


ML_04. 최적화와 경사하강법

🕒 생성일	@2022년 6월 7일 오전 10:35
📁 유형	머신러닝/딥러닝
👤 작성자	 동훈 오

최적화(Optimization)

경사하강법을 얘기하기 이전에 최적화에 대한 이해가 우선되어야 한다. 왜냐하면 경사하강법은 ‘최적화 문제 → 컨벡스 최적화 → 딥러닝을 위한 최적화 전략’에 속하기 때문이다.

최적화는 ‘주어진 함수를 최대/최소화 시키는 값을 선택하는 문제’로 정의되며 수식으로 나타내면 다음과 같다.

$$\begin{aligned} \text{maximize}(\text{max}) : f_0(w) & \quad f_0 : \text{목적 함수} \\ \text{subject}(\text{s.t.}) : f_i(w) = b_i, i = 1, \dots, m & \quad f_i : \text{제한 함수} \end{aligned}$$

제한 함수의 조건을 만족하면서, 목적함수를 최대 / 최소화 시키는 최적의 파라미터 w 를 찾는 것이 목표이다.

효율적으로 풀 수 있는 몇 가지 최적화 문제들이 있다.

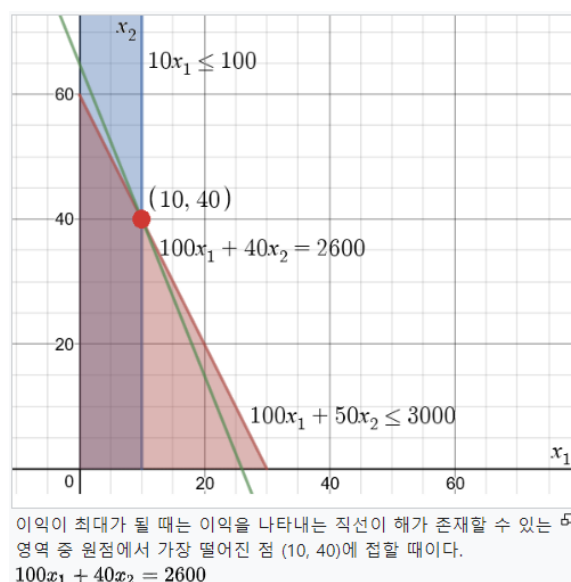
- **최소 자승법(least squares)**

비용 함수에 편미분을 통해서 minimum을 찾는 방식으로, 앞서 회귀와 분류에서 다루었다. 데이터에 적합하면서도 예측을 잘 하는 회귀 모델을 찾아가는 과정은 최적화 문제를 푸는 것이라고 생각할 수 있다.

- **선형 프로그래밍(linear programming)**

선형 계획법으로 알고 있는 문제이다. 가변 요소 사이에 일차 방정식이 성립하여 선형의 관계가 보장될 경우, 변화의 한계를 정할 때 사용하는 방법이다.

ex.

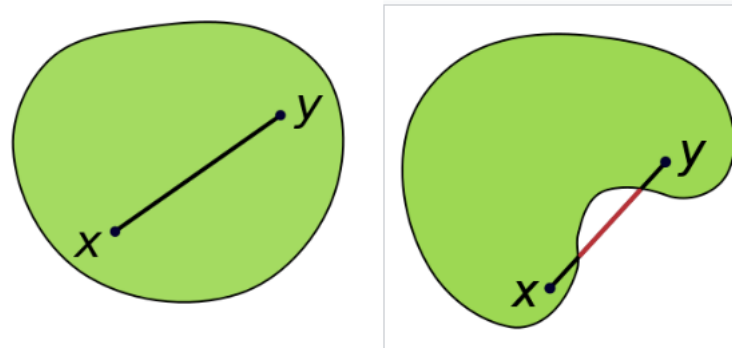


$$\begin{aligned} \text{목적: } \max_{\mathbf{x}} \quad & 100x_1 + 40x_2 \\ \text{조건: } \quad & 100x_1 + 50x_2 \leq 3000, \\ & 10x_1 \leq 100, \\ & x_1, x_2 \geq 0. \end{aligned}$$

- convex optimization

MSE 와 같은 볼록 함수를 convex function 이라고 한다.

convex set



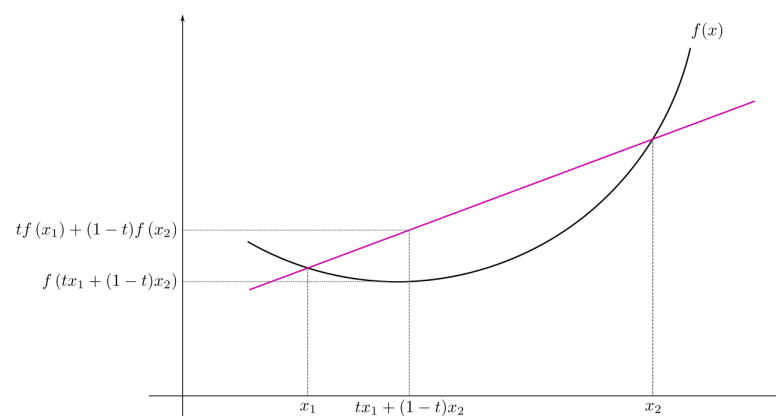
위 그림은 convex set 에 대한 직관적인 이해를 돕는다. 왼쪽은 convex set 에 해당하며, 오른쪽은 convex set에 해당하지 않는다. 특정 집합에 속한 두 점을 이은, line segment 의 도메인이 해당 집합 내에 위치하는 지가 중요하다.

convex function

$f: \mathbb{R}^n \rightarrow \mathbb{R}$ 일 때, f 의 도메인 구간이 convex set 일 때, 임의의 x, y 그리고 $0 \leq \alpha \leq 1$ 에 대해 다음의 식을 만족하는 함수를 말한다.

$$f(\alpha x + (1 - \alpha)y) \leq \alpha f(x) + (1 - \alpha)f(y)$$

- f 가 convex(볼록) 하면 f 는 concave(오목) 하다.



convex optimization

목적함수와 제한 함수가 모두 convex 함수 일 때의 최적화 문제를 컨벡스 최적화라고 한다. global optimum 을 찾는 것을 보장하며 많은 머신러닝 알고리즘은 컨벡스 최적화로 치환될 수 있다.

핸즈온 : (4.2) 경사하강법

경사 하강법(gradient descent ; GD) 의 기본 아이디어는 비용 함수를 최소화하기 위해 반복해서 파라미터를 조정해가는 것이다. 구체적으로, 파라미터 θ 를 임의의 값으로 시작해서 한 번에 조금씩 비용 함수가 감소되는 방향으로 진행하여 비용 함수값이 최소값에 수렴할 때 까지 점진적으로 향사시킨다. (다른 의미로 오차가 최소로 만들어진다는 것.)

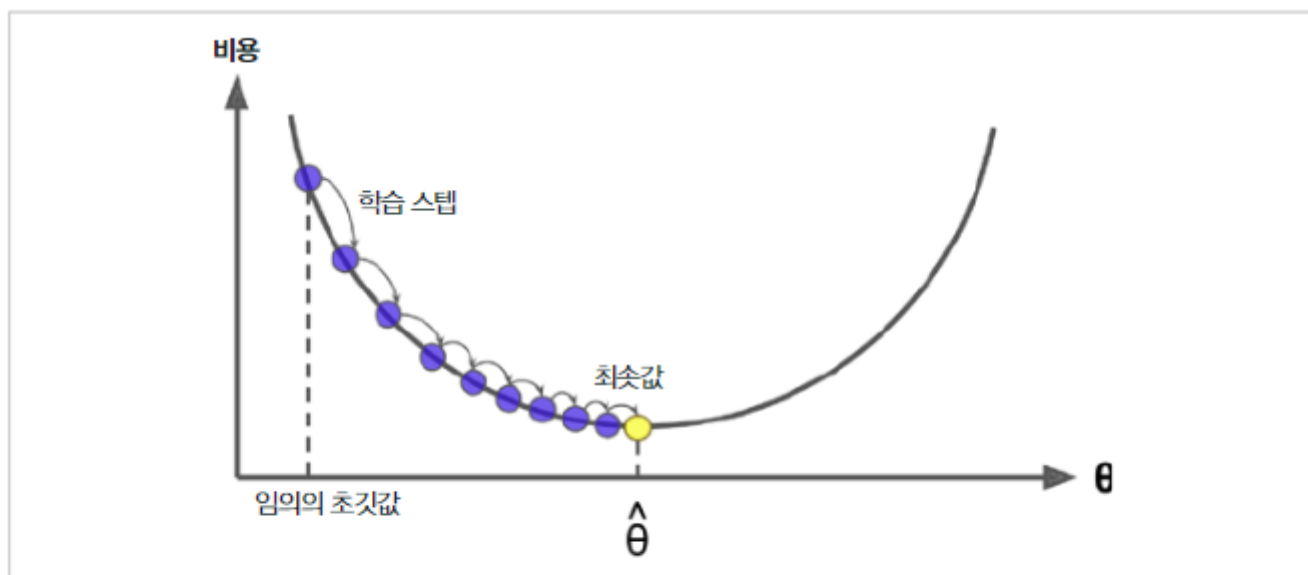
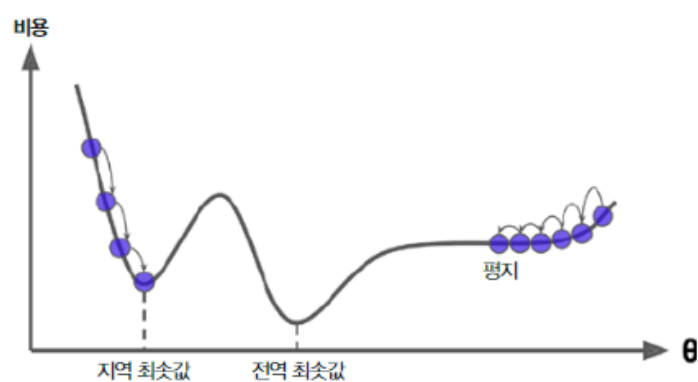


그림 4-3 이 경사 하강법 그림에서 모델 파라미터가 무작위하게 초기화된 후 반복적으로 수정되어 비용 함수를 최소화 합니다. 학습 스텝 크기는 비용 함수의 기울기에 비례합니다. 따라서 파라미터가 최솟값에 가까워질수록 스텝 크기가 점진적으로 줄어듭니다.

스텝의 크기는 학습률(learning rate) 하이퍼파라미터로 결정되는데, 학습률이 너무 작으면 알고리즘이 수렴하기 위해 반복을 많이 진행해야 한다.

경사하강법의 치명적인 문제점은 global minimum보다 local minimum 에 수렴할 가능성이 크다는 것이다.



위의 그래프는 비용 함수가 볼록 함수(convex function) 이지만, 특성들의 스케일이 매우 다르면 마치 평면과 같은 길쭉한 모양일 수 있다.

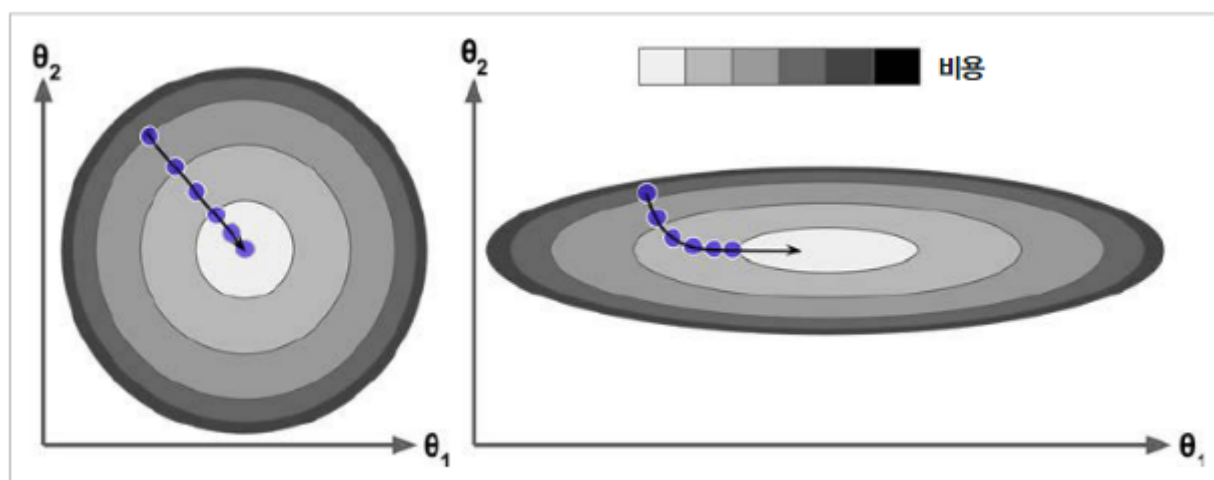


그림 4-7 특성 스케일을 적용한 경사 하강법(왼쪽)과 적용하지 않은 경사 하강법(오른쪽)

위의 그림은 모델 훈련이 비용 함수를 최소화하는 모델 파라미터의 조합을 찾는 일임을 설명한다. 모델의 파라미터 공간에서 작업은 수행되며, 파라미터가 많을 수록 공간의 차원은 커지게 된다.

배치 경사 하강법

경사 하강법을 구현하려면 각 모델 파라미터 θ_j 에 대해 비용 함수의 그레디언트를 계산해야 한다. 편미분을 이용할 수 있다. 예를 들어, 파라미터 θ_j 에 대한 MSE 비용 함수를 편미분 하면 다음과 같다.

$$\frac{\partial}{\partial \theta_j} \text{MSE}(\boldsymbol{\theta}) = \frac{2}{m} \sum_{i=1}^m (\boldsymbol{\theta}^T \mathbf{x}^{(i)} - y^{(i)}) x_j^{(i)}$$

여러 개의 파라미터에 대해 모두 편미분 하면 그레디언트 벡터의 집합으로 고려되고 이것은 행렬로 표현된다.

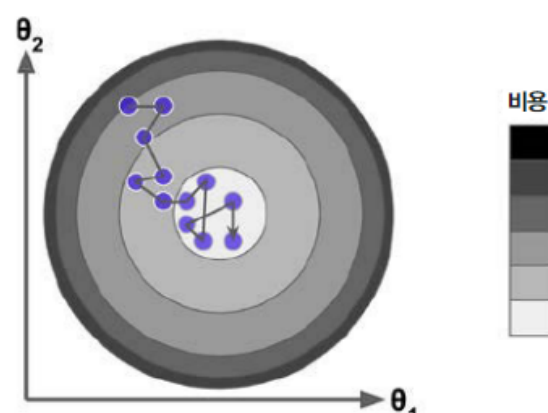
$$\nabla_{\boldsymbol{\theta}} \text{MSE}(\boldsymbol{\theta}) = \begin{pmatrix} \frac{\partial}{\partial \theta_0} \text{MSE}(\boldsymbol{\theta}) \\ \frac{\partial}{\partial \theta_1} \text{MSE}(\boldsymbol{\theta}) \\ \vdots \\ \frac{\partial}{\partial \theta_n} \text{MSE}(\boldsymbol{\theta}) \end{pmatrix} = \frac{2}{m} \mathbf{X}^T (\mathbf{X} \boldsymbol{\theta} - \mathbf{y})$$

이러한 편미분 계산이 매 경사 하강법 스텝에서 전체 훈련 세트 \mathbf{X} 에 대해 수행된다. 이 알고리즘을 배치 경사 하강법(batch gradient descent) 라고 한다. 매 스텝에서 훈련 데이터 전체를 사용해 업데이트를 수행하기에 훈련 세트가 크다면 시간이 오래 걸린다.

확률적 경사 하강법(Stochastic Gradient Descent ; SGD)

SGD 는 배치 경사 하강법과 달리, 매 스텝 한 개의 샘플을 무작위로 선택하고 그 하나의 샘플에 대한 그레디언트를 계산한다. 알고리즘이 확실히 빠르고, 매우 큰 훈련 세트도 훈련시킬 수 있다.

다만, 무작위성이 강하므로 배치 경사 하강법보다 불안정하다. 비용 함수가 최솟값에 다다를 때까지 부드럽게 감소하지 않고 요동치며 평균적으로 감소한다.



요동치며 최솟값을 찾아가는 것은 단점만 있는 것은 아니다. local minimum에 갇히지 않고 뛰어넘어 global minimum 을 찾을 수 있을 가능성이 오히려 커지게 된다.

SGD의 남아있는 불안정성을 해결하는 방법은 학습률을 점진적으로 감소시키는 것이다. 시작할 때는 학습률을 크게 하여 수렴을 빠르게 하고 지역 최솟값에서 쉽게 탈출할 수 있도록 한다. 이후, 학습률을 점차 줄여서 알고리즘이 global minimum에 도달하게 한다. 이러한 전략은 'Learning rate Schedule' 이라고 부른다.

Shuffle

SGD 를 사용할 때 훈련 샘플이 i.i.d. 를 만족해야 평균적으로 파라미터가 global minimum 을 향해 진행한다고 보장할 수 있다. 이렇게 만드는 간단한 방법은 훈련하는 동안 샘플을 섞는 것이다. 예를 들어 각 샘플을 랜덤하게 선택하거나 에포크를 시작할 때 훈련 세트를 섞어주는 것이다.

미니배치 경사 하강법 (mini-batch gradient descent)

미니배치 경사 하강법은 '미니배치' 라 부르는 임의의 작은 샘플 세트에 대해 그레디언트를 계산한다. 따라서 SGD 보다 덜 불안정하게 움직이지만, local minimum 에 빠질 가능성은 조금 더 크다.

세 가지 방식 비교

다음 그래프는 세 가지 경사 하강법 알고리즘이 훈련 과정 동안 파라미터 공간에서 움직인 경로이다.

