Machine learning : Sheet 3

Author : Djordje Zivanovic

1. 1.

$$\mathcal{L}(w) = \sum_{i=1}^{N} |\mathbf{w}^{\mathrm{T}}\mathbf{x}_i - y_i| + \lambda \sum_{i=1}^{D} |w_i| \tag{1}$$

Objective function $\mathcal{L}$ is equivalent to solving the following linear program with $2D + N$ variables $w_1, \ldots, w_D, \zeta_1, \ldots, \zeta_{N+D}$:

$$\text{minimize} \sum_{i=1}^{N} \zeta_i$$

subject to:

$$\mathbf{w}^{\mathrm{T}}\mathbf{x}_i - y_i \leq \zeta_i, i \in \{1, \ldots, N\} \tag{2}$$

$$y_i - \mathbf{w}^{\mathrm{T}}\mathbf{x}_i \leq \zeta_i, i \in \{1, \ldots, N\} \tag{3}$$

$$\lambda w_i \leq \zeta_i, i \in \{N+1, \ldots, N+D\} \tag{4}$$

$$-\lambda w_i \leq \zeta_i, i \in \{N+1, \ldots, N+D\} \tag{5}$$

We need to prove that the linear program is indeed the solution of the $\mathcal{L}(w)$ objective function. First we need to prove that a solution exists. We notice that, any vector $\mathbf{w}$ in $\mathbb{R}^D$ and $\zeta_i = |\mathbf{w}^{\mathrm{T}}\mathbf{x}_i - -y_i|$ for $1 \leq i \leq N$ and $\zeta_i = |\lambda w_i|$ for $N+1 \leq i \leq N+D$ are solutions of the linear program. By seeing that equations 2 and 3 are the definitions of absolute values for $\zeta_i, i \in \{1, \ldots, N\}$ and that equalities hold in all cases, this solution is valid for these inequalities. Similarly, equations 4 and 5 are the definitions of absolute values for $\zeta_i, i \in \{N+1, \ldots, N+D\}$ and equalities hold in all cases. Thus the linear program has at least one solution.
Further, we need to prove that the solution of the linear program is the solution of the loss function 1. Since ζ_is are limited by inequalities from bellow and for all of them limitations are absolute values which are nonnegative numbers, this means their sum is minimized when the sum of these nonnegative numbers is minimized. Thus, the solution of program has to be the solution of minimization of objective function, because a minimization over a sum of nonnegative numbers is done on the whole $\mathbb{R}^D$ vector space for $\mathbf{w}$ which is equivalent to the objective function 1.

2. First let us compute subgradient:

$$\nabla_{\mathbf{w}}\mathcal{L}(\mathbf{w} = \nabla_{\mathbf{w}} \sum_{i=1}^{N} (\mathbf{w}^{\mathrm{T}}\mathbf{x}_i - y_i)^2 + \lambda \sum_{i=1}^{D} \nabla_{\mathbf{w}}|w_i| \tag{6}$$

$$= \nabla_{\mathbf{w}}(\mathbf{X}\mathbf{w} - \mathbf{y})^{\mathrm{T}}(\mathbf{X}\mathbf{w} - \mathbf{y}) + \lambda \sum_{i=1}^{D} \nabla_{\mathbf{w}}|w_i| \tag{7}$$

$$= \nabla_{\mathbf{w}}(\mathbf{w}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}} - \mathbf{y}^{\mathrm{T}})(\mathbf{X}\mathbf{w} - \mathbf{y}) + \lambda \sum_{i=1}^{D} \nabla_{\mathbf{w}}|w_i| \tag{8}$$

$$= \nabla_{\mathbf{w}}\mathbf{w}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{w} - \nabla_{\mathbf{w}}\mathbf{w}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{y} - \nabla_{\mathbf{w}}\mathbf{y}^{\mathrm{T}}\mathbf{X}\mathbf{w} + \nabla_{\mathbf{w}}\mathbf{y}^{\mathrm{T}}\mathbf{y} + \lambda \sum_{i=1}^{D} \nabla_{\mathbf{w}}|w_i| \tag{9}$$

$$= 2\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{w} - 2\mathbf{X}^{\mathrm{T}} + 0 + \lambda \sum_{i=1}^{D} \nabla_{\mathbf{w}}|w_i| \tag{10}$$

In the previous derivation we used the gradient property to "go" through addition in equation 6. The equation 7 is compacted version of the sum from the equation 6. The equation 8 uses property that transposition of matrices exchanges the order of multipliers in multiplication. The equation 9 uses simple calculations and the property of gradient to pass through addition. The equation 10 uses gradient rules for derivations such as quadratic and multiplied by constant. A subgradient of $|w_i|$ is $g_{\mathbf{w}}|w_i| = [0, ..., sgnext(w_i), ..., 0]^{\mathrm{T}}$, where $sgnext$ represents extended sgn function which for values greater than zero is 1, less than zero is -1, and for the value 0 is either -1 or 1. The subgradient column vector has all cells 0 except the i-th which is $sgnext$.

$$g_{\mathbf{w}} = 2\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{w} - 2\mathbf{X}^{\mathrm{T}} + 0 + \lambda \sum_{i=1}^{D} g_{\mathbf{w}}|w_i|$$

Thus the calculation rule using modified version of gradient descent is:

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta g_{w_t}$$