[L1 Regularization 과 L2 Regularization]

1. L1 Regularization 과 L2 Regularization은 Overfitting(과적합) 을 막기 위해 사용된다.
2. Norm

Norm 은 벡터의 크기(혹은 길이)를 측정하는 방법(혹은 함수)를 뜻함(두 벡터 사이의 거리를 측정하는 방법)

$$\|\mathbf{x}\|_p := \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}$$

* p 는 Norm 의 차수를 의미. p = 1 이면 L1 Norm 이고, P = 2 이면 L2 Norm
* n은 해당 벡터의 원소 수

1) L1 Norm

$$d_1(\mathbf{p}, \mathbf{q}) = \|\mathbf{p} - \mathbf{q}\|_1 = \sum_{i=1}^{n} |p_i - q_i|, \text{ where } (\mathbf{p}, \mathbf{q}) \text{ are vectors } \mathbf{p} = (p_1, p_2, \ldots, p_n) \text{ and } \mathbf{q} = (q_1, q_2, \ldots, q_n)$$

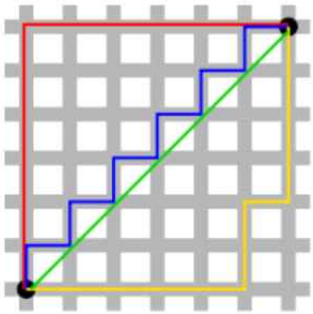
- L1 Norm 은 벡터 p, q 의 각 원소들의 차이의 절대값의 합을 뜻함
예를 들어 벡터 p =(3, 1, -3), q = (5, 0, 7) 이라면 p, q의 L1 Norm 은
|3-5| + |1-0| + |-3 -7| = 2 + 1 + 10 = 13 이 된다.

2) L2 Norm

$$\|\boldsymbol{x}\|_2 := \sqrt{x_1^2 + \cdots + x_n^2}.$$

L2 Norm 은 벡터 p, q 의 유클리디안 거리(직선 거리)이다. 여기서 q 가 원점이라면 벡터 p, q의 L2 Norm 은 벡터 p 의 원점으로부터의 직선거리라고 할 수 있다.
- 위식 p = (x_1, x_2, ... , x_n), q = (0, 0, ... , 0) 라고 할 수 있다.

3) L1 Norm 과 L2 Norm 의 차이



- 검정색 두 점사이의 L1 Norm 은 빨간색, 파란색, 노란색 선으로 표현 될 수 있고, L2 Norm 은 오직 초록색 선으로만 표현될 수 있다. L1 Norm 은 여러가지 path 를 가지지만 L2 Norm 은 Unique shortest path 를 가진다.
- 예를 들어 p = (1, 0), q = (0, 0) 일 때 L1 Norm = 1, L2 Norm = 1 로 값은 같지만 여전히 Unique shortest path 라고 할 수 있다.

3. L1 Loss

$$L = \sum_{i=1}^{n} |y_i - f(x_i)|$$

y_i 는 실제 값을, f(x_i)는 예측치를 의미합니다. 실제 값과 예측치 사이의 차이(오차) 값의 절대값을 구하고 그 오차들의 합을 L1 Loss 라고 한다. 이를 Least absolute deviations(LAD), Least absolute Errors(LAE), Least absolute value(LAV), Least absolute residual(LAR), Sum of absolute deviations 라고 부른다.

4. L2 Loss

$$L = \sum_{i=1}^{n} (y_i - f(x_i))^2$$

L2 Loss 는 오차의 제곱의 합으로 정의된다. 이를 Least squares error(LSE) 라고 부른다.

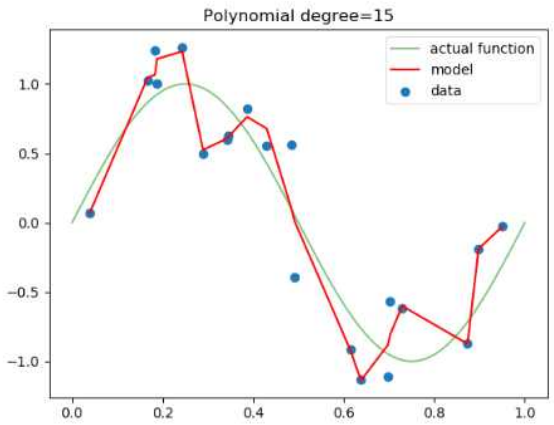
5. L1 Loss, L2 Loss 의 차이

L2 Loss 는 직관적으로 오차의 제곱을 더하기 때문에 Outlier 에 더 큰 영향을 받는다. "L1 Loss 가 L2 Loss 에 비해 Outlier 에 대하여 더 Robust(덜 민감 혹은 둔감) 하다."
Outlier 가 적당히 무시되길 원한다면 L1 Loss 를 사용하고, Outlier(이상치,이상치 때문에 평균값이 무너진다) 의 등장에 신경 써야 하는 경우라면 L2 Loss 를 사용하는 것이 좋다.
L1 Loss 는 0인 지점에서 미분이 불가능하다는 단점 또한 가지고 있다.

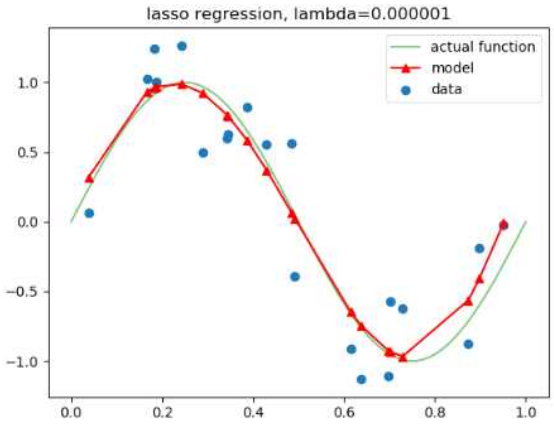
6. Regularization

- '정규화' 또는 '일반화' 라고 한다.

모델 복잡도에 대한 패널티로 정규화는 Overfitting 을 예방하고 Generalization(일반화) 성능을 높이는데 도움을 준다. Regularization 방법으로는 L1 Regularization, L2 Regularization, Dropout, Early stopping 등이 있다.



model 을 쉽게 만드는 방법은 단순하게 cost function 값이 작아지는 방향으로만 진행하는 것이다. 이럴 경우 특정 가중치가 너무 큰 값을 가지기 때문에 model 의 일반화 성능이 떨어지게 될 것이다. 위 그래프에서 actual function 이 target function 이라고 했을 때, model 이 overfitting 된 것을 알 수 있다.



Regularization 은 특정 가중치가 너무 과도하게 커지지 않도록 하여 모델을 위 그래프처럼 만들어 준다.

1) L1 Regularization

$$Cost = \frac{1}{n}\sum_{i=1}^{n}\{L(y_i, \hat{y_i}) + \frac{\lambda}{2}|w|\}$$

$L(y_i, \hat{y_i})$: 기존의 Cost function

- 분수로 붙는 1/n 이나 1/2 가 달라지는 경우가 있는데 L1 Regularization 의 개념에서 가장 중요한 것은 cost function 에 가중치의 절대값을 더해준다는 것이기 때문에 1/n 이나 1/2 가 달라지는 경우는 연구의 case 에 따라 다르다(이는 L2 Regularization 또한 같다)
- 기존의 cost function 에 가중치의 크기가 포함되면서 가중치가 너무 크지 않은 방향으로 학습 되도록 한다. 이때 λ 는 learning rate(학습률) 같은 상수로 0에 가까울수록 정규화의 효과는 없어진다.
- L1 Regularization 을 사용하는 Regression model 을 Least Absolute Shrinkage and Selection Operater(Lasso) Regression 이라고 부른다.

2) L2 Regularization

$$Cost = \frac{1}{n}\sum_{i=1}^{n}\{L(y_i, \hat{y_i}) + \frac{\lambda}{2}|w|^2\}$$

기존의 cost function 에 가중치의 제곱을 포함하여 더함으로써 L1 Regularization 과 마찬가지로 가중치가 너무 크지 않은 방향으로 학습되게 된다. 이를 Weight decay 라고도 한다.
- L2 Regularization 을 사용하는 Regression model 을 Ridge Regression 이라고 부릅니다.

3) L1 Regularization, L2 Regularization 의 차이와 선택 기준
Regularization에서 가중치 w 가 작아지도록 학습한 다는 것은 결국 Local noise 에 영향을 덜 받도록 하겠다는 것이며 이는 Outlier 의 영향을 더 적게 받도록 하겠다는 것이다.

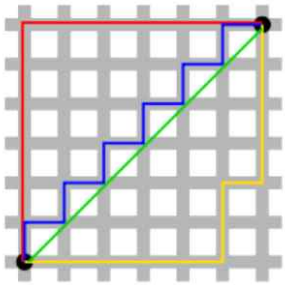
벡터 a 와 b에 대해서 L1 Norm 과 L2 Norm 을 계산)

$$a = (0.3, -0.3, 0.4)$$
$$b = (0.5, -0.5, 0)$$

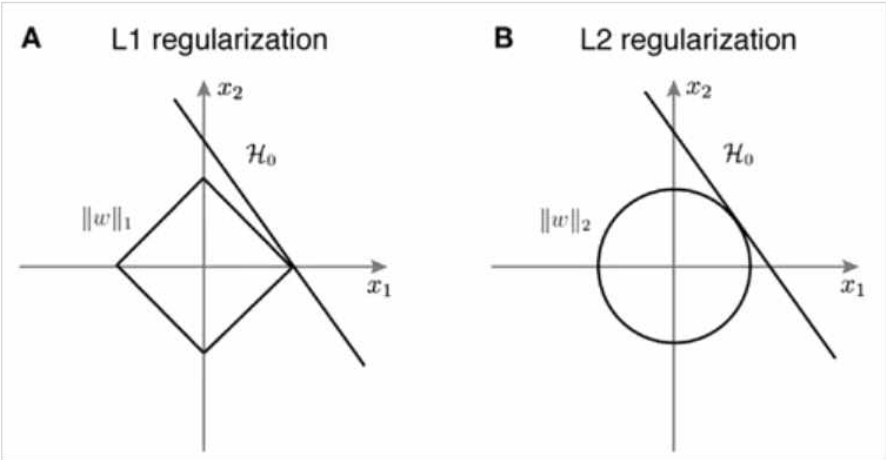
$$||a||_1 = |0.3| + |-0.3| + |0.4| = 1$$
$$||b||_1 = |0.5| + |-0.5| + |0| = 1$$

$$||a||_2 = \sqrt{0.3^2 + (-0.3^2) + 0.4^2} = 0.583095$$
$$||b||_2 = \sqrt{0.5^2 + (-0.5^2) + 0^2} = 0.707107$$

L2 Norm 은 각각의 벡터에 대해 항상 Unique 한 값을 내지만, L1 Norm 은 경우에 따라 특정 Feature(벡터의 요소) 없이도 같은 값을 낼 수 있다.



L1 Norm 은 파란색 선 대신 빨간색 선을 사용하여 특정 Feature 를 0으로 처리하는 것이 가능하다. 다시 말하자면 L1 Norm 은 Feature selection 이 가능하고 이런 특징이 L1 Regularization에 동일하게 적용 될 수 있는 것이다. 이러한 특징 때문에 L1 은 Sparse model(coding,희소모델, 차원/전체 공간에 비해 데이터가 있는 공간이 매우 협소한 데이터를 의미)에 적합하다. L1 Norm의 이러한 특징 때문에 convex optimization(어떤 함수의 optimum point, 즉 그것이 최소이거나 혹은 최대인 지점을 찾는 과정) 에 유용하게 쓰인다.



단, L1 Regularization 의 경우 위 그림처럼 미분 불가능한 점이 있기 때문에 Gradient-base learning(딥러닝에서 Learning Algorithm에는 batch GD, SGD, Momentum SGD, AdaGrad, RMSprop, Adam 등이 있다) 에는 주의가 필요하다.