Venkata Pradeep Kadubandi
3011069060

Homework 3

CS 567
July 27, 2019

1 Problem

1.1 Solution

The slack variables ζ_i^+, ζ_i^- allow the prediction $f(\mathbf{x}) = \hat{y}$ to be less than $y_i - \epsilon$ and greater than $y_i + \epsilon$ respectively for each data point in N training samples. All the constraints written in the form $h_j(x) < 0$ are (for $i \in [1, N]$) :

$$\begin{aligned} -\zeta_i^+ &\leq 0 \\ -\zeta_i^- &\leq 0 \\ y_i - \hat{y}_i - \epsilon - \zeta_i^+ &\leq 0 \\ \hat{y}_i - y_i - \epsilon - \zeta_i^- &\leq 0 \\ \text{where } \hat{y}_i = f(\mathbf{x}_i) = &\mathbf{w}^T\phi(\mathbf{x}_i) + b \end{aligned}$$

1.2 Solution

Intuitively, we can see that when $\hat{y}_i > y_i + \epsilon$, we expect $\zeta_i^- > 0, \zeta_i^+ = 0$. In this case, $E_\epsilon(\hat{y}_i, y_i) = \zeta_i^-$. Similarly, when $\hat{y}_i < y_i - \epsilon$, we expect $\zeta_i^- = 0, \zeta_i^+ > 0$ and $E_\epsilon(\hat{y}_i, y_i) = \zeta_i^+$. In all other scenarios where $y_i - \epsilon \leq \hat{y}_i \leq y_i + \epsilon$, $\zeta_i^+ = \zeta_i^- = E_\epsilon(\hat{y}_i, y_i) = 0$. A common expression from all these scenarios for $E_\epsilon(\hat{y}_i, y_i) = \zeta_i^+ + \zeta_i^-$.

Replacing the above in the minimization objective and adding the Lagrange multipliers for the constraints from section 1.1, our Lagrangian is:

$$\begin{aligned} L(\mathbf{w}, b, \zeta_i^+, \zeta_i^-, \lambda_i^+, \lambda_i^-, \alpha_i^+, \alpha_i^-) = C\sum_{i=1}^{N}(\zeta_i^+ + \zeta_i^-) + \frac{1}{2}\|\mathbf{w}\|^2 \\ - \sum_{i=1}^{N}\lambda_i^+\zeta_i^+ - \sum_{i=1}^{N}\lambda_i^-\zeta_i^- \\ + \sum_{i=1}^{N}\alpha_i^+(y_i - \hat{y}_i - \epsilon - \zeta_i^+) \\ + \sum_{i=1}^{N}\alpha_i^-(\hat{y}_i - y_i - \epsilon - \zeta_i^-) \end{aligned}$$

where all the Lagrangian multipliers are non-negative i.e., $\forall i \in [1, N], \lambda_i^+ \geq 0, \lambda_i^- \geq 0, \alpha_i^+ \geq 0, \alpha_i^- \geq 0$. Note that $\hat{y}_i = f(\mathbf{x}_i) = \mathbf{w}^T\phi(\mathbf{x}_i) + b$.

1.3 Solution

Taking the partial derivatives w.r.t. primal variables $\mathbf{w}, b, \zeta_i^+, \zeta_i^-$ and setting them to zero, we get (All summations are from 1 to N):

$$\frac{\partial L}{\partial w} = 0 \Rightarrow \mathbf{w} - \sum_i \alpha_i^+ \phi(\mathbf{x}_i) + \sum_i \alpha_i^- \phi(\mathbf{x}_i) = 0 \text{ i.e.,}$$

$$\mathbf{w} = \sum_i (\alpha_i^+ - \alpha_i^-) \phi(\mathbf{x}_i)$$

$$\frac{\partial L}{\partial b} = 0 \Rightarrow \sum_i (\alpha_i^- - \alpha_i^+) = 0$$

$$\frac{\partial L}{\partial \zeta_i^+} = 0 \Rightarrow C - \lambda_i^+ - \alpha_i^+ = 0$$

$$\frac{\partial L}{\partial \zeta_i^-} = 0 \Rightarrow C - \lambda_i^- - \alpha_i^- = 0$$

Using the above we can simplify Lagrangian:

$$
\begin{aligned}
L(\mathbf{w}, b, \zeta_i^+, \zeta_i^-, \lambda_i^+, \lambda_i^-, \alpha_i^+, \alpha_i^-) =& \frac{1}{2}\|\mathbf{w}\|^2 + \sum_i (C - \lambda_i^+ - \alpha_i^+)\zeta_i^+ + \sum_i (C - \lambda_i^- - \alpha_i^-)\zeta_i^- \\
&+ \sum_i (\alpha_i^+ - \alpha_i^-) y_i - \epsilon \sum_i (\alpha_i^+ + \alpha_i^-) \\
&- \sum_i (\alpha_i^+ - \alpha_i^-) \hat{y}_i \quad \text{...rearranging terms} \\
=& \frac{1}{2}\|\mathbf{w}\|^2 + \sum_i (\alpha_i^+ - \alpha_i^-) y_i - \epsilon \sum_i (\alpha_i^+ + \alpha_i^-) \\
&- \sum_i (\alpha_i^+ - \alpha_i^-) \hat{y}_i \quad \text{...partial derivative results for slack variables} \\
=& \frac{1}{2}\mathbf{w}^T\mathbf{w} - \sum_i (\alpha_i^+ - \alpha_i^-)(\mathbf{w}^T \phi(\mathbf{x}_i) + b) \\
&+ \sum_i (\alpha_i^+ - \alpha_i^-) y_i - \epsilon \sum_i (\alpha_i^+ + \alpha_i^-) \quad \text{...rewriting } \hat{y}_i \text{ and norm} \\
=& \frac{1}{2} \left\{ \sum_i (\alpha_i^+ - \alpha_i^-)\phi(\mathbf{x}_i) \right\} \left\{ \sum_j (\alpha_j^+ - \alpha_j^-)\phi(\mathbf{x}_j) \right\} \\
&- \sum_i (\alpha_i^+ - \alpha_i^-)(\phi(\mathbf{x}_i) \left\{ \sum_j (\alpha_j^+ - \alpha_j^-)\phi(\mathbf{x}_j) \right\} + b) \\
&+ \sum_i (\alpha_i^+ - \alpha_i^-) y_i - \epsilon \sum_i (\alpha_i^+ + \alpha_i^-) \quad \text{...substituting } \mathbf{w} \\
=& -\frac{1}{2} \sum_i \sum_j (\alpha_i^+ - \alpha_i^-)(\alpha_j^+ - \alpha_j^-)\phi(\mathbf{x}_i)\phi(\mathbf{x}_j) \\
&+ \sum_i (\alpha_i^+ - \alpha_i^-)(y_i - b) - \epsilon \sum_i (\alpha_i^+ + \alpha_i^-) \quad \text{...substitute, simplify} \\
=& -\frac{1}{2} \sum_i \sum_j (\alpha_i^+ - \alpha_i^-)(\alpha_j^+ - \alpha_j^-) k(\mathbf{x}_i, \mathbf{x}_j) \\
&+ \sum_i (\alpha_i^+ - \alpha_i^-) y_i - \epsilon \sum_i (\alpha_i^+ + \alpha_i^-) \quad \text{...kernel trick}
\end{aligned}
$$

Since we have $\alpha_i^+ \geq 0$, $\lambda_i^+ \geq 0$ and $C = \alpha_i^+ + \lambda_i^+$, we can deduce that $0 \leq \alpha_i^+ \leq C$, same logic goes for α_i^- variables as well.

So the final expression for dual is:

$$L(\alpha_i^+, \alpha_i^-) = -\frac{1}{2}\sum_i \sum_j (\alpha_i^+ - \alpha_i^-)(\alpha_j^+ - \alpha_j^-)k(\mathbf{x}_i, \mathbf{x}_j) \\ + \sum_i (\alpha_i^+ - \alpha_i^-)y_i - \epsilon \sum_i (\alpha_i^+ + \alpha_i^-)$$

subject to

$$0 \leq \alpha_i^+ \leq C \\ 0 \leq \alpha_i^- \leq C$$

1.4 Solution

Using the value for $\mathbf{w}$ in the expression for $f(\mathbf{x})$, we get:

$$f(\mathbf{x}) = \mathbf{w}^T\phi(\mathbf{x}) + b = \sum_i (\alpha_i^+ - \alpha_i^-)\phi(\mathbf{x}_i)\phi(\mathbf{x}) + b \\ = \sum_i (\alpha_i^+ - \alpha_i^-)k(\mathbf{x}_i, \mathbf{x}) + b$$

1.5 Solution

Using the complimentary slackness conditions for $\alpha_i^+, \alpha_i^-, \lambda_i^+, \lambda_i^-$ with substitution for latter, $\forall i \in [1, N]$ we will have:

$$\alpha_i^+(y_i - \hat{y}_i - \epsilon - \zeta_i^+) = 0 \\ \alpha_i^-(\hat{y}_i - y_i - \epsilon - \zeta_i^-) = 0 \\ (C - \alpha_i^+)\zeta_i^+ = 0 \\ (C - \alpha_i^-)\zeta_i^- = 0$$

Given the slack variables ζ_i^+ and ζ_i^- are non-negative $\forall i$, for any point $\mathbf{x}_i$ such that $y_i - \epsilon < \hat{y}_i < y_i + \epsilon$ (note the strict inequality) both the terms $y_i - \hat{y}_i - \epsilon - \zeta_i^+$ and $\hat{y}_i - y_i - \epsilon - \zeta_i^-$ are less than 0, so both the corresponding α_i^+ and α_i^- values must be 0 (from first two conditions). So all such training points do not participate in the prediction and need not be stored. Also from the last two conditions, it follows that corresponding slack variables ζ_i^+, ζ_i^- for such variables must be 0 (since $C \neq 0$).

So a necessary condition for a point to be a support vector is either $\hat{y}_i \geq y_i + \epsilon$ or $\hat{y}_i \leq y_i - \epsilon$. Also, in the case where $\hat{y}_i \geq y_i + \epsilon$, while α_i^- can be non-zero from second condition, we can see that clearly $y_i - \hat{y}_i - \epsilon - \zeta_i^+ \leq -2\epsilon - \zeta_i^+ < 0$, so α_i^+ for this point must be 0 from the first condition. With a similar argument, we can see that when $\hat{y}_i \leq y_i - \epsilon$, α_i^+ can be non-zero but α_i^- must be 0. So for any training point only one of α_i^+ or α_i^- can be non-zero. Finally from the last two conditions, we can also validate our initial intuition that, when $\hat{y}_i \geq y_i + \epsilon$, ζ_i^+ must be 0 and when $\hat{y}_i \leq y_i - \epsilon$, ζ_i^- must be 0.

2 Problem

2.1 Solution

From the results of problem 2.2, it's pretty clear that all 3 points are support vectors because all of them lie on the margin from hyper plane. (The absolute distance from hyper plane is 1 for all points).

Graphically, if we plot the $\phi(x)$ for all 3 points (either in a 3-D system OR just 2-D system with the last 2 values as first value 1 is common for all the points), we can see that the positive labelled point (for $x_1 = 0$) and the two negative labelled points are linearly separable. Also, we can guess that the maximum margin classifier will be parallel to the line/plane between the two negative points. If we imagine a parallel line/plane going through the positive point, we can see that the maximum margin hyper plane will be right in the middle of these 2 hyper planes (because the distance between given negative points and the one positive point are equal). Hence all 3 points will lie on the corresponding margin line/planes for positive line/plane and negative line/planes accordingly and hence all 3 of them will be support vectors.

2.2 Solution

The Lagrangian formulation for the optimization problem is:

$$L(\mathbf{w}, b, \{\lambda_i\}) = \frac{1}{2}\|\mathbf{w}\|_2^2 + \sum_{i=1}^{3} \lambda_i(1 - y_i\hat{y_i}) \quad \text{...where } \hat{y_i} = \mathbf{w}^T\phi(x_i) + b$$

Using the stationary conditions from KKT, we get:

$$\frac{\partial L}{\partial \mathbf{w}} = 0 \Rightarrow \mathbf{w} - \sum_{i=1}^{3} \lambda_i y_i \phi(x_i) = 0$$

$$\frac{\partial L}{\partial b} = 0 \Rightarrow \sum_{i} \lambda_i y_i = 0 \Rightarrow \lambda_1 - \lambda_2 - \lambda_3 = 0$$

$$\frac{\partial L}{\partial \lambda_i} = 0 \Rightarrow 1 - y_i\hat{y_i} = 0 \quad \forall i \in [1, 3]$$

By putting the given data points and using $\mathbf{w} = [w_1, w_2, w_3]^T$, we will get:

$$\begin{aligned}
\text{Note : } \phi(x_1) = [1,0,0]^T, \phi(x_2) =& [1, -\sqrt{2}, 1]^T, \phi(x_3) = [1, \sqrt{2}, 1]^T \\
\text{for the last equality } i = 1 \Rightarrow & w_1 + b = 1 \\
i = 2 \Rightarrow & -w_1 + \sqrt{2}w_2 - w_3 - b = 1 \\
i = 3 \Rightarrow & -w_1 - \sqrt{2}w_2 - w_3 - b = 1 \\
\text{...from 2 and 3, } w_2 = 0; & \text{ ...from 1 and 2, } w_3 = -2; \\
\text{Using } \frac{\partial L}{\partial w_1} = 0 \Rightarrow & w_1 - \lambda_1 + \lambda_2 + \lambda_3 = 0 \\
\text{and since } \lambda_1 - \lambda_2 - \lambda_3 = 0 \Rightarrow & w_1 = 0 \\
\text{...and from } w_1 + b = 1 \Rightarrow & b = 1 \\
\text{Therefore...} \quad \mathbf{w}^* =& [0, 0, -2]^T; b^* = 1
\end{aligned}$$

Finally the margin is given by distance of support vectors from the hyper plane which is:

$$l = \frac{\left|\mathbf{w}^T\phi(x) + b\right|}{\|\mathbf{w}\|_2} = \frac{1}{\|\mathbf{w}^*\|_2}$$

because $\left|\mathbf{w}^{*T}\phi(x) + b\right| = 1$ for all given data points (support vectors).

3 Problem

3.1 Solution

The given loss function:

$$\begin{aligned}
L_t &= (e^{\beta_t} - e^{-\beta_t}) \sum_n w_t(n) \mathbb{1}[y_n \neq h_t(\mathbf{x}_n)] + e^{-\beta_t}) \sum_n w_t(n) \\
&= \frac{1}{\sum_n w_t(n)} [(e^{\beta_t} - e^{-\beta_t})\epsilon_t + e^{-\beta_t})] \quad \text{...from } \epsilon_t = \frac{\sum_n w_t(n) \mathbb{1}[y_n \neq h_t(\mathbf{x}_n)]}{\sum_n w_t(n)}
\end{aligned}$$

For minimizing the above loss, we take derivative w.r.t. β_t and set it to zero:

$$\begin{aligned}
\frac{\partial L_t}{\partial \beta_t} = 0 &\Rightarrow (e^{\beta_t} + e^{-\beta_t})\epsilon_t - e^{-\beta_t} = 0 \\
&\Rightarrow e^{2\beta_t}\epsilon_t = 1 - \epsilon_t \\
&\Rightarrow \beta_t = \frac{1}{2} log \ \frac{1 - \epsilon_t}{\epsilon_t}
\end{aligned}$$

3.2 Solution

Writing the error function as product of terms, we will get:

$$\begin{aligned}
\epsilon_{exp} &= \sum_n e^{-y_n\hat{y_n}} = -log\ e^{-\sum_n e^{-y_n\hat{y_n}}} \\
&= -log\ \prod_n e^{-e^{-y_n\hat{y_n}}} \\
i.e., p(y_n|\mathbf{x}_n) &\propto e^{-e^{-y_n\hat{y_n}}} \quad \text{...after normalization assuming binary classification, we get} \\
p(y_n|\mathbf{x}_n) &= \frac{e^{-e^{-y_n\hat{y_n}}}}{e^{-e^{-y_n\hat{y_n}}} + e^{-e^{y_n\hat{y_n}}}} \quad \text{...using this to calculate loss, we get} \\
\epsilon_{new} &= -log\ \prod_n \frac{e^{-e^{-y_n\hat{y_n}}}}{e^{-e^{-y_n\hat{y_n}}} + e^{-e^{y_n\hat{y_n}}}} \\
&= \sum_n -log \frac{e^{-e^{-y_n\hat{y_n}}}}{e^{-e^{-y_n\hat{y_n}}} + e^{-e^{y_n\hat{y_n}}}} \\
&= \sum_n log(1 + e^{e^{-y_n\hat{y_n}} - e^{y_n\hat{y_n}}})
\end{aligned}$$

which is not the same as given error function. So the given error function does not correspond to the log likelihood of any well-behaved probabilistic model. (Though the proof was done for binary classification, it can be easily extended to multi-class by taking sum of terms for k classes instead of 2 classes).

3.3 Solution

In a similar way as above:

$$\begin{aligned}
\epsilon_{log} &= \sum_n log\ (1 + e^{-2y_n\hat{y_n}}) = -log\ e^{-\sum_n log\ (1+e^{-2y_n\hat{y_n}})} \\
&= -log\ \prod_n e^{-log\ (1+e^{-2y_n\hat{y_n}})} = -log\ \prod_n \frac{1}{1+e^{-2y_n\hat{y_n}}} \\
\text{from this, we have: } p(y_n|\mathbf{x}_n) &\propto \frac{1}{1+e^{-2y_n\hat{y_n}}} \quad \text{...and after normalization, we get} \\
p(y_n|\mathbf{x}_n) &= \frac{\frac{1}{1+e^{-2y_n\hat{y_n}}}}{\frac{1}{1+e^{-2y_n\hat{y_n}}} + \frac{1}{1+e^{2y_n\hat{y_n}}}} \quad \text{...denominator value is 1, so} \\
&= \frac{1}{1+e^{-2y_n\hat{y_n}}} = \frac{e^{y_n\hat{y_n}}}{e^{y_n\hat{y_n}} + e^{-y_n\hat{y_n}}}
\end{aligned}$$

For validating if we put the expression back in likelihood for calculating error function:

$$
\begin{aligned}
\epsilon_{new} &= -log \prod_n p(y_n|\mathbf{x}_n) = -log \prod_n \frac{1}{1 + e^{-2y_n \hat{y_n}}} \\
&= \sum_n -log \frac{1}{1 + e^{-2y_n \hat{y_n}}} \\
&= \sum_n log \ (1 + e^{-2y_n \hat{y_n}})
\end{aligned}
$$

which is same as ϵ_{log} we started with.