Machine Learning Assignment 5

B08611035

December 2020

1 Hard-Margin SVM and Large Margin

(1) **d.** $w_1^* = 0$

$$\phi(\mathbf{x}) = \begin{bmatrix} 1 & x & x^2 \end{bmatrix}^T$$

tell that the hyper-plane is $w_0^* + w_1^* x_n + w_2^* x^2 = f(x)$ where $w_0^* = b^*$. Thus

$$\min_{b,\mathbf{w}} \frac{1}{2}\mathbf{w}^{*\mathbf{T}}\mathbf{w}^*$$

subject to the constraint $y_n(w_0^* + w_1^* x_n + w_2^* x^2) \geq 1$ for $n = 1, 2, 3,\ w^* = [w_1^*, w_2^*]$. Observe $y_n(w_0^* + w_1^* x_n + w_2^* x^2) \geq 1$.

$$\begin{cases} w_1^* - 2w_2^* \geq 1 \\ -b^* - 4w_2^* \geq 1 \\ -w_1^* - 2w_2^* \geq 1 \end{cases}$$

Thus, $-4w_2^* \geq 2 \Rightarrow w_2^* \leq -1/2$. If $w_2^* \leq 1/2$ and $b^* \geq 1$, then w_1^* must equal to zero to holds the inequality. So that we can find that:

$$w_1^* = 0$$

(2) **b.** 2

Use the results in problem (1), take $w_1^* = 0$, $w_2^* = -1/2$ and $b^* = 1$, then we could get the hyper-plane:

$$1 - \frac{1}{2}x^2 = f(x)$$

$$\Rightarrow \text{margin}(b^*, \mathbf{w}^*) = \frac{1}{||\mathbf{w}||} = \frac{1}{\sqrt{\frac{1}{4}}} = 2$$

(3) **e.** $(x_{m+1} - x_m)/2$

$[(\mathbf{x}_n, y_N)]_n^N = 1$ for $\mathbf{x}_n \in \mathbb{R}$ for one dimensional examples

$$x_1 \le 1_2 \le \cdots \le x_m < x_{m+1} \le \cdots \le x_N$$

Because the examples are linear separable, the lines must pass through some where between x_m, x_{m+1}. And since the linear margin is symmetric to the hyper-plane, the largest symmetric margin is the middle between x_m and x_{m+1}. Thus, the margin will be:

$$\frac{1}{2}(x_{m+1} - x_m)$$

(4) **a.** $2 + 2(1-2\rho)^2$

For case 1, if $|x_1 - x_2| < 2\rho$, then we might use the line with margin at least ρ to separate it. But the line is either at right side of both x_1 and x_2; or the left side of both x_1 and x_2.

In case 2, if $|x_1 - x_2| \ge 2\rho$, then we have $m_{\mathcal{H}} = 2^2 = 4$. If $x_1 < x_2$ or $x_2 < x_1$, then

$$E(x_2) = E(1 - x_1 - 2\rho) = \int_0^{1-2\rho} (1 - x_1 - 2\rho)dx_1 = \frac{(1-2\rho)^2}{2}$$

$$\Rightarrow E(m_{\mathcal{H}}) = 2 + 4\Big(\frac{(1-2\rho)^2}{2}\Big) = 2 + 2(1-2\rho)^2$$

2 Dual Problem of Quadratic Programming

(5) **c.**

$$\min_{b,\mathbf{w}} \frac{1}{2}\mathbf{w}^{\mathbf{T}}\mathbf{w}$$

subject to constraint

$$\begin{cases} y_n(\mathbf{w}^{\mathbf{T}}\mathbf{x}_n + b) \ge \rho_+ & \text{for } n \text{ s.t. } y_n = +1 \\ y_n(\mathbf{w}^{\mathbf{T}}\mathbf{x}_n + b) \ge \rho_- & \text{for } n \text{ s.t. } y_n = -1 \end{cases}$$

Lagrange function with multiple α_n:

$$\mathcal{L}(b, \mathbf{w}, \alpha) = \frac{1}{2}\mathbf{w}^{\mathbf{T}}\mathbf{w} + \sum_n \alpha_n [[y_n = +1]](\rho_+ - y_n(\mathbf{w}^{\mathbf{T}}\mathbf{x}_n + b)) + \sum_n \alpha_n [[y_n = -1]](\rho_- - y_n(\mathbf{w}^{\mathbf{T}}\mathbf{x}_n + b))$$

solving Lagrange dual with:

$$\max_{\alpha_n \ge 0}\left(\min_{b,\mathbf{w}} \mathcal{L}(b, \mathbf{w}, \alpha)\right)$$

$$\frac{\partial \mathcal{L}}{\partial b} = 0 \Rightarrow \sum_n \alpha_n y_n \Big([[y_n = +1]] + [[y_n = -1]]\Big) = 0 \Rightarrow \sum_n \alpha_n y_n = 0$$

Therefore, now the dual is:

$$\max_{\alpha_n \geq 0, \sum \alpha_n y_n = 0} \left(\min_{\mathbf{w}} \mathcal{L}(\mathbf{w}, \alpha) \right)$$

$$\frac{\partial \mathcal{L}}{\partial w_i} = 0 = w_i - \sum \alpha_n [[y_n = +1]] y_n x_{n,i} + \sum \alpha [[y_n = -1]] y_n x_{n,i}$$

$$\Rightarrow \mathbf{w} = \sum \alpha_n y_n [[y_n = +1]] x_n + \sum \alpha_n [[y_n = -1]] y_n x_n$$

Thus,

$$\mathcal{L}(\mathbf{w}, \alpha) = -\frac{1}{2} \mathbf{w}^{\mathbf{T}} \mathbf{w} + \sum \alpha_n [[y_n = +1]] \rho_+ + \sum \alpha_n [[y_n = -1]] \rho_-$$

$$\Rightarrow \min_{\alpha} \frac{1}{2} \sum_n \sum_n \alpha_n \alpha_m y_n y_m \mathbf{x_n^T} \mathbf{x_m} - \sum_n [[y_n = +1]] \rho_+ \alpha_n - \sum_n [[y_n = -1]] \rho_- \alpha_n$$

Thus, the answer is:

$$-\sum_{n=1}^{N} \rho_+ [[y_n = +1]] \alpha_n - \sum_{n=1}^{N} \rho_- [[y_n = -1]] \alpha_n$$

(6) **e.**

In dual-inner optimal, we have that $\sum y_n \alpha_n = 0$. Thus, we have

$$\sum_n \alpha_n y_n = \sum_n \alpha_n [[y_n = +1]] - \sum_n \alpha_n [[y_n = -1]] = 0$$

$$\Rightarrow \sum_n \alpha_n [[y_n = +1]] = \sum_n \alpha_n [[y_n = -1]]$$

And we also have $\sum \alpha_n = 2 \sum \alpha [[y_n = +1]]$. By

$$\sum_n \alpha_n^{new} = 2 \sum_n \alpha_n^{new} [[y_n = +1]]$$

$$\Rightarrow \sum_n \alpha_n^{new} = 2 \sum_n \alpha_n^{new} [[y_n = +1]] = \sum_n (\rho_+ + \rho_-) \alpha_n [[y_n = +1]]$$

$$2\alpha_n^{new} = (\rho_+ + \rho_-) \alpha_n^*$$

Thus,

$$\alpha_n^{new} = \frac{(\rho_+ + \rho_-)}{2} \alpha^* \forall n$$

3 Properties of Kernels

(7) **d.**

For (a), (b), (c) and (e) options, both eigenvalue in each option is greater than zero. Thus, option (d) is not always a valid kernel.

(8) **c.** 2

$$K(\mathbf{x}, \mathbf{x}') = \exp(-\gamma||\phi(\mathbf{x}) - \phi(\mathbf{x}')||^2)$$

If $\phi(\mathbf{x}) = \phi(\mathbf{x}')$, then $\phi(\mathbf{x})^t\phi(\mathbf{x}) = K(\mathbf{x}, \mathbf{x}) = 1, \forall\gamma > 0$.

$$||\phi(\mathbf{x}) - \phi(\mathbf{x}')||^2 = \phi(\mathbf{x})^t\phi(\mathbf{x}) - 2\phi(\mathbf{x})^t\phi(\mathbf{x}') + \phi(\mathbf{x}')^t\phi(\mathbf{x}')$$

$$= K(\mathbf{x}, \mathbf{x}) - 2K(\mathbf{x}, \mathbf{x}') + K(\mathbf{x}', \mathbf{x}') = 1 + 1 - 2\exp(-\gamma||\phi(\mathbf{x}) - \phi(\mathbf{x}')||^2)$$

which $2\exp(-\gamma||\phi(\mathbf{x}) - \phi(\mathbf{x}')||^2) \in [0, 2]$. Thus,

$$0 \leq ||\phi(\mathbf{x}) - \phi(\mathbf{x}')||^2 \leq 2$$

(9) **d.**

$$\alpha = 1, b = 0 \Rightarrow h_{1,0}(\mathbf{x}) = h(\mathbf{x}) = \text{sign}\Big(\sum_n y_n K(\mathbf{x}_n, \mathbf{x})\Big)$$

$$\Rightarrow E_{in}(\hat{h}) = \frac{1}{N}\sum_i [[h(\mathbf{x}_i) \neq y_i]] = \frac{1}{N}\sum_i [[\text{sign}(\sum y_n K(x_n, x_i)) \neq y_i]]$$

Thus, $E_{in}(\hat{h}) = 0$ if $\text{sign}(\sum y_n K(x_n, x_i)) = y_i, \forall i$

$$\Rightarrow \sum_i \sum_n y_n K(x_n, x_i) \leq \sum_n Y_i + \sum Y_i(N-1)\exp(-\gamma\epsilon^2)$$

because $||x - x'|| \geq \epsilon^2$, and we need:

$$\sum_i \sum_n y_n K(x_n, k_i) \to \sum_i y_i$$

Thus, $\sum Y_i \geq \sum Y_i(N-1)\exp(-\gamma\epsilon^2)$ as γ larger enough.

$$\frac{1}{N-1} \geq \exp(-\gamma\epsilon^2) \Rightarrow \ln(N-1) \leq \gamma\epsilon^2$$

$$\Rightarrow \gamma \geq \frac{\ln(N-1)}{\epsilon^2}$$

4 Kernel Perceptron Learning Algorithm

(10) **c.**

When the current $\mathbf{w}_t$ makes a mistake on $(\phi(\mathbf{x}_{n(t)}), y_{n(t)}$, update $\mathbf{w}_t$ to $\mathbf{w}_{t+1}$.

$$\mathbf{w}_{t+1} = \mathbf{w}_t + y_{n(t)}\phi(\mathbf{x}_{n(t)}) = \sum_i \alpha_{t,i}\phi(\mathbf{x}_i) + y_{n(t)}\phi(\mathbf{x}_{n(t)})$$

(11) **a.**

$$\mathbf{w}_t = sum_n \alpha_{t,n}\phi(\mathbf{x}_n) \Rightarrow w_t^T = \sum_n \alpha_{t,n}\phi(\mathbf{x}_n^T)$$

$$\mathbf{w}^T\phi(\mathbf{x}) = \Big(\sum_n \alpha_{t,n}\phi(\mathbf{x}_n)^T\Big)\phi(\mathbf{x}) = \sum_n \alpha_{t,n}\phi(\mathbf{x}_n)^T\phi(\mathbf{x})$$

$$\Rightarrow \mathbf{w}^T\phi(\mathbf{x})\sum_{n=1}^{N}\alpha_{t,n}K(\mathbf{x}_n, \mathbf{x})$$

5 Soft-Margin SVM

(12) **b.**

Consider the complementary slackness and the condition below:

$$\alpha_n(1 - \xi_n - y_n(\mathbf{w}^\mathbf{T}\mathbf{z}_n + b)) = 0, \xi_n \geq 0$$

And due to we get that optimal α^* so that $\alpha_n^* \forall n$

$$\alpha^*(1 - \xi_n - y_n(\mathbf{w}^\mathbf{T}\mathbf{z}_n + b^*)) = 0 \Rightarrow 1 - \xi_n - y_n(\mathbf{w}^\mathbf{T}\mathbf{z}_n + b^*) = 0$$

$$\xi_n = 1 - y_n(\mathbf{w}_{\mathbf{z}}^{\mathbf{T}}{}_n + b^*) \geq 0 \Rightarrow 1 \geq y_n(\mathbf{w}^\mathbf{T}\mathbf{z}_n + b^*)$$

Here, it's going to find upper bound of b^*, thus let's consider if $y_n > 0$,

$$b^* \leq \frac{1}{y_n} - \mathbf{w}^\mathbf{T}\mathbf{z}_n = \frac{1}{y_n} - \sum_m y_m\alpha_m k(\mathbf{x}_n, \mathbf{x}_m)$$

Thus, b^* has n choices for $n = 1, 2, \cdots, N$:

$$b^* = \min_{n:y_n>0}\Big(1 - \sum_{m=1}^{M} y_m\alpha_m K(\mathbf{x}_n, \mathbf{x}_m)\Big)$$

(13) **e.**

$$\min_{b,\mathbf{w},\xi}\frac{1}{2}\mathbf{w}^\mathbf{T}\mathbf{w} + C\cdot\sum_n \xi_n^2$$

subject to the constraint:

$$y_n(\mathbf{w}^T\phi(\mathbf{x}_n)+b)\geq 1-\xi_n$$

Give into Lagrange multiplier for the constraint optimization problems

$$\mathcal{L}(b,\mathbf{w},\xi,\alpha)=\frac{1}{2}\mathbf{w}^{\mathbf{T}}\mathbf{w}+C\cdot\sum_n\xi_n^2+\sum_n\alpha_n\Big(1-\xi_n-y_n(\mathbf{w}^{\mathbf{T}}\phi(\mathbf{x}_n)+b)\Big)$$

and we need

$$\max_{\alpha\geq 0}\Big(\min_{b,\mathbf{w},\xi}\mathcal{L}(b,\mathbf{w},\xi,\alpha)\Big)$$

subject to the condition that:

$$\frac{\partial\mathcal{L}}{\partial b}-0=-\sum_n\alpha_n y_n\Rightarrow\sum_n\alpha_n y_n=0$$

Thus,

$$\mathcal{L}(b,\mathbf{w},\xi,\alpha)=\mathcal{L}(\mathbf{w},\xi,\alpha)=\frac{1}{2}\mathbf{w}^{\mathbf{T}}\mathbf{W}+C\cdot\sum_n\xi_n^2+\sum_n\alpha_n\Big(1-\xi_n-y_n(\mathbf{w}^{\mathbf{T}}\phi(\mathbf{w}_n))\Big)$$

and for ξ to $\mathcal{L}$:

$$\frac{\partial\mathcal{L}}{\partial\xi_n}=0=2C\xi_n-\alpha_n\Rightarrow\xi_n=\frac{\alpha_n}{2C}$$

Get into the Lagrange expression, and also for the weight:

$$\frac{\partial\mathcal{L}}{\partial w_i}=0\Rightarrow\mathbf{w}=\sum_n\alpha_n y_N\phi(\mathbf{x}_n)$$

Thus,

$$\max_\alpha\mathcal{L}(\alpha)\Rightarrow\min_\alpha-\mathcal{L}(\alpha)=\min_\alpha\frac{1}{2}\sum_n\sum_m\alpha_n\alpha_m y_n y_m\Big(\phi(\mathbf{x}_n)^T\phi(x\mathbf{x}_m)+\frac{1}{2}[[n=m]]\Big)+\sum_n\alpha_n$$

$$=\min_\alpha\frac{1}{2}\sum_n\sum_m\alpha_n\alpha_m y_n y_m\Big(K(\mathbf{x}_N,\mathbf{x}_m)+\frac{1}{2C}[[n=m]]\Big)+\sum_n\alpha_n$$

subject to $\sum_n y_n\alpha_n=0$, $\alpha_n\geq 0$ for $n=1,2,\cdots,N$, Thus, answer is:

$$K(\mathbf{x}_N,\mathbf{x}_m)+\frac{1}{2C}[[n=m]]$$

(14) **e.** $\xi^* = \alpha^*/2C$

(P_2) written in Lagrange dual form: $\mathcal{L}(b, \mathbf{w}, \xi, \alpha)$, and find the best ξ^* that

$$\frac{\partial \mathcal{L}}{\partial \xi_n} = 0 = \xi_n = \frac{\alpha_n}{2C}$$

$$\Rightarrow \xi \begin{bmatrix} \xi_1 \\ \vdots \\ \xi_n \end{bmatrix} = \frac{1}{2C} \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix} = \frac{1}{2C}\alpha$$

and the optimal α is α^*, thus the optimal ξ would be:

$$\xi^* = \frac{1}{2C}\alpha^*$$

6 Programming*

(15) **d.** 8.5

(16) **b** "2" versus "not 2"

(17) **c.** 700

(18) **d.** 10

(19) **b.** 1

(20) **b.** 1