Optimization Algorithms for Machine/Deep Learning

Machine Learning Models

Classifiacation

- Logistic regression
- Support vector machine(SVM)

Regression

- Ordinary least square
- Lasso
- Deep learning

Clustering

Dimension Reduction

models (with parameters)

s.t. minimize $f(x)\, x \in X$

Examples

1. Ordinary least square

   $(u^{(i)}, v^{(i)}) \quad i = 1, \ldots, N$ where u is input and v is output

   $v^{(i)} \approx \theta^{\mathrm{T}} u^{(i)} = \sum_{j=1}^{n} \theta_j u_j^{(i)}$

   minimize $\sum_{i=1}^{N} (u^{(i)} - \theta^{\mathrm{T}} u^{(i)})^2 \quad \theta \in \mathbb{R}^n$

2. Logistic regression

   $(u^{(i)}, v^{(i)}) \quad i = 1, \ldots, N$

   $v^{(i)} \in \{0, 1\}$

   $v^{(i)} = \theta^{\mathrm{T}} u^{(i)}$

   $g(z) = \frac{1}{1+e^{-z}}$

   $v^{(i)} = h(u^{(i)}) = g(\theta^{\mathrm{T}} u^{(i)}) = \frac{1}{1+e^{-\theta^{\mathrm{T}} u^{(i)}}}$

   Probability distribution of $v^{(i)} \in \{0, 1\}$

   $max_\theta log \prod_{i=1}^{N} (1 - h(u^{(i)}))^{1-v^{(i)}} (h(u^{(i)}))^{v^{(i)}}$

   $= max_\theta \sum_{i=1}^{N} log(1 - h(u^{(i)}))^{1-v^{(i)}} (h(u^{(i)}))^{v^{(i)}}$

   $= max_\theta \sum_{i=1}^{N} (1 - v^{(i)}) log(1 - h(u^{(i)})) + v^{(i)} logh(u^{(i)})$

$= max_\theta \sum_{i=1}^{N} (1 - v^{(i)}) log \frac{e^{-\theta^T u^{(i)}}}{1+e^{-\theta^T u^{(i)}}} + v^{(i)} log \frac{1}{1+e^{-\theta^T u^{(i)}}}$

3. Support vector machine

   $min \frac{1}{2} ||w||^2 + \sum_{i=1} v^{(i)} max\{0, 1 - w^T u^{(i)} + b\}$

   $min \frac{1}{2} ||w||^2$ s.t.

   $\frac{w^T u^{(i)} + b}{||w||} \geq 0 \quad v^{(i)} = 1$

   $\frac{w^T u^{(i)} + b}{||w||} \leq 0 \quad v^{(i)} = -1$

   $d^{(i)} = \frac{v^{(i)} (w^T u^{(i)} + b)}{||w||}$

   $max_{w,b} min_{i=1,...,N} \frac{v^{(i)} (w^T u^{(i)} + b)}{||w||}$

   $\Leftrightarrow max \frac{min_{i=1,...,N} v^{(i)} (w^T u^{(i)} + b)}{||w||}$

   $\Leftrightarrow max \frac{r}{||w||}$ s.t. $v^{(i)} (w^T u^{(i)} + b) \geq r, \quad i = 1, \ldots, N$

$\Leftrightarrow max \frac{1}{||w||}$ s.t. $v^{(i)} (w^T u^{(i)} + b) \geq 1, \quad i = 1, \ldots, N$

$\Leftrightarrow min ||w||^2$ s.t. $v^{(i)} (w^T u^{(i)} + b) \geq 1, \quad i = 1, \ldots, N$

$min \frac{\rho}{2} ||w||^2 + \sum_{i=1}^{N} max\{1 - v^{(i)} (w^T u^{(i)} + b), 0\}$

4. Neural Network

   $(u^{(i)}, v^{(i)})$

   $$wu^i \in \mathbb{R}^m \Rightarrow g(wu^{(i)}) = \begin{pmatrix} \frac{1}{1+e^{(-wu^{(i)})_1}} \\ \frac{1}{1+e^{(-wu^{(i)})_2}} \\ \cdots \\ \frac{1}{1+e^{(-wu^{(i)})_m}} \end{pmatrix}, \quad w \in \mathbb{R}^{m \times n}, \quad g : \mathbb{R}^m \to \mathbb{R}^m$$

   $\theta \in \mathbb{R}^m$

   $v^{(i)} \approx \theta^T g(wu^{(i)})$

   $min \sum_{i=1}^{N} (v^{(i)} - \theta^T g(wu^{(i)}))^2$

   multi-layer:

   $min \sum_{i=1}^{N} (v^{(i)} - \theta^T g(w_e g(w_{e-1} g(w_{e-2} \ldots w_2 g(w_1, u^{(i)})))))^2$

5. Lasso regression

   $min \sum_{i=1}^{N} (v^{(i)} - \theta^T u^{(i)})^2 + \rho ||\theta||_1$

   ★ Stochastic Optimization Formulation

   $min_\theta E_{(u,v)} [(v - \theta^T u)^2] + \rho ||\theta||_1$

   $(u^{(i)}, v^{(i)}) \quad i = 1, \ldots, N$

$$min_\theta \frac{1}{N}\sum_{i=1}^{N}(v^{(i)}-\theta^{\mathrm{T}}u^{(i)})^2+\rho||\theta||_1$$

General Form

$minf(\theta)+\gamma(\theta),\ \ \theta\in X$

$$\begin{cases} f(\theta)\approx\frac{1}{N}\sum_{i=1}^{N}F(\theta,u^{(i)},v^{(i)}) \\ f(\theta)=E[F(\theta,u^{(i)},v^{(i)})] \end{cases}$$

Review of Convex Analysis

Convex functions

`todo:`

Subgradient

Let $X\subseteq\mathbb{R}^n$ be a convex set

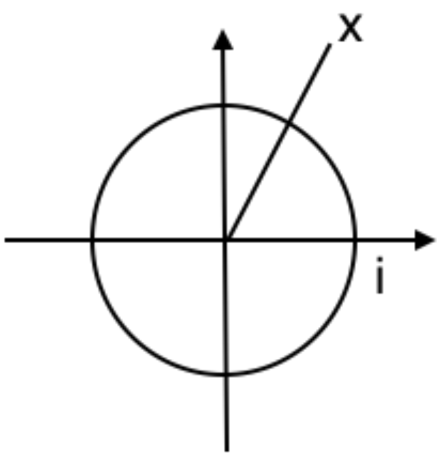
$f:X\rightarrow\mathbb{R}$ be a convex function

$g\in\mathbb{R}^n$ is called a subgradient of f at $x\in X$ if $f(y)\geq f(x)+<g,y-x>,\ \ \forall y\in X$

The subgradient of f at x exists if $x\in Int(X)$

Projection

$Proj_X(x)=argmin||y-x||^2,\ \ y\in X$

e.g. 

$$Proj_X(x)=\begin{cases} x, & ||x||\leq 1 \\ \frac{x}{||x||}, & ||x||>1 \end{cases}$$

e.g. $X=\{x\in\mathbb{R}^n:\sum_{i=1}^{n}x_i\leq 1,x_i\geq 0\}$

$min||a - x||^2$, $\sum_{i=1}^{n} x_i = 1, x_i \geq 0$

Use (KKT) Optimality conditions to solve the problem.

Review of Optimality Conditions

$minf(x)$, $x \in X$

Simple Optimality Condition

x^* is an optimal solution if ∃ subgradient $g(x^*)$, s.t. $< g(x*), x - x^* >\geq 0$, $\forall x \in X$

Review of Convex Analysis

If f is differentiable, $< \nabla f(x^*), x - x^* >\geq 0$, $\forall x \in X$

$x = x^* - \varepsilon \frac{\nabla f(x^*)}{||\nabla f(x^*)||}$

$< \nabla f(x^*), x - x^* >=< \nabla f(x^*), -\varepsilon \frac{\nabla f(x^*)}{||\nabla f(x^*)||} >= -\varepsilon ||\nabla f(x^*)|| \geq 0$

e.g. $min \sum_{i=1}^{n} [a_i x_i + \frac{1}{2} x_i^2]$, $x_i > 0, i = 1, \ldots, n$

$min(a_i x_i + \frac{1}{2} x_i^2)$, $x_i \geq 0$

$\nabla f(x^*) = a_i + x_i^*$

$< a_i, x_i >\geq 0$, $\forall x_i \geq 0$

$< a_i + x_i^*, x_i - x_i^* >\geq 0$, $\forall x_i \geq 0$

Suppose $x_i* \geq 0$

$a_i + x_i^* = 0$

$x_i^* = -a_i$

If $a_i < 0$, $x_i^* = -a_i$

If $a_i \geq 0$, $x_i^* = 0$

$$x_i^* = \begin{cases} -a_i, & a_i < 0 \\ 0, & otherwise \end{cases}$$