 스크린샷, 라인, 폰트이(가) 표시된 사진

자동 생성된 설명

예측하는 모델은 직선이 아니라 여러가지 비 선형적인 모델이 될 수도 있다. 하지만 선형회귀 모델의 가장 기본적인 식은 다음과 같이 쓸 수 있다.

기본적인 선형회귀 모델 식을 보면 알 수 있지만 직선의 방정식과 같은 모양을 가지고 있다. 따라서 w는 모델의 기울기에 영향을 주고, b는 모델의 위치에 영향을 준다.

텍스트, 라인, 폰트, 도표이(가) 표시된 사진

자동 생성된 설명

Cost Function : 단순히 말하자면 모델의 오류이다. 모델이 예측하는 값과 실제 관측된 값 간의 차이를 나타내는 측정 지표이기 때문이다. 조금 더 구체적으로 설명하자면 모델이 예측을 잘못하면 비용 함수 값이 증가하며, 모델이 예측을 정확하게 할수록 비용 함수 값이 감소한다.

우리의 목표는 이 비용 함수 값을 최소화하는 방향으로 모델을 조정하고 학습시키는 것이다. 비용 함수를 최소화하는 것은 모델이 실제 데이터에 더 잘 적합 되도록 하여 예측의 정확성을 향상시키는 것을 의미한다.

선형회귀 모델에서 Cost Function은 MSE(Mean Square error) 평균 제곱 오차를 사용한다.

라인, 폰트, 도표, 그래프이(가) 표시된 사진

자동 생성된 설명

여기서 2를 나눠주는 이유는 나중에 나오지만 Cost Function을 미분할 때 상수를 없애기 위함이다.

이 식의 가장 중요한 점은 라는 것을 기억해야 한다.

이제 선형회귀 모델에 대한 정리를 해보자.

Cost Function은 모델의 오류라고 설명을 했다. 그러면 우리의 궁국적인 목표는 Cost Function의 값을 최소화 하는 것이다. 이것을 와 의 연관성을 보며 알아보자. 그러기 위해서는 우선 식을 간단하게 b=0 설정하고 시작하자.

텍스트, 도표, 라인, 폰트이(가) 표시된 사진

자동 생성된 설명

간단히 설명하면 y가 1, 2, 3이고 w=1, b=0일 때 Cost Function을 구한 것이다.

텍스트, 라인, 도표, 폰트이(가) 표시된 사진

자동 생성된 설명

이것은 같은 조건에서 w=0.5, b=0일 때 Cost Function을 구한 것이다.

텍스트, 도표, 라인, 폰트이(가) 표시된 사진

자동 생성된 설명

W 값을 바꿔가며 Cost Function을 구하는 과정을 계속 반복하면 다음과 같은 결과를 얻을 수 있다.

라인, 폰트, 도표이(가) 표시된 사진

자동 생성된 설명

이 결과가 말해주는 것은 가 낮은 값을 가질수록 선형모델이 예측한 값()과 실제 데이터 값(y)가 유사하다는 것을 나타낸다.

그렇다면 가 가장 낮은 값을 가지기 위해 w와 b 값을 어떻게 찾아야 할까?

정답은 경사 하강법이다. Gradient Descent(경사 하강법)의 궁국적인 목표는 가 가장 낮은 값을 가지도록 w와 b 값을 찾는 것이다.

우선 Gradient Descent의 식을 살펴보자. 자세히 보면 = 기울기, = 상수 값을 가지는 선형 방정식이다. 그러면 그래프로 살펴보자.

텍스트, 도표, 라인, 폰트이(가) 표시된 사진

자동 생성된 설명

가 양수가 나오면 w – 양수 즉, 기존 w 값보다 작은 크기의 w가 만들어진다.

가 음수가 나오면 w – 음수 즉, 기존 w 값보다 큰 크기의 w가 만들어진다.

이렇게 가장 낮은 위치에 존재하는 w값을 찾아 다니는 것이 경사 하강법이다.

이 조건을 대입해서 w의 Gradient Descent 식이 어떻게 나오는지 자세히 살펴보자.

식을 간단하게 하기 위해 m의 값을 1로 두고 식을 풀겠다.

이렇게 나오게 된다. 최종적으로 w, b의 Gradient Descent식을 정리하면 다음과 같다.

그리고 는 학습 속도라고 생각하면 된다. 하지만 값이 너무 작으면 학습을 너무 천천히 하는 단점이 생기고, 너무 크면 over shoot이라는 최솟값을 찾아가는 도중에 지그재그로 움직이거나, 최솟값을 지나쳐서 반대 방향으로 이동하는 경우가 발생한다.

이제 집값 예측 모델의 학습 조건을 여러 개 가져와 보자.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Data Table | | | | | | |
| m | (size) | (room) | (age) | (floor) | y(price) |  |
| 1 | 2107 | 3 | 23 | 1 | 1343 | 1000 |
| 2 | 2342 | 4 | 34 | 2 | 2433 | 2400 |
| 3 | 3432 | 4 | 43 | 3 | 45332 | 45000 |
| 4 | 3424 | 2 | 23 | 3 | 453345 | 440000 |
| 5 | 4324 | 4 | 34 | 2 | 53453 | 50000 |
| 6 | 4532 | 4 | 12 | 1 | 453423 | 450000 |

이때 데이터 스케일링을 하지 않고 모델을 만들게 되면 스케일이 큰 특성이 모델에 더 큰 영향을 미치므로, 작은 스케일의 특성은 모델의 예측에 거의 영향을 미치지 못하는 경우가 발생한다. 그래서 정규화 과정을 해 줘야 한다.

스케일 방법은 3가지가 있다. 그중 첫번째는 단순히 최댓값으로 나눠주는 것이다.

만약 과 를 사용한다고 생각해보자.

*텍스트, 폰트, 라인, 도표이(가) 표시된 사진

자동 생성된 설명*

두번째는 평균 정규화다. 이때, 는 평균을 의미한다.

텍스트, 폰트, 라인, 스크린샷이(가) 표시된 사진

자동 생성된 설명

세번째는 z-score 정규화다. 이때 는 표준편차를 의미한다.

텍스트, 폰트, 라인, 도표이(가) 표시된 사진

자동 생성된 설명

정규화는 값이 큰 거 같으면 바로 해주는 것이 좋다.

Classification : 의사결정 트리, 로지스틱 회귀 등등 다양하지만 우리는 로지스틱 회귀를 사용할 것이다. 예측되는 값으로는 0 또는 1으로 출력된다. 이때 보통 질문에 대한 참값을 1로, 거짓을 0으로 설정한다. 예를 들어 ‘스팸 메일인가?’ 라는 질문에는 1 = ’스팸 메일’, 0 = ’스팸 메일이 아님’

Logistic Regression : sigmoid function 즉, logistic function을 사용하기 때문이다. logistic function을 사용하는 이유는 그래프가 다음과 같이 나오기 때문이다.

라인, 도표, 그래프, 평행이(가) 표시된 사진

자동 생성된 설명

이 식을 모델식으로 바꾸면 다음과 같다.

모델식을 알았으니 Cost Function을 구해보자. 분류 모델에서는 MSE를 사용하지 않는다. MSE만 사용할 경우 다음과 같이 출력되기 때문이다.

라인, 그래프, 폰트, 도표이(가) 표시된 사진

자동 생성된 설명

이런 그래프가 출력되는 이유는 MSE는 예측 값과 실제 값 간의 제곱 오차를 사용하여 손실을 계산하는데, 분류 문제에서는 클래스의 확률 분포를 고려해야 하며, MSE를 사용하면 클래스 확률을 잘 반영하지 못하는 문제가 발생하기 때문이다.

그래서 기존 MSE에 Loss Function을 더해줘야 한다.

이때, Loss 값은 y의 실제 값이 1일때, 0일때 2가지로 나뉜다.

이 식이 출력되는 이유는 의 그래프를 알면 이해하기 쉽다.

*텍스트, 라인, 도표, 폰트이(가) 표시된 사진

자동 생성된 설명*

가로축은 y값, 세로축은 Loss Function 값이다. 이때, y의 실제 값이 1이고 예측한 가 0과 가깝다면 Loss Function은 무한히 커지게 된다. 대신 가 1과 가깝게 나오면 Loss Function은 0과 가깝게 출력된다.

텍스트, 라인, 폰트, 도표이(가) 표시된 사진

자동 생성된 설명

이때, y의 실제 값이 0이고 예측한 가 1과 가깝다면 Loss Function은 무한히 커지게 된다. 대신 가 0과 가깝게 나오면 Loss Function은 0과 가깝게 출력된다.

이것을 정리하면 다음과 같은 식이 나온다.

이러면

정리하면 분류 모델의 Cost Function 식은 다음과 같다.

구해진 Cost Function을 바탕으로 분류 모델의 Gradient Descent 식을 구하면 다음과 같다.

식은 선형 회귀와 같지만 가 다르다.

Linear Regression의 Overfitting : 선형회귀 모델에서 오버 피팅이 발생하면 모델은 학습 데이터에 대해 높은 정확도를 보일 수 있으나, 실제 데이터에서는 제대로 일반화되지 못해 성능이 저하된다.

라인, 도표, 그래프, 경사이(가) 표시된 사진

자동 생성된 설명

해결 방법은 대표적으로 4가지가 있다. 데이터 확보, 모델 단순화, 규제화, 교차검증이다.

* 데이터 확보 : 더 많은 학습 데이터를 확보하여 모델이 실제 패턴을 더 잘 학습하도록 한다.
* 모델 단순화 : 모델의 복잡도를 줄이는 방향으로 조절하여 불필요한 파라미터나 특성을 제거하거나 줄인다.
* 정규화 : L1 또는 L2 규제를 적용하여 모델의 가중치를 제한하고 일반화 한다.
* 교차 검증 : 교차 검증을 통해 모델의 성능을 평가하고 오버 피팅을 감지한다.

이중 우리가 할 것은 정규화다. 정규화는 단순히 말하면 가중치(w)의 값을 줄여서 고차 함수의 존재감을 작게 만드는 것이다.

정규화를 하기 전 Cost Function의 성질에 대해 이해를 하자. Cost Function은 컴퓨터가 무조건 작은 값을 만들기 위해 노력을 한다. 그러면 Cost Function에 MSE와 같이 가중치(w)를 넣어주면 가중치를 줄여준다는 것이 기본적인 생각이다.

이 식을 쉽게 설명하자면 Cost Function은 0에 가까워지도록 노력하는데 가 커져서 의 존재감이 MSE보다 커지게 되면 Cost Function은 MSE를 무시하고 를 줄이는데 노력할 것이다.

하지만 이 정규화의 방법은 모든 가중치(w)의 값이 줄어들게 된다. 이제 이렇게 구해진 를 이용해서 경사 하강을 진행하면 다음과 같다.

이때 식을 간단하게 보기 위해 m=1이라고 설정하자.

이때, , , m = 5일때, 다음과 같이 나오게 된다.

즉, 가중치 는 점점 작아지는 값이 된다.

Classification에서 정규화를 하면 만들어지는 식은 선형회귀와 같지만 중요한 점은 가 다르다는 것이다.

Cost Function은 다음과 같다.

수정된 Cost Function으로 경사 하강법을 하면 다음과 같다.

하지만 이것도 마찬가지로 식은 똑같지만 가 다르다.