Gradient Descent for Support Vector Machines and Subgradients

Objectives

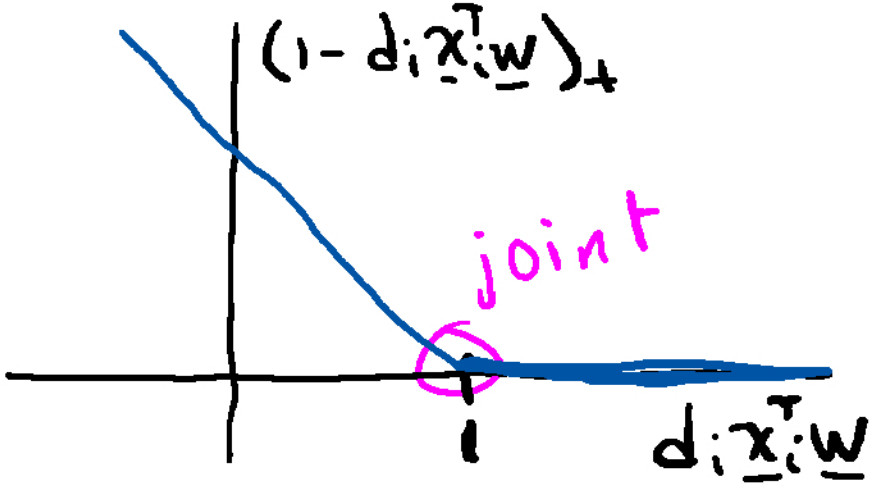
- develop a gradient descent algorithm for SVMs

- introduce subgradients for convex but non differentiable cost functions

Support vector machines require iterative algorithms

$$\min_{\underline{w}} \sum_{i=1}^{N} (1 - d_i \underline{x}_i^T \underline{w})_+ + \lambda \|\underline{w}\|_2^2$$

labels: d_i — features: $\underline{x}_i$ — hinge loss: $(\cdot)_+$ — regularization: $\lambda \|\underline{w}\|_2^2$



No closed form solution

Convex

$\Rightarrow$ gradient descent

Problem: hinge loss not differentiable

☆ (Not differentiable, but convex → use a gradient-descent based solution to solve)

Subderivatives generalize derivatives

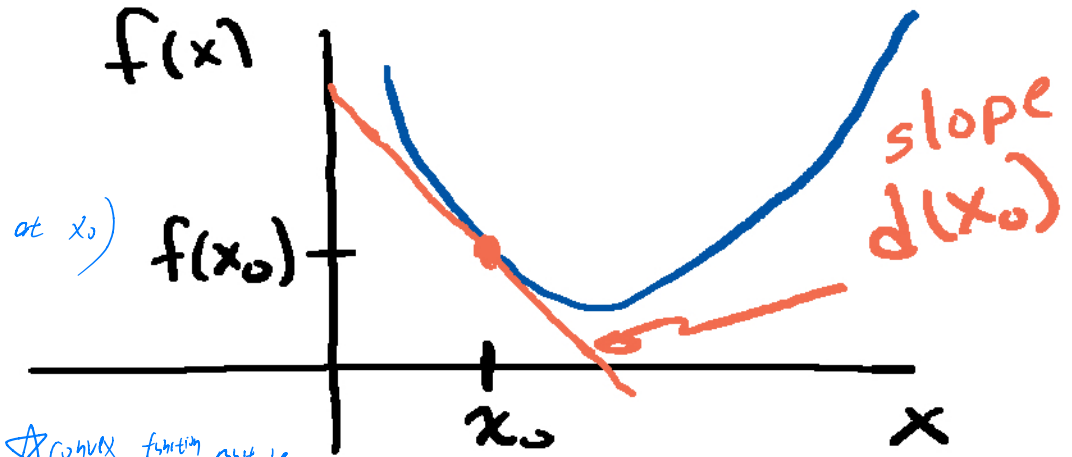
– Convex, but nondifferentiable $f(x)$

☆(次导数是在处理不可微导数的凸函数时，对值域导数的推广)

☆核心思想：当 $f(x)$ 在某点可微 → 次导数就是导数；不可微 → 导数不存在，定义次导数

– Derivatives –

$$d(x_0) = \lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0}$$ (Tangent line at x_0)



Convex:

$$f(x) \geq f(x_0) + d(x_0)(x - x_0)$$

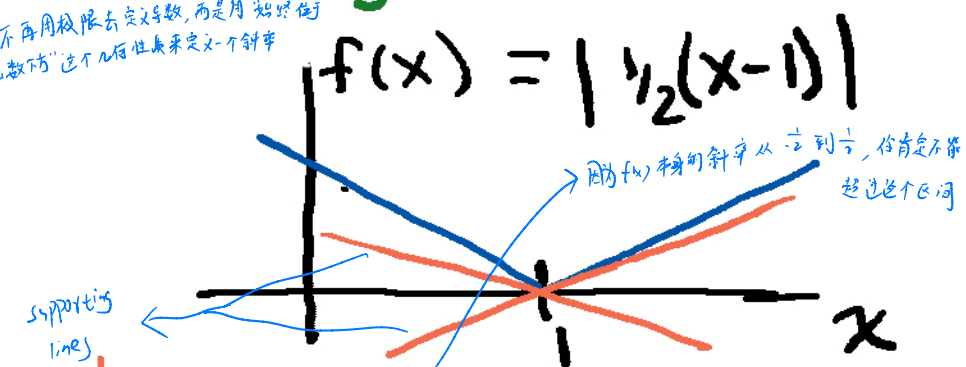
☆convex function must be "above tangent line"

Subderivative (convex)

☆不再用极限去定义导数，而是用"始终位于函数下方"这个几何性质来定义一个斜率

Any $d_s(x_0)$: (sub derivative) $f(x) \geq f(x_0) + d_s(x_0)(x - x_0)$

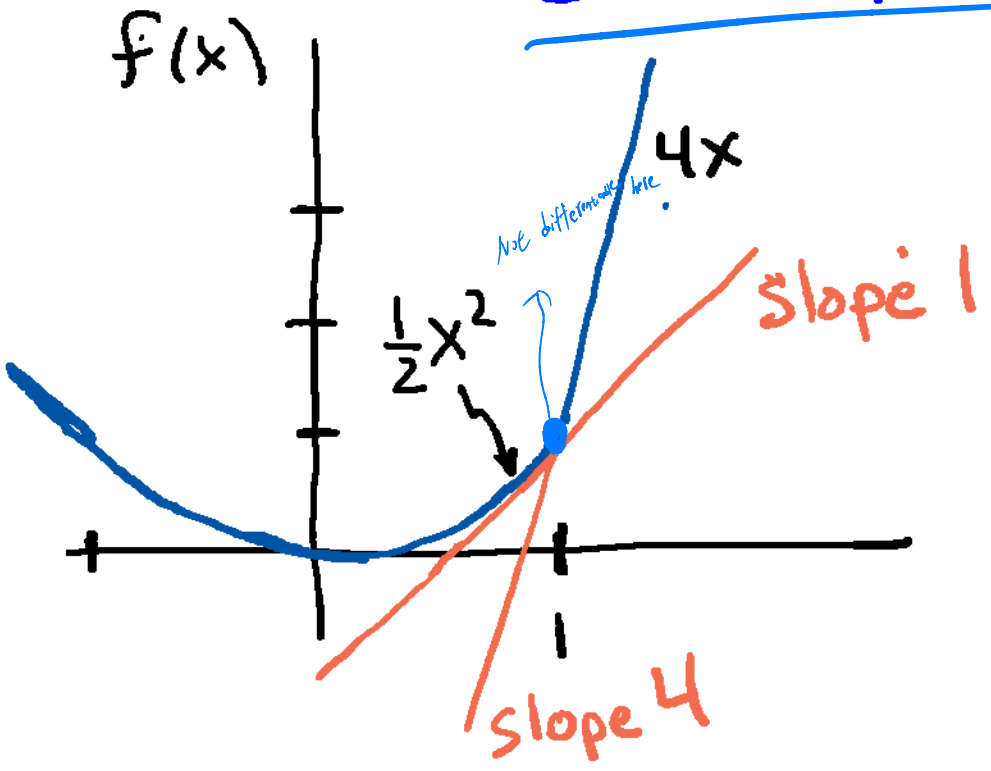
$x < 1$: $d_s(x) = -\frac{1}{2}$; $x > 1$: $d_s(x) = \frac{1}{2}$



☆次导数是一个数，而是满足 $m \in [-\frac{1}{2}, \frac{1}{2}]$ 的所有斜率的集合

$-\frac{1}{2} \leq d_s(1) \leq \frac{1}{2}$

Subderivatives produce "reasonable" downhill directions



Example: $f(x) = \begin{cases} \frac{1}{2}x^2 & x<1 \\ 4x & x>1 \end{cases}$

convex

Subderivative

$d_s(x) = \begin{cases} x \; (= \frac{d}{dx}(\frac{1}{2}x^2) = x) & x<1 \\ 4 \; (= \frac{d}{dx}(4x) = 4) & x>1 \\ [1,4] & x=1 \end{cases}$

(Any interval between 1 and 4)

Subgradients generalize gradients

- Convex, nondifferentiable $\ell(\underline{w})$

就是 subderivative 在多维中的叫法 (subderivative 只在一维空间)

(切有 "supporting plane" 的梯度向量的集合)

在 $\underline{w}_0$ 点的梯度向量 $[\frac{\partial \ell}{\partial w_1} \frac{\partial \ell}{\partial w_2} \cdots]$

Tangent plane

Gradients-

$$\ell(\underline{w}) \geq \ell(\underline{w}_0) + (\underline{w}-\underline{w}_0)^T \underline{v}(\underline{w}_0) \qquad \underline{v}(\underline{w}) = \nabla_{\underline{w}} \ell(\underline{w})$$

"above tangent plane" $\left(\sum_{i=1}^{n} (w_i - w_{0i}) \frac{d}{dw_i} \ell(\underline{w}) \right)$

Subgradients-

在 $\underline{w}_0$ 点, 任何一个向量 $\underline{v}(\underline{w}_0)$, 如果由它定义的 "支撑平面" $y = \ell(\underline{w}_0) + (\underline{w}-\underline{w}_0)^T \underline{v}(\underline{w}_0)$ 始终低于 $\ell(\underline{w})$ 之下, 那么这个 $\underline{v}(\underline{w}_0)$ 就是 $\underline{w}_0$ 点的一个 subgradient

Any $\underline{v}(\underline{w})$: $\ell(\underline{w}) \geq \ell(\underline{w}_0) + (\underline{w}-\underline{w}_0)^T \underline{v}(\underline{w}_0)$

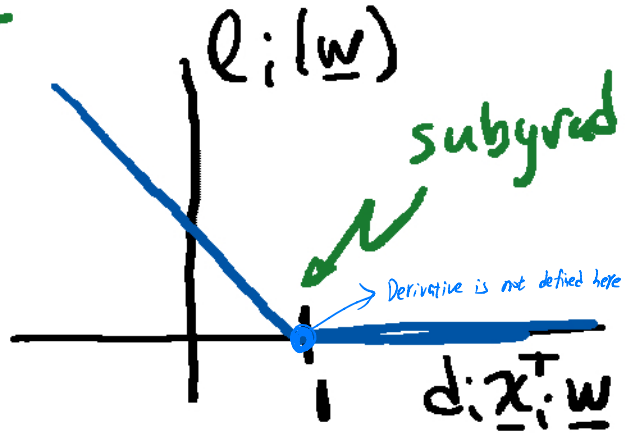
Gradient descent optimization: replace gradient with subgradient

Gradient descent for SVMs

$$\ell(\underline{w}) = \sum_{i=1}^{N} (1 - d_i \underline{x}_i^T \underline{w})_+$$ Hinge Loss → subgradient

$$\ell_i(\underline{w}) = (1 - d_i \underline{x}_i^T \underline{w})_+ = \begin{cases} 1 - d_i \underline{x}_i^T \underline{w} & d_i \underline{x}_i^T \underline{w} < 1 \\ 0 & d_i \underline{x}_i^T \underline{w} > 1 \end{cases}$$

for each sample i



Subgradient

$$\underline{v}_i(\underline{w}) = \begin{cases} -d_i \underline{x}_i & d_i \underline{x}_i^T \underline{w} < 1 \\ 0 & d_i \underline{x}_i^T \underline{w} \geq 1 \end{cases} = -d_i \underline{x}_i \mathbb{I}_{\{d_i \underline{x}_i^T \underline{w} < 1\}}$$

$\nabla_{\underline{w}}(1 - d_i \underline{x}_i^T \underline{w}) = -d_i \underline{x}_i$

如果{}内条件为真，等于1，否则等于0

indicator function

Cost

$$f(\underline{w}) = \ell(\underline{w}) + \lambda \|\underline{w}\|_2^2$$

$$\Rightarrow \nabla f(\underline{w})\big|_{\underline{w}^{(k)}} = \sum_{i=1}^{N} (-d_i \underline{x}_i \mathbb{I}_{\{d_i \underline{x}_i^T \underline{w}^{(k)} < 1\}}) + 2\lambda \underline{w}^{(k)}$$

Gradient of $\lambda \|\underline{w}\|_2^2$

Gradient descent

$$\underline{w}^{(k+1)} = \underline{w}^{(k)} - \tau \nabla f(\underline{w})\big|_{\underline{w}^{(k)}}$$

最陡的方向，下坡