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 gradient as much as we can:

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

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

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

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

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

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, the value 0 is either -1 or 1. The subgradient column vector has all cells 0 except the i-th which is $sgnext$. If we sum subgradients for each $w_i, i \in \{1, ..., D\}$,

we get a column vector whose cells are values of *sgnext* of specific w_i. Finally, subgradient of the objective function is:

$$g_{\mathbf{w}} = 2\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{w} - 2\mathbf{X}^{\mathrm{T}}\mathbf{y} + \lambda[sgnext(w_1), ..., sgnext(w_D)]^{\mathrm{T}}$$

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

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

2. 1.

$$\sigma'(z) = \frac{\partial}{\partial z}\frac{1}{1+e^{-z}} = \frac{\partial}{\partial z}(1+e^{-z})^{-1} \tag{11}$$

$$= \frac{e^{-z}}{(1+e^{-z})^2} \tag{12}$$

$$= \sigma(z)\frac{e^{-z}}{1+e^{-z}} == \sigma(z)\left(\frac{1+e^{-z}-1}{1+e^{-z}}\right)\sigma(z)\left(1-\frac{1}{1+e^{-z}}\right) \tag{13}$$

$$= \sigma(z)(1-\sigma(z)) \tag{14}$$

The equation 11 is definition for derivation of function in a point. The equation 12 uses derivation properties (exponential function, constant, power function). Equation 13 uses simple tricks so we could get the end result. The equation 14 is what is requested.

2.

$$Likelihood(\mathbf{y}) = p(\mathbf{y}|\mathbf{X},\mathbf{w}) = \prod_{i=1}^{N} p(y_i|\mathbf{x}_i,\mathbf{w}) = \prod_{i=1}^{N}(\sigma(\mathbf{x}_i^{\mathrm{T}}\mathbf{w}))^{y_i}(1-\sigma(\mathbf{x}_i^{\mathrm{T}}\mathbf{w}))^{(1-y_i)}$$

The previous is correct because for $y_i = 1$ just the first multiplier will be part of the product because $1 - y_1$ will set the second multiplier to 1. Similarly, stands for $y_i = 0$.

3. Taking negative log of the likelihood we get:

$$NLL(\mathbf{y}) - \log(Likelihood(\mathbf{y}) = -\log\prod_{i=1}^{N}\sigma(\mathbf{x}_i^{\mathrm{T}}\mathbf{w})^{y_i}(1-\sigma(\mathbf{x}_i^{\mathrm{T}}\mathbf{w}))^{(1-y_i)}$$

$$= -\sum_{i=1}^{N} y_i\log(\sigma(\mathbf{x}_i^{\mathrm{T}}\mathbf{w})) + (1-y_i)\log(1-\sigma(\mathbf{x}_i^{\mathrm{T}}\mathbf{w})) \tag{15}$$

Equation 15 uses basic properties of logarithm: $\log(ab) = \log(a) + \log(b)$, and $\log(a^b) = b\log(a)$.

$$\nabla_{\mathbf{w}}NLL(y_i|\mathbf{x}_i,\mathbf{w}) = \nabla_{\mathbf{w}} - \left[y_i\log(\sigma(\mathbf{x}_i^{\mathrm{T}}\mathbf{w})) + (1-y_i)\log(1-\sigma(\mathbf{x}_i^{\mathrm{T}}\mathbf{w}))\right] \tag{16}$$

$$= -\left[y_i\frac{\sigma(\mathbf{x}_i^{\mathrm{T}}\mathbf{w})\cdot(1-\sigma(\mathbf{x}_i^{\mathrm{T}}\mathbf{w}))\cdot\mathbf{x}_i}{\sigma(\mathbf{x}_i^{\mathrm{T}}\mathbf{w})} + (1-y_i)\frac{(-1)\cdot\sigma(\mathbf{x}_i^{\mathrm{T}}\mathbf{w})\cdot(1-\sigma(\mathbf{x}_i^{\mathrm{T}}\mathbf{w}))\cdot\mathbf{x}_i}{1-\sigma(\mathbf{x}_i^{\mathrm{T}}\mathbf{w})}\right] \tag{17}$$

$$= -y_i\cdot(1-\sigma(\mathbf{x}_i^{\mathrm{T}}\mathbf{w}))\cdot\mathbf{x}_i + (1-y_i)\cdot\sigma(\mathbf{x}_i^{\mathrm{T}}\mathbf{w})\cdot\mathbf{x}_i \tag{18}$$

$$= (-y_i + y_i\cdot\sigma(\mathbf{x}_i^{\mathrm{T}}\mathbf{w}) + \sigma(\mathbf{x}_i^{\mathrm{T}}\mathbf{w}) - y_i\cdot\sigma(\mathbf{x}_i^{\mathrm{T}}\mathbf{w}))\mathbf{x}_i \tag{19}$$

$$= (\sigma(\mathbf{x}_i^{\mathrm{T}}\mathbf{w}) - y_i)\mathbf{x}_i \tag{20}$$

The equation 16 is using negative log likelihood for one label got in the equation 15. The equation 17 uses rules for derivatives and gradients (gradient behaves the same as derivative until we "reach" the variable that is dependent on gradient operand). The equations 18, 19, 20 use simple additions and multiplications. Gradient of the negative log likelihood is the sum of gradients from 20 for each label:

$$\nabla_{\mathbf{w}}NLL(Y|\mathbf{X},\mathbf{w}) = \sum_{i=1}^{N}(\sigma(\mathbf{x}_i^{\mathrm{T}}\mathbf{w}) - y_i)\mathbf{x}_i$$

First we will calculate i-th row of Hessian matrix:

$$\frac{\partial}{\partial w_j}\nabla_{\mathbf{w}}NLL(Y|\mathbf{X},\mathbf{w}) = \sum_{i=1}^{N}\mathbf{x}_i^{\mathrm{T}}\cdot\sigma(\mathbf{x}_i^{\mathrm{T}}\mathbf{w})\cdot(1-\sigma(\mathbf{x}_i^{\mathrm{T}}\mathbf{w}))x_{i,j} \tag{21}$$

Notice that the equation 21 is derived using basic derivation properties. However, let us notice that for all the elements of first rows we have that different rows are multiplied by $[x_{1,1}, x_{2,1}, ..., x_{N,1}]^{\mathrm{T}}$ and notice that the multipliers are respectively $[x_{1,1}, ..., x_{1,N}], ..., [x_{D,1}, ..., x_{D,N}]$. Similar rule we can notice for the other rows of Hesisan Matrix, just instead of $[x_{1,1}, x_{2,1}, ..., x_{N,1}]^{\mathrm{T}}$ $[x_{1,j}, x_{2,j}, ..., x_{N,j}]^{\mathrm{T}}$ is, where j represents the row of the Hessian matrix. Thus, our Hessian matrix should be the product of the if we did not have σs:

$$\begin{bmatrix} x_{1,1} & x_{2,1} & x_{3,1} & \dots & x_{N,1} \\ x_{1,2} & x_{2,2} & x_{3,2} & \dots & x_{N,2} \\ \dots & \dots & \dots & \dots & \dots \\ x_{1,D} & x_{2,D} & x_{3,D} & \dots & x_{N,D} \end{bmatrix} \cdot \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,D} \\ x_{2,1} & x_{2,2} & \dots & x_{2,D} \\ x_{3,1} & x_{3,2} & \dots & x_{3,D} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{N,1} & x_{N,2} & \dots & x_{N,D} \end{bmatrix} = \mathbf{X}^{\mathrm{T}}\mathbf{X} \tag{22}$$

Sigma issue is easily solved because they are tied to the X matrix, we just need to insert diagonal matrix whose elements are, respectively: $\sigma(\mathbf{x}_i^{\mathrm{T}}\mathbf{w}) \cdot (1 - \sigma(\mathbf{x}_i^{\mathrm{T}}\mathbf{w}))$. Hence, Hessian becomes $H(NLL(\mathbf{y}|\mathbf{X}, \mathbf{w})) = \mathbf{X}^{\mathrm{T}} \cdot \mathbf{S} \cdot \mathbf{X}$, where $\mathbf{S}$ is diagonal matrix whose elements are described, previously. In order to prove that Hessian is positive-semi definite we have to prove that for every vector $\mathbf{z}$ from $\mathbb{R}^D$ in :

$$\begin{aligned} \mathbf{z}\mathbf{X}^{\mathrm{T}}\mathbf{D}\mathbf{X}\mathbf{z}^{\mathrm{T}} \geq 0 \iff & \\ (\mathbf{X}\mathbf{z}^{\mathrm{T}})^{\mathrm{T}}\mathbf{D}\mathbf{X}\mathbf{z}^{\mathrm{T}} \geq 0 \iff & \quad (23) \\ (\mathbf{X}\mathbf{z}^{\mathrm{T}})^{\mathrm{T}}\sqrt{\mathbf{D}}\sqrt{\mathbf{D}}\mathbf{X}\mathbf{z}^{\mathrm{T}} \geq 0 \iff & \quad (24) \\ (\sqrt{\mathbf{D}}\mathbf{X}\mathbf{z}^{\mathrm{T}})^{\mathrm{T}}(\sqrt{\mathbf{D}}\mathbf{X}\mathbf{z}^{\mathrm{T}}) \geq 0 \iff & \quad (25) \\ ||\sqrt{\mathbf{D}}\mathbf{X}\mathbf{z}^{\mathrm{T}}||_2^2 \geq 0 & \quad (26) \end{aligned}$$

The equation 23 is the simple usage of property of transposition of matrices. The equation 24 is possible because the entries of diagonal matrix are positive due to the fact that $\sigma(\mathbf{x}_i^{\mathrm{T}}\mathbf{w})$ is in segment $(0, 1)$, and $1 - \sigma(\mathbf{x}_i^{\mathrm{T}}\mathbf{w})$ is in $(0, 1)$ and we are always able to calculate square roots of diagonal entries since product of values of two positive functions is strictly positive. The equation 25 is the simple usage of property of transposition of matrices. The final equation 35 gives is possible because norm of the vector $\mathbf{x}$ is equal to $\mathbf{X}^{\mathrm{T}}\mathbf{X}$. Further, Eucledian norm is always positive, and this concludes the proof.

4. I would use Newton update rule. Derivations are given on the lectures:

$$\mathbf{w}_{t+1} = (\mathbf{X}^{\mathrm{T}}\mathbf{D}_t\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{D}_t(\mathbf{X}\mathbf{w}_t + \mathbf{D}_t^{-1}(\mathbf{y} - \sigma(\mathbf{X}\mathbf{w}))$$

The obtained solution would be correct because Hessian is semidefinite positive, and thus negative log likelihood will be convex function. The only problem with this approach is it would take immense amount of time, because of the calculations, and the problem should be reformulated.

3. 3.1 Because data set is linearly separable, $\alpha > 0$. Let us assume there is tuple $(\mathbf{w}^*, w_0^*)$ and $||\mathbf{w}^*|| = \beta < 1$ and it satisfies the solution. By dividing both sides of inequality by β for $i \in \{1, ..., N\}$ inequalities hold, but on the rhs of inequalities we have $\frac{\alpha}{\beta} > \alpha$ and $\frac{||\mathbf{w}^*||}{|\beta||} = 1$. Hence, we have a solution that maximizes α that is not $\mathbf{w}^*$. Contradiction! This means $||\mathbf{w}|| = 1$ holds for the solution of the program.
If we divide inequalities in the task paper formulation by α they are transformed into:

$$y_i(\mathbf{x}_i \cdot \frac{\mathbf{w}}{\alpha} + \frac{w_0}{\alpha}) \geq 1,\ i \in \{1, ..., N\}$$

Thus we need to maximize α or in our case to minimize $\frac{\mathbf{w}}{\alpha}$. If we change $(\frac{\mathbf{w}}{\alpha}, \frac{w_0}{\alpha})$ with $(\hat{\mathbf{w}}, \hat{w_0})$, we get that our objective is to minimize $\frac{1}{\alpha} = ||\hat{\mathbf{w}}||$ (norm of solution of $\mathbf{w}$ is 1, which we get from the previous part of the task) which is equivalent to the formulation in the lectures. w_0 is not part of the norm of $||\mathbf{w}||$, this guarantees that we could set it as much as we want to "lower" the influence of α on inequalities.

3.2 1. Maximum number of support vectors is N, because if all the points from one and the other set are located on two distinct coordinates then the support vectors will be formed from the full set of points. Minimum number of support vectors is 2. If we have all the aligned on the line, with points from one class on one side of the line, and the points from the other class on the other side of the line, then hyperplane that would separate the classes is limited by two "closest" points from different sets and they are at the same time support vectors.

2. First we will prove that an optimal solution in linearly separable formulation of the algorithm satisfies the formulation for non-separable case formulation. Setting up $\zeta_i = 0$, we get the same quadratic program as for linearly separable case, and an optimal solution is this solution because of the formulation of the task.
   Further, assume there is some solution where all $0 \leq \zeta_i < 1$. This leads to the case explained in the lectures for linearly separable algorithm. The idea is if we divide both sides of inequalities by $1 - \zeta_i$ we get the same case as when $\zeta_i = 0$. Just this time instead of $||\mathbf{w}_{linear}||$ it might be $\frac{||\mathbf{w}_{linear}||}{\zeta_i} \geq ||\mathbf{w}_{linear}||$, where $\mathbf{w}_{linear}$ is the optimal solution when $\zeta_i = 0$. But we assumed that $||\mathbf{w}_{linear}||$ is the optimal solution to the linearly separable case. Contradiction!
   Let us assume there is a $\zeta_i \geq 1$. But we could set up the constant C to be the $\frac{1}{2}||\mathbf{w}_{linear}||^2$ and the expression that we have to minimize will be at least $\frac{1}{2}||\mathbf{w}_{linear}||^2$ because

$$\sum_{i=1}^{N} C\zeta_i \geq \frac{1}{2}||\mathbf{w}_{linear}||^2$$

   However, this means we cannot minimize the expression more than we minimized with the previous solution. Thus, the algorithm will give the right solution if we set up C correctly.
3. Notice that the algorithm will output the tuple $(\mathbf{w}, w_0)$ that sets lhs of the inequality of 0 only when it misclassifies specific training example, thus $1 - \zeta_i \leq 0$. Further, more $\zeta_i \geq 1$. Summing all ζ_i we can notice this is the upper bound of the misclassified training examples, for the specific solution.

4. 1.

$$\begin{aligned}
\mathbf{w} = \mathbf{X}^{\mathrm{T}}\lambda^{-1}(\mathbf{y} - \mathbf{X}\mathbf{w}) &\iff \\
\mathbf{w} = \mathbf{X}^{\mathrm{T}}\lambda^{-1}\mathbf{y} - \mathbf{X}^{\mathrm{T}}\lambda^{-1}\mathbf{X}\mathbf{w} &\iff \quad (27)\\
(\mathbf{I}_D + \mathbf{X}^{\mathrm{T}}\lambda^{-1}\mathbf{X})\mathbf{w} = \mathbf{X}^{\mathrm{T}}\lambda^{-1}\mathbf{y} &\iff \quad (28)\\
(\lambda\mathbf{I}_D + \mathbf{X}^{\mathrm{T}}\mathbf{X})\mathbf{w} = \mathbf{X}^{\mathrm{T}}\mathbf{y} & \quad (29)
\end{aligned}$$

   The equation 27, 28, 29 because we use simple vector spaces of matrices properties (distributivity, division by scalar).

   2.

$$\begin{aligned}
(\mathbf{X}\mathbf{X}^{\mathrm{T}} + \lambda\mathbf{I}_D)\mathbf{X}^{\mathrm{T}}(\mathbf{X}\mathbf{X}^{\mathrm{T}} + \lambda\mathbf{I}_N)^{-1}\mathbf{y} = \mathbf{X}^{\mathrm{T}}\mathbf{y} &\iff \quad (30)\\
(\mathbf{X}\mathbf{X}^{\mathrm{T}} + \lambda\mathbf{I}_D)\mathbf{X}^{\mathrm{T}}(\mathbf{X}\mathbf{X}^{\mathrm{T}} + \lambda\mathbf{I}_N)^{-1} = \mathbf{X}^{\mathrm{T}} &\iff \quad (31)\\
(\mathbf{X}\mathbf{X}^{\mathrm{T}} + \lambda\mathbf{I}_D)\mathbf{X}^{\mathrm{T}} = \mathbf{X}^{\mathrm{T}}(\mathbf{X}\mathbf{X}^{\mathrm{T}} + \lambda\mathbf{I}_N) &\iff \quad (32)\\
\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}} + \lambda\mathbf{X}^{\mathrm{T}} = \mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{X}^{\mathrm{T}} + \lambda\mathbf{X}^{\mathrm{T}} &\iff \quad (33)\\
\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}} = \mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{X}^{\mathrm{T}} &\iff \quad (34)\\
\mathbf{X}\mathbf{X}^{\mathrm{T}} = \mathbf{X}^{\mathrm{T}}\mathbf{X} & \quad (35)
\end{aligned}$$

   The equation 30 is the starting assumption. The equation 31 is correct or $\mathbf{y}$ is zero vector. If y is zero vector we do not need to prove further; If y is not zero than we have this equation. The equation 32 is correct because $\mathbf{X}\mathbf{X}^{\mathrm{T}} + \lambda\mathbf{I}_N$ has determinant different than zero. The equations 33 and 34 use simple matrix calculations. The equation 35 is correct because matrix multiplication in this case produces symmetric matrix. Besides, if $\mathbf{X}$ was 0, this would still hold. Thus, our $\mathbf{w}$ satisfies the beginning condition and it is a solution of normal equation.