Derivative

1. Chain Rule (합성함수의 미분)
$(g(f(x)))' = (g \circ f)'(x) = g'(f(x)) \cdot f'(x) = \frac{\partial g}{\partial f} \circ \frac{\partial f}{\partial x}$

Gradient (다변수 함수의 미분)

1. $\nabla_x f(x) = \left[\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \cdots, \frac{\partial f}{\partial x_n}\right]^T$ (변수마다 미분한 것을 벡터로 쌓는다.)

Jacobian (벡터 함수의 미분)

1. 벡터 함수: Input, Output이 모두 벡터인 함수를 벡터 함수라고 한다.

2. $J(x_1, x_2, \cdots, x_m) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_2}{\partial x_1} & \cdots & \frac{\partial f_n}{\partial x_1} \\ \frac{\partial f_1}{\partial x_2} & \frac{\partial f_2}{\partial x_2} & & \\ \vdots & & \ddots & \\ \frac{\partial f_1}{\partial x_m} & & & \frac{\partial f_n}{\partial x_m} \end{bmatrix}$ 전치 $(\nabla f_1)^T = \begin{bmatrix} \nabla_{\vec{x}} f_1(\vec{x})^T \\ \vdots \\ \nabla_{\vec{x}} f_n(\vec{x})^T \end{bmatrix}$

Hessian (벡터 함수의 2차 미분)

1. $H(f(x)) = \frac{\partial}{\partial x} \nabla_x f(x) = J(\nabla_x (f(x)))$
2. Hessian은 항상 대칭행렬이다.

알아두면 좋은 공식들

$\nabla_{\vec{x}} \vec{b}^T \vec{x} = \vec{b} \Leftrightarrow \frac{d}{dx} ax = a$

$\nabla^2_{\vec{x}} \vec{b}^T \vec{x} = 0 \Leftrightarrow \frac{d^2}{dx^2} ax = 0$

If A가 대칭행렬 If not,

$\nabla_{\vec{x}} \vec{x}^T A \vec{x} = 2A\vec{x} \Leftrightarrow \frac{d}{dx} ax^2 = 2ax$

$\nabla^2_{\vec{x}} \vec{x}^T A \vec{x} = 2A \Leftrightarrow \frac{d^2}{dx^2} ax^2 = 2a \quad \Rightarrow (A^T + A)\vec{x}$

Iterative Optimization

1. 목표: $\hat{\theta} = \arg\min_\theta J(\theta)$ cf) $\min J(\theta)$: $J(\theta)$의 최솟값, $\arg\min J(\theta)$: $J(\theta)$가 최소일때의 θ 값

2. 방법: 1) $J(\theta)$가 작아지도록 $\Delta\theta$를 구한다.
   2) θ를 업데이트 한다. $\theta \rightarrow \theta + \Delta\theta$

3. 종류: 1) Gradient Descent (경사 하강법)
   Batch mode (모든 훈련 데이터들의 Gradient의 평균을 구한 뒤 한꺼번에 Update)
   Online mode (한 개의 훈련 데이터마다 Gradient를 구한 뒤 즉시 갱신)
   Mini-Batch mode (훈련 집합을 여러 개의 MiniBatch로 나누고 Mini Batch 안의 훈련 데이터들의 Gradient의 평균을 구한 뒤 Mini Batch마다 갱신)

2) Stochastic Gradient Descent

결과가 훈련 데이터의 순서에 의존하지 않도록 훈련 데이터의 순서를 임의적으로 선택한다.

Shuffling : 훈련 데이터를 섞는 방법

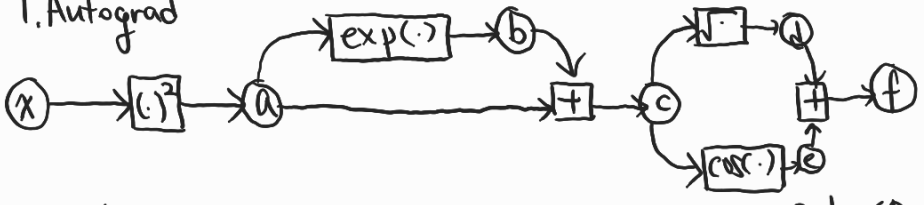
Sampling : 훈련 집합에서 임의로 훈련 데이터를 뽑는 방법

Automatic Differentiation

1. Backpropagation (역전파)

ex) $f(x)=\sqrt{x^2+\exp(x^2)}+\cos(x^2+\exp(x^2))$

1. Autograd



2. Calculate

$a=x^2$ $\quad \frac{\partial a}{\partial x}=2x$

$b=\exp(a)$ $\quad \frac{\partial b}{\partial a}=\exp(a)$

$c=a+b$

$d=\sqrt{c}$ $\quad \frac{\partial d}{\partial c}=\frac{1}{2\sqrt{c}}$

$e=\cos(c)$ $\quad \frac{\partial e}{\partial c}=-\sin(c)$

$f=d+e$

3. Chain Rule (Reverse)

$$\frac{\partial f}{\partial c}=\frac{\partial d}{\partial c}+\frac{\partial e}{\partial c}$$

$$\frac{\partial f}{\partial a}=\frac{\partial f}{\partial c}\cdot\frac{\partial c}{\partial a}=\frac{\partial f}{\partial c}\left(\frac{\partial b}{\partial a}+1\right)$$

$$\frac{\partial f}{\partial x}=\frac{\partial f}{\partial a}\cdot\frac{\partial a}{\partial x}$$