Optimization (1:30 ~ 2:50)

Goal 최적화 이론에 대한 전문적인 지식을 습득한다.

최적화 이론이 데이터분석에 있어 어느부분에서 쓰이고 있는지 이해한다.

---

0. 개관.
1. Gradient Descent Method.
2. Optimization in Machine Learning Algorithm.
3. Heuristic Algorithm.

---

0. 개요

Problems

* 발의 사이즈와 키 사이에 어떤 관계가 있을까.
* 날씨의 변화를 예측할 수 있을까.
* 자율주행 자동차를 위해 ~~차선인식 기술~~ 이미지 혹은 동영상 데이터에서 사람 / 도로 / 다른 차 등을 인식하게 할 수 없을까.
* 어떤 상품을 생산함에 있어서 이익을 최대로 하는 재료의 수를 구할 수 있을까.

⋮

수학적인 식을 세울 수 있는 경우

- 미분방정식
- 수리계획법 (Mathematical Programming)
- + α ...

수학적인 식을 세울 수 없는 경우

- Machine Learning
- Heuristic Algorithm
- ⋮

mathematical programming

Unconstrained Optimization.

$$\min_{x \in X} f(x)$$

Gradient Method

Newton's Method.

Conjugate ...

Constrained Optimization.

$$\min_{x \in X} f(x)$$

Subject to $x \geq 0$.

⋮

Penalty Method.

feasible direction Method.

Lagrangian Method.

Linear Programming — **Simplex method.**

minimize. $f(x) = x_1 + 3x_2 - x_3$

subject to $x_1 + x_2 + x_3 \geq 0$

$x_1 < 0$ ...

$x_2 > 0$

Non-Linear Programming

minimize $f(x) = e^{-x^2} x^2 + x$.

Subject to $x^2 + y^2 = 1$

$x \geq -0.5$.

⋮

1. Gradient Descent Method

1.1. chain rule.
1.2. ~~Parametrization~~ ?? Parametric Curve.
1.3 directional derivative.
1.4 gradient.
1.5. gradient descent method.

1.1. Chain rule.

Theorem 1.

$x = x(t)$가 t에 대해 미분 가능하고
$f = f(x)$가 x에 대해 미분 가능한 함수일때,

$$\frac{df(x(t))}{dt} = \frac{df(x(t))}{dx}\frac{dx}{dt}$$ 가 성립한다.

Theorem 2

$x = x(t)$, $y = y(t)$가 각각 t에 대해 미분가능하고.
$f = f(x, y)$가 x와 y에 대해 미분가능한 함수일때.

$$\frac{df(x(t),y(t))}{dt} = \frac{\partial f}{\partial x}\frac{dx}{dt} + \frac{\partial f}{\partial y}\frac{dy}{dt}$$ 가 성립한다

Theorem 3

$x = x(t, s)$, $y = y(t, s)$가 각각 x와 y에 대해 미분가능하고.
$f = f(x, y)$가 x와 y에 대해 미분가능한 함수일때,

$$\frac{\partial f}{\partial s} = \frac{\partial f}{\partial x}\frac{\partial x}{\partial s} + \frac{\partial f}{\partial y}\frac{\partial y}{\partial s}, \quad \frac{\partial f}{\partial t} = \frac{\partial f}{\partial x}\frac{\partial x}{\partial t} + \frac{\partial f}{\partial y}\frac{\partial y}{\partial t}$$
가 성립한다.

1.2. Parametric Curve

Consider Parametric equation.

$$\begin{cases} x = x(t) \\ y = y(t) \end{cases} \quad \cdots (*)$$

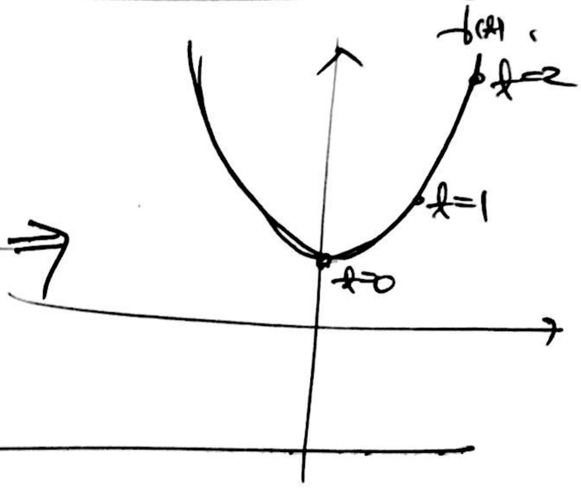
We have a Curve as the solution set of $(*)$

$$\gamma : \{ \gamma(t) = (x(t), y(t)) \mid a \le t \le b. \}.$$

ex). ① $\gamma(t) = (x(t), y(t))$
$= (t, t^2+2)$

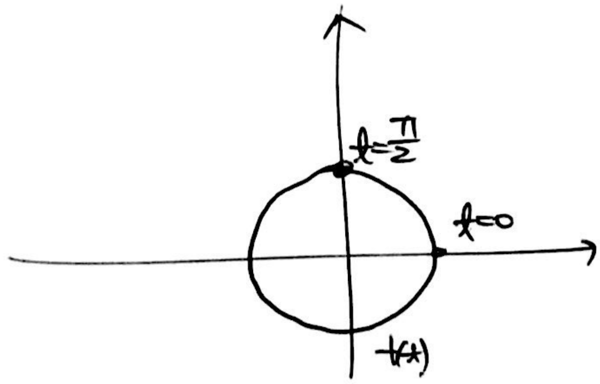
or. $x = x(t) = t$
$y = y(t) = t^2+2.$

$\Rightarrow$



② $x = x(t) = \cos t$
$y = y(t) = \sin t$

$\Rightarrow$

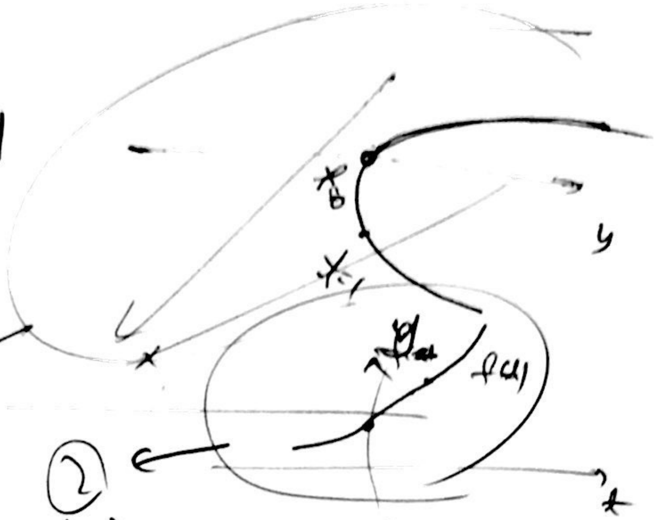


We can calculate the derivative of a function along a curve.

ex) $f = f(x,y) = x^2 + y$

$x(t) = t$

$y(t) = t^2 + 2.$



① ②

$$\frac{df}{dt} = \frac{d}{dt}(x^2+y) = \frac{d}{dt}\left(t^2 + t^2 + 2\right)$$

$$= 4t.$$

$$\frac{df}{dt} = \frac{\partial f}{\partial x}\frac{dx}{dt} + \frac{\partial f}{\partial y}\frac{dy}{dt}$$

$$= 2x \cdot 1 + 1 \cdot 2t$$

$$= 4t$$

1.3. Directional Derivative. (방향도함수)

We define the Directional Derivative $D_u f(P)$

as $$D_u f(P) = \lim_{h \to 0} \frac{f(a+hu_1, b+hu_2) - f(a,b)}{h}$$

where $P = (a,b)$, $\vec{u} = \langle u_1, u_2 \rangle$.
$\|\vec{u}\| = 1$.

Theorem.

$$D_u f(P) = \cancel{\frac{\partial f}{\partial x}\frac{dx}{dt}} \quad \frac{\partial f}{\partial x} u_1 + \frac{\partial f}{\partial y} u_2 .$$

Consider.

$f = f(x,y) = f(x(t), y(t))$

$x(t) = a + tu_1$

$y(t) = b + tu_2$

$$\left.\frac{df}{dt}\right|_{t=0} = \lim_{h\to 0} \frac{f(x(0+h), y(0+h)) - f(x(0), y(0))}{h}$$

$$= \lim_{h\to 0} \frac{f(a+hu_1, b+hu_2) - f(a,b)}{h}$$

$$\left.\frac{df}{dt}\right|_{t=0} = \left.\frac{\partial f}{\partial x}\right|_{\substack{x=a\\y=b}} \left.\frac{dx}{dt}\right|_{t=0} + \left.\frac{\partial f}{\partial y}\right|_{\substack{x=a\\y=b}} \left.\frac{dy}{dt}\right|_{t=0}$$

$$= \frac{\partial f}{\partial x} u_1 + \frac{\partial f}{\partial y} u_2 . \qquad \blacksquare$$

1.4. Gradient

<u>Def</u>. the gradient of function f is defined by $\nabla f := \left\langle \frac{\partial f}{\partial x}, \frac{\partial f}{\partial y} \right\rangle$.

---

ex). $f = f(x, y) = x^2 + xy - y + 1$.

$$\Rightarrow \nabla f = \langle 2x + y, x - 1 \rangle.$$

$$\nabla f = \nabla f(x, y) = \langle 2x + y, x - 1 \rangle.$$

$$\nabla f(1, 1) = \langle 3, 0 \rangle$$

Recall $D_u f(a, b) = \frac{\partial f}{\partial x} u_1 + \frac{\partial f}{\partial y} u_2$, $\quad \vec{u} = \langle u_1, u_2 \rangle$. $\|\vec{u}\| = 1$.

$$\Rightarrow D_u f(a, b) = \nabla f_{(a,b)} \cdot \vec{u}$$

$$= \|\nabla f_{(a,b)}\| \, \|\vec{u}\| \cos\theta.$$

$$= \|\nabla f_{(a,b)}\| \cos\theta.$$

$$\boxed{D_u f(1, 1)} = \|\langle 3, 0 \rangle\| \cos\theta$$

Q. 어떤 방향벡터 $\vec{u}$를 잡아야 함수가 가장 큰 변화율을 갖을까.

1.5. Gradient Descent Method.

$\boxed{\text{Idea}}$. $f = f(x,y)$는 $-\nabla f$의 방향으로 가장빨리 감소한다.

$\alpha \leftarrow \alpha_0$.

$\Rightarrow$ $(x, y) \longleftarrow (x_0, y_0)$

Iterate {

~~Until~~ [If $f(x,y)$ does not decrease then break.]

$(x, y) \longleftarrow (x, y) - \alpha \nabla f(x,y)$.

}

α is called by Learning rate
Step size.

Theorem For a function $f: \mathbb{R}^n \to \mathbb{R}$,
The gradient descent algorithm with fixed step size η Converge to a stationary point. If $\eta < \frac{2}{L}$, where L ~~is Constant $\in \mathbb{R}$~~ is scalar $\in \mathbb{R}$ satisfying

$$\|\nabla f(x) - \nabla f(y)\|_2 \le L \|x - y\|_2 \quad \forall x, y \in \mathbb{R}^n$$

Scanned by CamScanner

2. Optimization in Machine Learning Algorithm

Machine Learning Algorithm

- Unsupervised Learning
  - Dimensional Reduction.
  - Clustering.
- Supervised Learning
  - Regression.
  - Classification.
    - Linear Regression
    - Logistic Regression
    - Support Vector Machine
    - Random Forest.
    - Neural Network.

$h_\theta(x)$: hypothesis function
with parameter θ, variable x.

Cost : Cost function to minimize.

Linear Regression

$\theta = (\theta_0, \theta_1, \theta_2, \cdots, \theta_n)$.

$x = (x_0, x_1, x_2, \cdots, x_n)$, $y = \text{target}$.

$x_0 = 1$

$h_\theta(x) = \theta^T x = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n$.

$$Cost = \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2$$

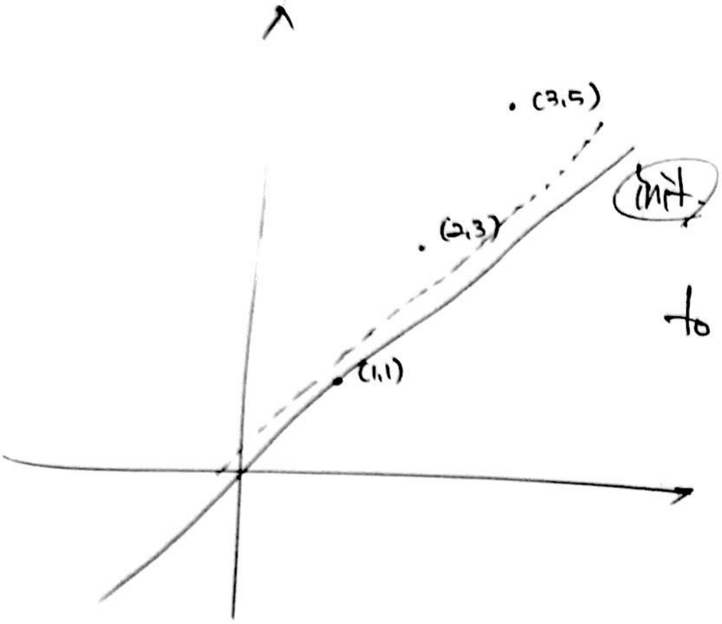
We want to minimize Cost.

$\Rightarrow$ need to calculate gradient.

~~f denote~~

$$f = Cost = \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2$$

$$\Rightarrow \nabla f = \langle 2x_0^{(i)} \sum_{i=1}^{m} h_\theta(x^{(i)}) - y^{(i)}),$$
$$2x_1^{(i)} \sum_{i=1}^{m} h_\theta(x^{(i)}) - y^{(i)}),$$
$$\vdots \quad \rangle.$$



to find. $y = 2x - 1$.

$X = (x_0, x_1)$
$\quad\; \overset{\shortparallel}{1}$.

$X^{(1)} = (1, 1) \qquad y^{(1)} = 1$

$X^{(2)} = (1, 2) \qquad y^{(2)} = 3$

$X^{(3)} = (1, 3) \qquad y^{(3)} = 5$.

$h_\theta(X) = \theta_0 + \theta_1 x_1$.

$$Cost = \sum_{i=1}^{3} \left(\theta_0 + \theta_1 x_1^{(i)} - y^{(i)} \right)^2.$$

by updating θ, we will minimize the Cost.

Init $\theta = (\theta_0, \theta_1) = (0, 1)$.

$\Rightarrow \quad h_\theta(X) = x_1$.

$\nabla f = \langle -6, -16 \rangle$.

$\vdots$
$\downarrow$
$\theta = (-1, 2)$.

next $\theta = \langle 0, 1 \rangle - 0.001 \langle -6, -16 \rangle$.

$= \langle 0.006, 1.016 \rangle$

3. Heuristic Algorithm.

* harmonic search Algorithm.

want to minimize $(x_1-2)^2+(x_2-4)^4$

$\Rightarrow x_1=2,\ x_2=4$ 일때 최솟값 0을 갖는다.

The Algorithm has the following steps.

Step 1. generate harmonic memory.

memory size:
보통 30~100

x_1	x_2	Cost.
1	4	1
2	5	1
3	5	2
⋮	⋮	⋮
1	2	17

Step 2. generate a new memory.

x_1	x_2	Cost.
1	3.	

90%의 확률로 이미 있던 값중에,
10% 확률로 임의의 값을 생성.

70% 확률로 그대로 사용
30% 확률로 ±3 으로 사용.

Step 3. If the new memory is better than the worst memory in harmonic memory, then replace.

Step 4. Repeat step 2, step 3 until criterion is satisfied.