Labs
Machine Learning Course
Fall 2019

EPFL
School of Computer and Communication Sciences
Martin Jaggi & Rüdiger Urbanke
www.epfl.ch/labs/mlo/machine-learning-cs-433

Problem Set 2, Sept 26, 2019 (Solutions to Theory Questions)

$\nabla(y_n - x_n^\top w)$
$= -x_n^\top$

1 MAE Subgradient (Exercise 6)

$\partial h =$

The subgradient for the function $h : \mathbb{R} \to \mathbb{R}, h(e) := |e|$ is given as

$$g : \mathbb{R} \to \{-1, 0, 1\} \ \ , g(e) := sign(e).$$

$x_n^1 w^1 + x_n^2 w^2 \dots$

The MAE cost function is defined as $\mathcal{L}(\boldsymbol{w}) = \frac{1}{N}\sum_{n=1}^{N} |y_n - \boldsymbol{x}_n^\top \boldsymbol{w}|$.

As given in the annotated lecture notes 2, we can use the **chain-rule for subgradients**, for $\mathcal{L}(\boldsymbol{w}) := h(q(\boldsymbol{w}))$, when the outer function h is not differentiable, but q is differentiable. We write $\partial h(\boldsymbol{y})$ for the set of all subgradients of h at $\boldsymbol{y}$. Then any vector $\boldsymbol{g}$ of the following form is a subgradient of $\mathcal{L}$ at $\boldsymbol{w}$:

$$\boldsymbol{g} \in \partial h(q(\boldsymbol{w})) \cdot \nabla q(\boldsymbol{w})$$

where we can pick any element of the left, and multiply with the vector on the right (the gradient).

We now find the (sub)gradient update for a single component w_i and conclude by generalizing to the whole vector $\boldsymbol{w}$. In our case here, ∂h is the sign function. Then for $\frac{\partial \mathcal{L}(\boldsymbol{w})}{\partial w_i} = \frac{1}{N}\sum_{n=1}^{N} \frac{\partial |y_n - \boldsymbol{x}_n^\top \boldsymbol{w}|}{\partial w_i}$ we have that:

$$\frac{\partial \mathcal{L}(\boldsymbol{w})}{\partial w_i} = \frac{1}{N}\sum_{n=1}^{N} -(x_n)_i \, sign(y_n - \boldsymbol{x}_n^\top \boldsymbol{w}).$$

Finally we can conclude that $\frac{-1}{N}\boldsymbol{X}^\top \cdot sign(\boldsymbol{e})$ is a subgradient to $\mathcal{L}$ at $\boldsymbol{w}$, where $\boldsymbol{e} := \boldsymbol{y} - \boldsymbol{X} \cdot \boldsymbol{w}$ and $sign$ applied element-wise to $\boldsymbol{e}$, and $\boldsymbol{X}$ is the matrix collecting all datapoints as its rows.