Machine Learning Techniques HW 5

2020 Fall Semester, NTU | R09946006 | 何青儒 | HO, Ching-Ru | Dec 24th, 2020

1. [**d**] $w_1^* = 0$
    - The optimal function and restriction: $\begin{cases} \min \dfrac{1}{2}\mathbf{w}^T\mathbf{w} \\ \text{subject to } y_n(\mathbf{w}^T \cdot \phi(\mathbf{x}_n) + b) \geq 1,\ n = 1, 2, 3 \end{cases}$.
    - Noticed that due to $\phi(\mathrm{x}) = [1, x, x^2]^T$, we have $\mathbf{w} = [w_1, w_2]$ and $w_0 = b$.
    - For each sample:
        - $\begin{cases} n = 1,\ (-2, -1) \xrightarrow{\phi(\cdot)} ([1, -2, 4]^T, -1) \xrightarrow{\text{substitute into}} -1(-2w_1 + 4w_2 + b) \geq 1 \\ n = 2,\ (0, +1) \xrightarrow{\phi(\cdot)} ([1, 0, 0]^T, +1) \xrightarrow{\text{substitute into}} +1(b) \geq 1 \\ n = 3,\ (2, -1) \xrightarrow{\phi(\cdot)} ([1, 2, 4]^T, -1) \xrightarrow{\text{substitute into}} -1(2w_1 + 4w_2 + b) \geq 1 \end{cases}$.
        - $\begin{cases} 2w_1 - 4w_2 - b \geq 1 \cdots\cdots (1) \\ b \geq 1 \cdots\cdots (2) \\ -2w_1 - 4w_2 - b \geq 1 \cdots\cdots (3) \end{cases}$.
        - $\begin{cases} \text{From (1)\&(2), } w_1 - 2w_2 \geq 1 \cdots\cdots (4) \\ \text{From (2)\&(3), } -w_1 - 2w_2 \geq 1 \cdots\cdots (5) \\ \text{From (1)\&(3), } -4w_2 - b \geq 1 \\ \text{From (4)\&(5), } w_2 \leq -\frac{1}{2} \end{cases}$.
        - Substituted $w_2 \leq -\frac{1}{2}$ into (4) and (5), $\begin{cases} (4) :\ w_1 \geq 0 \\ (5) :\ w_1 \leq 0 \end{cases}$.
    - Union both condition, $w_1^* = 0$, then $w_2^* \leq -\frac{1}{2}$ and $b^* \leq 1$.
2. [**b**] 2
    - Use the marginal value in the previous inequality, let $w_1^* = 0, w_2^* = -\frac{1}{2}, b = 1$, the function of separation line of SVM is $\left(-\dfrac{1}{2}x^2 + 1\right)$, and the margin is $\dfrac{1}{||\mathbf{w}||} = \dfrac{1}{\sqrt{0^2 + \frac{1}{2}^2}} = \dfrac{1}{\sqrt{\frac{1}{4}}} = 2$.
3. [**e**] $\dfrac{1}{2}(x_{M+1} - x_M)$
    - Due to the samples are ordered, the "perfect" separation line of SVM should be between x_M (the largest negative sample) and X_{M+1} (the smallest positive sample). And the largest margin with happen **when the separation line is orthogonal to the coordinate** of samples. Thus, the margin value equals to $\dfrac{1}{2}(x_{M+1} - x_M)$.
4. [**a**] $2 + 2 \cdot (1 - 2\rho)^2$
    - The 1D perceptron with margin at least $\rho \in [0, 0.5]$, means the distance of x_1 and x_2 should be at least 2ρ. If it not holds, the "separation line" can not between both sample points.
    - Consider two cases: $\begin{cases} |x_1 - x_2| < 2\rho \cdots\cdots (1) \\ |x_1 - x_2| > 2\rho \cdots\cdots (2) \end{cases}$
    - In the (1) cases, because the "separation line" can not between both points, it can only be in the area left than the smallest sample point, or right than the largest one, making for the two classification $\{(\times, \times), (\bigcirc, \bigcirc)\}$ (no matter x_1 and x_2 which one is bigger).
    - In the (2) cases, if $x_1 < x_2$, the range of x_2 is $x_2 = 1 - x_1 - 2\rho$. To calculate the expected value of x_2:
    - $$\begin{aligned} \mathrm{E}(x_2) &= \mathrm{E}(1 - x_1 - 2\rho) \\ &= \int_0^{1-2\rho} (1 - x_1 - 2\rho) dx_1 \\ &= x_1 - \frac{x_1^2}{2} - 2\rho x_1 \Bigg|_{x_1=0}^{x_1=1-2\rho} \\ &= \frac{(1-2\rho)^2}{2} \end{aligned}.$$
    - Noticed that the integral range of x_1 is from 0 (where $x_2 \in [2\rho, 1]$) and $1 - 2\rho$ (where $x_2 = 1$). Because (2) cases has four possible events (if $x_1 > x_2$ for $\{(\times, \bigcirc), (\bigcirc, \times)\}$, and if $x_2 > x_1$ for $\{(\times, \bigcirc), (\bigcirc, \times)\}$), the expected value should be $\dfrac{(1-2\rho)^2}{2} \times 4 = 2 \cdot (1 - 2\rho)^2$

- Combine (1) and (2), the expected number of dichotomies is $2 + 2 \cdot (1 - 2\rho)^2$.

5. **[c]** $-\sum_{n=1}^{N} \rho_+ \llbracket y_n = +1 \rrbracket \alpha_n - \sum_{n=1}^{N} \rho_- \llbracket y_n = -1 \rrbracket \alpha_n$

- $$\begin{cases} \min \frac{1}{2}\mathbf{w}^T\mathbf{w} \\ \text{subject to } y_n(\mathbf{w}^T\mathbf{x}_n + b) \geq \rho_+, \ \forall n \text{ such that } y_n = +1 \\ y_n(\mathbf{w}^T\mathbf{x}_n + b) \geq \rho_-, \ \forall n \text{ such that } y_n = -1 \end{cases}.$$
- Then we combine both inequality, add the Lagrange multipliers α_n:
- $$\mathcal{L}(\alpha, \mathbf{w}, b) = \frac{1}{2}\mathbf{w}^T\mathbf{w} + \sum_{n=1}^{N} \alpha_n \llbracket y_n = +1 \rrbracket \Big(\rho_+ - y_n(\mathbf{w}^T\mathbf{x}_n + b) \Big) + \sum_{n=1}^{N} \alpha_n \llbracket y_n = -1 \rrbracket \Big(\rho_- - y_n(\mathbf{w}^T\mathbf{x}_n + b) \Big)$$
  .
- $$\begin{aligned} \frac{\partial \mathcal{L}(\alpha, \mathbf{w}, b)}{\partial b} &= -\sum_{n=1}^{N} \alpha_n \llbracket y_n = +1 \rrbracket y_n - \sum_{n=1}^{N} \alpha_n \llbracket y_n = -1 \rrbracket y_n \\ &= -\left(\sum_{n=1}^{N} \alpha_n y_n (\llbracket y_n = +1 \rrbracket + \llbracket y_n = -1 \rrbracket) \right) \\ &= 0 \end{aligned}.$$
- Due to the restriction $\sum_{n=1}^{N} y_n \alpha_n = 0$, the result always holds, meaning we can ignore b.
- $$\begin{aligned} \frac{\partial \mathcal{L}(\alpha, \mathbf{w}, b)}{\partial \mathbf{w}} &= \frac{\partial \mathcal{L}(\alpha, \mathbf{w})}{\partial \mathbf{w}} \\ &= \mathbf{w} - \sum_{n=1}^{N} \alpha_n \llbracket y_n = +1 \rrbracket y_n \mathbf{x}_n - \sum_{n=1}^{N} \alpha_n \llbracket y_n = -1 \rrbracket y_n \mathbf{x}_n \\ &= 0 \end{aligned}.$$
- We know $\mathbf{w} = \sum_{n=1}^{N} \alpha_n \llbracket y_n = +1 \rrbracket y_n \mathbf{x}_n - \sum_{n=1}^{N} \alpha_n \llbracket y_n = -1 \rrbracket y_n \mathbf{x}_n$.
- Ignore b, substitute $\mathbf{w} = \sum_{n=1}^{N} \alpha_n \llbracket y_n = +1 \rrbracket y_n \mathbf{x}_n - \sum_{n=1}^{N} \alpha_n \llbracket y_n = -1 \rrbracket y_n \mathbf{x}_n$ into $\mathcal{L}(\cdot)$:
- $$\begin{aligned} \mathcal{L}(\alpha, \mathbf{w}, b) &= -\frac{1}{2}\mathbf{w}^T\mathbf{w} + \sum_{n=1}^{N} \alpha_n \rho_+ \llbracket y_n = +1 \rrbracket + \sum_{n=1}^{N} \alpha_n \rho_- \llbracket y_n = -1 \rrbracket \\ &= -\frac{1}{2} \left\| \sum_{n=1}^{N} \alpha_n y_n \mathbf{x}_n \right\|^2 + \sum_{n=1}^{N} \alpha_n \rho_+ \llbracket y_n = +1 \rrbracket + \sum_{n=1}^{N} \alpha_n \rho_- \llbracket y_n = -1 \rrbracket \end{aligned}$$
- The function is try to maximize, therefore multiply a minus in front of the function to make it become a minimization problem:
- $$\begin{aligned} &\min - \left(-\frac{1}{2} \left\| \sum_{n=1}^{N} \alpha_n y_n \mathbf{x}_n \right\|^2 + \sum_{n=1}^{N} \alpha_n \rho_+ \llbracket y_n = +1 \rrbracket + \sum_{n=1}^{N} \alpha_n \rho_- \llbracket y_n = -1 \rrbracket \right) \\ &\longrightarrow \min \frac{1}{2} \left\| \sum_{n=1}^{N} \alpha_n y_n \mathbf{x}_n \right\|^2 + \sum_{n=1}^{N} \alpha_n \rho_+ \llbracket y_n = +1 \rrbracket + \sum_{n=1}^{N} \alpha_n \rho_- \llbracket y_n = -1 \rrbracket \\ &\longrightarrow \min \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} \alpha_n \alpha_m y_n y_m \mathbf{x}_n^T \mathbf{x}_m - \sum_{n=1}^{N} \alpha_n \rho_+ \llbracket y_n = +1 \rrbracket - \sum_{n=1}^{N} \alpha_n \rho_- \llbracket y_n = -1 \rrbracket \end{aligned}.$$

6. **[c]** $\frac{2}{\rho_+ + \rho_-}\alpha^*$

- From problem 5, we know $\sum_{n=1}^{N} y_n \alpha_n = 0 = \sum_{n=1}^{N} \alpha_n \llbracket y_n = +1 \rrbracket - \sum_{n=1}^{N} \alpha_n \llbracket y_n = -1 \rrbracket$, thus $\sum_{n=1}^{N} \llbracket y_n = +1 \rrbracket = \sum_{n=1}^{N} \llbracket y_n = -1 \rrbracket$.
- About α_n, we know $\sum_{n=1}^{N} \alpha_n = \sum_{n=1}^{N} \alpha_n \llbracket y_n = +1 \vee y_n = -1 \rrbracket = \sum_{n=1}^{N} \alpha_n \llbracket y_n = +1 \rrbracket + \sum_{n=1}^{N} \alpha_n \llbracket y_n = -1 \rrbracket$.
- For an even margin SVM (normal definition):
- $$\begin{aligned} &\min_{\alpha} \frac{1}{2} \sum_{n-1}^{N} \sum_{m=1}^{N} \alpha_n \alpha_m y_n y_m \mathbf{x}_n^T \mathbf{x}_m - \sum_{n=1}^{N} \alpha_n \\ &\longrightarrow \min_{\alpha} \frac{1}{2} \sum_{n-1}^{N} \sum_{m=1}^{N} \alpha_n \alpha_m y_n y_m \mathbf{x}_n^T \mathbf{x}_m - \sum_{n=1}^{N} \alpha_n \llbracket y_n = +1 \rrbracket + \sum_{n=1}^{N} \alpha_n \llbracket y_n = -1 \rrbracket \\ &\longrightarrow \min_{\alpha} \frac{1}{2} \sum_{n-1}^{N} \sum_{m=1}^{N} \alpha_n \alpha_m y_n y_m \mathbf{x}_n^T \mathbf{x}_m - \underbrace{2 \times \sum_{n=1}^{N} \alpha_n \llbracket y_n = +1 \rrbracket}_{\blacktriangle} \end{aligned}.$$

- For uneven margin SVM (by result in Problem 5), substitute $\sum_{n=1}^{N} \alpha_n [\![y_n = +1]\!] = \sum_{n=1}^{N} \alpha_n [\![y_n = -1]\!]$ into:
$$\min \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} \alpha_n \alpha_m y_n y_m \mathbf{x}_n^T \mathbf{x}_m - \sum_{n=1}^{N} \alpha_n \rho_+ [\![y_n = +1]\!] - \sum_{n=1}^{N} \alpha_n \rho_- [\![y_n = -1]\!]$$
- $$\longrightarrow \min \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} \alpha_n \alpha_m y_n y_m \mathbf{x}_n^T \mathbf{x}_m - \sum_{n=1}^{N} \alpha_n \rho_+ [\![y_n = +1]\!] - \sum_{n=1}^{N} \alpha_n \rho_- [\![y_n = +1]\!]$$
$$\longrightarrow \min \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} \alpha_n \alpha_m y_n y_m \mathbf{x}_n^T \mathbf{x}_m - \underbrace{(\rho_+ + \rho_-) \sum_{n=1}^{N} \alpha_n [\![y_n = +1]\!]}_{\blacksquare\text{, let } \alpha_n = \alpha_n'}.$$
- Since α^* is an optimal solution of even margin SVM, and ▲ part should be equal to ■ part, we have:
- $$\underbrace{(\rho_+ + \rho_-) \sum_{n=1}^{N} \alpha_n' [\![y_n = +1]\!]}_{\blacksquare} = \underbrace{2 \times \sum_{n=1}^{N} \alpha_n^* [\![y_n = +1]\!]}_{\blacktriangle}.$$
- As above, $\alpha' = \dfrac{2}{\rho_+ + \rho_-} \alpha^*$

7. [**d**] $\log_2 K(\mathbf{x}, \mathbf{x}')$
   - To find a counter-example. Assume $x_1 = \frac{1}{2}, x_2 = \frac{1}{3}, K(\alpha, \beta) = \phi(\alpha)^T \phi(\beta) = \begin{bmatrix} \alpha\alpha & \alpha\beta \\ \beta\alpha & \beta\beta \end{bmatrix}$ (symmetric).
   - Thus, $K = \begin{bmatrix} x_1 x_1 & x_1 x_2 \\ x_2 x_1 & x_2 x_2 \end{bmatrix} = \begin{bmatrix} \frac{1}{4} & \frac{1}{6} \\ \frac{1}{6} & \frac{1}{9} \end{bmatrix}$, the eigenvalue $\lambda_1 = 0 \geq 0$, $\lambda_2 = \frac{13}{36} \geq 0$, satisfy the property of positive semi-definite matrix.
   - [**a**], $K^a = \begin{bmatrix} 2^{x_1 x_1} & 2^{x_1 x_2} \\ 2^{x_2 x_1} & 2^{x_2 x_2} \end{bmatrix} = \begin{bmatrix} 2^{\frac{1}{4}} & 2^{\frac{1}{6}} \\ 2^{\frac{1}{6}} & 2^{\frac{1}{9}} \end{bmatrix}$, the eigenvalue $\lambda_1^a \sim 2.25 \geq 0$, $\lambda_2^a \sim 0.01 \geq 0$, satisfy the property of positive semi-definite matrix.
   - [**b**], $K^b = \begin{bmatrix} (2 - x_1 x_1)^{-2} & (2 - x_1 x_2)^{-2} \\ (2 - x_2 x_1)^{-2} & (2 - x_2 x_2)^{-2} \end{bmatrix} = \begin{bmatrix} \frac{16}{49} & \frac{36}{121} \\ \frac{36}{121} & \frac{81}{289} \end{bmatrix}$, the eigenvalue $\lambda_1^b = \frac{8593}{28322} + \frac{\sqrt{1045849201489}}{3426962} \geq 0$, $\lambda_2^b = \frac{8593}{28322} - \frac{\sqrt{1045849201489}}{3426962} \geq 0$, satisfy the property of positive semi-definite matrix.
   - [**c**], $K^c = \begin{bmatrix} 2 + x_1 x_1 & 2 + x_1 x_2 \\ 2 + x_2 x_1 & 2 + x_2 x_2 \end{bmatrix} = \begin{bmatrix} \frac{9}{4} & \frac{13}{6} \\ \frac{13}{6} & \frac{19}{9} \end{bmatrix}$, the eigenvalue $\lambda_1^c = \frac{157 - \sqrt{24361}}{72} \geq 0$, $\lambda_2^c = \frac{157 + \sqrt{24361}}{72} \geq 0$, satisfy the property of positive semi-definite matrix.
   - [**d**], $K^d = \begin{bmatrix} \log_2(x_1 x_1) & \log_2(x_1 x_2) \\ \log_2(x_2 x_1) & \log_2(x_2 x_2) \end{bmatrix} = \begin{bmatrix} \log_2(\frac{1}{4}) & \log_2(\frac{1}{6}) \\ \log_2(\frac{1}{6}) & \log_2(\frac{1}{9}) \end{bmatrix}$, the eigenvalue $\lambda_1^d = \frac{-\sqrt{2 \log^2 2 + \log^2 3} + \log 6}{\log 2} < 0$, $\lambda_2^d = \frac{\sqrt{2 \log^2 2 + \log^2 3} - \log 6}{\log 2} \geq 0$, **does not** satisfy the property of positive semi-definite matrix.
   - [**e**], $K^e = \begin{bmatrix} (x_1 x_1)^2 & (x_1 x_2)^2 \\ (x_2 x_1)^2 & (x_2 x_2)^2 \end{bmatrix} = \begin{bmatrix} 2^{\frac{1}{16}} & 2^{\frac{1}{36}} \\ 2^{\frac{1}{36}} & 2^{\frac{1}{81}} \end{bmatrix}$, the eigenvalue $\lambda_1^e = 0 \geq 0$, $\lambda_2^e = \frac{97}{1296} \geq 0$, satisfy the property of positive semi-definite matrix.
8. [**c**] 2
   - Apply the transform $\phi(\cdot)$ to kernel $K(\mathbf{x}, \mathbf{x}') = \exp\Big(\gamma ||\phi(\mathbf{x}) - \phi(\mathbf{x}')||^2\Big)$.
   - If $\phi(\mathbf{x}) = \phi(\mathbf{x}')$, for any γ use in Gaussian kernel, the value of squared distance with the kernel trick equals to $\exp(-\gamma \cdot 0) = 0$.
   - If $\phi(\mathbf{x}) \neq \phi(\mathbf{x}')$, expand the squared distance, $||\phi(\mathbf{x}) - \phi(\mathbf{x}')||^2 = \phi(\mathbf{x})^T \phi(\mathbf{x}) - \phi(\mathbf{x})^T \phi(\mathbf{x}') - \phi(\mathbf{x}')^T \phi(\mathbf{x}) + \phi(\mathbf{x}')^T \phi(\mathbf{x}')$, since the sample $\mathbf{x}$ and $\mathbf{x}'$ are transform from the Gaussian kernel, where $\phi(\mathbf{x})^T \phi(\mathbf{x}) = \phi(\mathbf{x}')^T \phi(\mathbf{x}') = 1$. And the $\phi(\mathbf{x})^T \phi(\mathbf{x}')$ and $\phi(\mathbf{x}')^T \phi(\mathbf{x})$ should be > 0, therefore $||\phi(\mathbf{x}) - \phi(\mathbf{x}')||^2 < 2$. The tightest upper bound equals to 2.
9. [**d**] $\dfrac{\ln(N-1)}{\epsilon^2}$
   - Since $\alpha = \mathbf{1}, b = 0$, $h(\mathbf{x}) = \text{sign}\left(\sum_{n=1}^{N} y_n K(\mathbf{x}_n, \mathbf{x})\right)$. The error in sample becomes:

- $E_{\text{in}}(\hat{h}) = \frac{1}{N}\sum_{i=1}^{N}[\![h(x_i) \neq y_i]\!] = \frac{1}{N}\sum_{i=1}^{N}\left[\!\!\left[\text{sign}\left(\sum_{n=1}^{N} y_n K(\mathbf{x}_n, \mathbf{x}_i)\right) \neq y_i\right]\!\!\right].$
- If $E_{\text{in}}(\hat{h}) = 0$, means that all prediction must be all correct (no classification error), where $\left[\!\!\left[\text{sign}\left(\sum_{n=1}^{N} y_n K(\mathbf{x}_n, \mathbf{x}_i)\right) = y_i\right]\!\!\right]$, $\forall i$, this can be represent as $\left[\!\!\left[\left(\sum_{n=1}^{N} y_n K(\mathbf{x}_n, \mathbf{x}_i)\right) \cdot y_i > 0\right]\!\!\right]$, $\forall i$.
- Expand (noticed that if $n = i$, $K(\mathbf{x}_n, \mathbf{x}_i) = \exp(-\gamma||\mathbf{x}_n - \mathbf{x}_i)||^2) = \exp(0) = 1$):
  - When $i = 1$, $y_1 \sum_{n=1}^{N} y_n K(\mathbf{x}_n, \mathbf{x}_1) = y_1 \left(y_1 + y_2 K(\mathbf{x}_2, \mathbf{x}_1) + \cdots + y_N K(\mathbf{x}_N, \mathbf{x}_1)\right) > 0$.
  - When $i = 2$, $y_2 \sum_{n=1}^{N} y_n K(\mathbf{x}_n, \mathbf{x}_2) = y_2 \left(y_1 K(\mathbf{x}_1, \mathbf{x}_2) + y_2 + \cdots + y_N K(\mathbf{x}_N, \mathbf{x}_2)\right) > 0.$
  - $\vdots$
  - When $i = N$, $y_N \sum_{n=1}^{N} y_n K(\mathbf{x}_n, \mathbf{x}_2) = y_N \left(y_1 K(\mathbf{x}_1, \mathbf{x}_N) + y_2 K(\mathbf{x}_2, \mathbf{x}_N) + \cdots + y_N\right) > 0.$
- Sum up above, the result should still be large than 0.
- $$\begin{aligned} \blacksquare = \sum_{i=1}^{N}\left(y_i \sum_{n=1}^{N} y_n K(\mathbf{x}_n, \mathbf{x}_i)\right) =& \left[y_1^2 + y_1\left(y_2 K(\mathbf{x}_2, \mathbf{x}_1) + \cdots + y_N K(\mathbf{x}_N, \mathbf{x}_1)\right)\right] \\ &+ \left[y_2^2 + y_2\left(y_1 K(\mathbf{x}_1, \mathbf{x}_2) + \cdots + y_N K(\mathbf{x}_N, \mathbf{x}_2)\right)\right] \\ &+ \cdots \\ &+ \left[y_N^2 + y_N\left(y_1 K(\mathbf{x}_1, \mathbf{x}_N) + \cdots + y_{N-1} K(\mathbf{x}_{N-1}, \mathbf{x}_N)\right)\right] \\ >& 0 \end{aligned}$$
- Since $\gamma > 0$ (large enough), $||\mathbf{x}_n - \mathbf{x}_m|| \geq \epsilon$, $\forall n \neq m$, $K(\mathbf{x}_n, \mathbf{x}_m) = \exp(-\gamma||\mathbf{x}_n - \mathbf{x}_m||^2) \leq \exp(-\gamma\epsilon^2)$, we have:
- $$\begin{aligned} 0 < \blacksquare \leq \blacktriangle =& \left[y_1^2 + y_2^2 + \cdots + y_N^2\right] \\ &+ \underbrace{\left[y_1\left(\underbrace{y_2 \exp(-\gamma\epsilon^2) + \cdots + y_N \exp(-\gamma\epsilon^2)}_{N-1}\right) + \cdots + y_N\left(\underbrace{y_1 \exp(-\gamma\epsilon^2) + \cdots + y_{N-1} \exp(-\gamma\epsilon^2)}_{N-1}\right)\right]}_{N} \end{aligned}$$
- Noticed that $y_i = (+1) \vee (-1)$, $\forall i$, so $y_i^2 = 1$, $\forall i$. On the other hand, $-1 < y_n y_m < 1$, $\forall n \neq m$, assume all $y_n y_m = -1$ (use the smallest value to find a lower bound of $\blacktriangle$), by this inequality, we have
- $$\begin{aligned} 0 < \blacktriangle \leq& \sum_{i=1}^{N} y_i^2 + \sum_{i=1}^{N}(-1) \times (N-1)\exp(-\gamma\epsilon^2) \\ =& \sum_{i=1}^{N} y_i^2 + N \cdot -(N-1)\exp(-\gamma\epsilon^2) \\ =& N - N \cdot (N-1)\exp(-\gamma\epsilon^2) \end{aligned}.$$
- From $0 < N - N \cdot (N-1)\exp(-\gamma\epsilon^2)$, then:
- $$\begin{aligned} 1 >& (N-1)\exp(-\gamma\epsilon^2) \\ \longrightarrow \frac{1}{N-1} >& \exp(-\gamma\epsilon^2) \\ \longrightarrow \ln(N-1) <& \gamma\epsilon^2 \\ \longrightarrow \frac{\ln(N-1)}{\epsilon^2} <& \gamma \end{aligned}.$$

10. [**c**] $\alpha_{t+1} \leftarrow \alpha_t$ expect $\alpha_{t+1,n(t)} \leftarrow \alpha_{t,n(t)} + y_{n(t)}$

    - Substituted $\mathbf{w}_t$ in the second equation into the first equation:
      $$\begin{aligned} \mathbf{w}_{t+1} \leftarrow& \mathbf{w}_t + y_n\phi(\mathbf{x}_n) \\ =& \sum_{n=1}^{N} \alpha_{t,n}\phi(\mathbf{x}_n) + y_n\phi(\mathbf{w}_n) \\ =& \alpha_{t,1}\phi(\mathbf{x}_1) + \alpha_{t,2}\phi(\mathbf{x}_2) + \cdots + \alpha_{t,N}\phi(\mathbf{x}_N) + y_n\phi(\mathbf{x}_n) \\ =& \alpha_{t,1}\phi(\mathbf{x}_1) + \alpha_{t,2}\phi(\mathbf{x}_2) + \cdots + \underbrace{(\alpha_{t,n} + y_n)\,\phi(\mathbf{x}_n)}_{\text{update term on } \phi(\mathbf{x}_n)} + \cdots + \alpha_{t,N}\phi(\mathbf{x}_N) \end{aligned}.$$
    - Use the definition given by the second equation on $\mathbf{w}_{t+1}$:

- $$\begin{aligned}\mathbf{w}_{t+1} =& \sum_{n=1}^{N} \alpha_{t+1,n}\phi(\mathbf{x}_n) \\ =&\alpha_{t+1,1}\phi(\mathbf{x}_1) + \alpha_{t+1,2}\phi(\mathbf{x}_2) + \cdots + \underbrace{\alpha_{t+1,n}\phi(\mathbf{x}_n)}_{\text{update term on } \phi(\mathbf{x}_n)} + \cdots + \alpha_{t+1,N}\phi(\mathbf{x}_N)\end{aligned}.$$
- Compare with the term $(\alpha_{t,n} + y_n)\,\phi(\mathbf{x}_n)$ and $\alpha_{t+1,n}\phi(\mathbf{x}_n)$, the α term should be updated by $\alpha_{t+1} \leftarrow \alpha_t$, except by $\alpha_{t+1,n(t)} \leftarrow \alpha_{t,n(t)} + y_{n(t)}$.

11. [**a**] $\sum_{n=1}^{N} \alpha K(\mathbf{x}_n, \mathbf{x})$

- From $\mathbf{w}_t \leftarrow \sum_{n=1}^{N} \alpha_{t,n}\phi(\mathbf{x}_n)$.
- $$\begin{aligned}\mathbf{w}_t^T\phi(\mathbf{x}) &= \left(\sum_{n=1}^{N} \alpha_{t,n}\phi(\mathbf{x}_n)\right)^T \phi(\mathbf{x}) \\ &= \left(\sum_{n=1}^{N} \alpha_{t,n}\phi(\mathbf{x}_n)^T\right) \phi(\mathbf{x}) \\ &= \sum_{n=1}^{N} \alpha_{t,n}\left(\phi(\mathbf{x}_n)^T\phi(\mathbf{x})\right) \\ &= \sum_{n=1}^{N} \alpha K(\mathbf{x}_n, \mathbf{x})\end{aligned}.$$

12. [**b**] $\min_{n:y_n>0}\left(1 - \sum_{m=1}^{N} y_m\alpha_m K(x_n, x_m)\right)$

- Consider a soft-margin SVM where $\alpha_n(1 - \xi_n - y_n(\mathbf{w}^T\mathbf{z}_n + b)) = 0$, if $\alpha^*_{\mathbf{n}} = C \neq 0$, $(1 - \xi_n - y_n(\mathbf{w}^T\mathbf{z}_n + b))$ term should be 0.
- Another restriction is that $\xi \geq 0$, $\xi_n = (1 - y_n(\mathbf{w}^T\mathbf{z}_n + b)) \geq 0$, so $y_n(\mathbf{w}^T\mathbf{z}_n + b) \leq 1$.
- Consider $\begin{cases} y_n > 0,\ b \leq \dfrac{1}{y_n} - \mathbf{w}^T\mathbf{z}_n = \dfrac{1}{y_n} - \sum_{m=1}^{N} y_m\alpha_m K(x_n, x_m) \cdots\cdots (1) \\ y_n < 0,\ b \geq \dfrac{1}{y_n} - \mathbf{w}^T\mathbf{z}_n = \dfrac{1}{y_n} - \sum_{m=1}^{N} y_m\alpha_m K(x_n, x_m) \cdots\cdots (2) \end{cases}$
- However, In the (2) situation, $b \geq \square \in \mathbb{R}$, has no upper bound, b can be divergence. Only the (1) situation, $b \leq \frac{1}{y_n} - \sum_{m=1}^{N} y_m\alpha_m K(x_n, x_m)$ has. To avoid of $n < 0$, we need to use the lower bound (find the smallest one) of $\frac{1}{y_n} - \sum_{m=1}^{N} y_m\alpha_m K(x_n, x_m)$. Thus the largest b can be represented as $\min_{n:y_n>0}\left(1 - \sum_{m=1}^{N} y_m\alpha_m K(x_n, x_m)\right)$.

13. [**e**] $K(\mathbf{x}_n, \mathbf{x}_m) + \frac{1}{2C}[\![n = m]\!]$

- The optimal function and restriction in (P_2): $\begin{cases} \min \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{n=1}^{N}\xi_n^2 \\ \text{subject to } y_n\left(\mathbf{w}^T\phi(\mathbf{x}_n) + b\right) \geq 1 - \xi_n,\ \forall n \in \{1, N\} \end{cases}.$
- Add the Lagrange multipliers α_n:
- $\mathcal{L}(\alpha, \mathbf{w}, b) = \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{n=1}^{N}\xi_n^2 + \sum_{n=1}^{N}\alpha_n\left(1 - \xi_n - y_n\left(\mathbf{w}^T\phi(\mathbf{x}_n) + b\right)\right)$.
- $\frac{\partial\mathcal{L}(\alpha, \mathbf{w}, b)}{\partial b} = -\sum_{n=1}^{N}\alpha_n y_n = 0$, due to $\sum_{n=1}^{N}\alpha_n y_n = 0$, we can ignore b ever.
- $\frac{\partial\mathcal{L}(\alpha, \mathbf{w}, b)}{\partial\xi_n} = 2C\xi_n - \alpha_n$, so $\xi_n = \frac{\alpha_n}{2C}$.
- $\frac{\partial\mathcal{L}(\alpha, \mathbf{w}, b)}{\partial\mathbf{w}} = \mathbf{w} - \sum_{n=1}^{N}\alpha_n y_n\phi(\mathbf{x}_n) = 0$, so $\mathbf{w} = \sum_{n=1}^{N}\alpha_n y_n\phi(\mathbf{x}_n)$.
- Ignore b, substitute $\xi_n = \frac{\alpha_n}{2C}$ and $\mathbf{w} = \sum_{n=1}^{N}\alpha_n y_n\phi(\mathbf{x}_n)$ into $\mathcal{L}(\cdot)$:

- $$\begin{aligned}\mathcal{L}(\alpha, \mathbf{w}, b) &= \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{n=1}^{N}\xi_n^2 + \sum_{n=1}^{N}\alpha_n\left(1 - \xi_n - y_n\left(\mathbf{w}^T\phi(\mathbf{x}_n) + b\right)\right) \\ &= \frac{1}{2}\mathbf{w}^T\mathbf{w} - \frac{1}{4C}\sum_{n=1}^{N}\alpha_n^2 + \sum_{n=1}^{N}\alpha_n - \mathbf{w}^T\mathbf{w} \\ &= -\frac{1}{2}\left\|\sum_{n=1}^{N}\alpha_n y_n\phi(\mathbf{x}_n)\right\|^2 - \frac{1}{4C}\left\|\sum_{n=1}^{N}\alpha_n\right\|^2 + \sum_{n=1}^{N}\alpha_n \\ &= -\frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}\alpha_n\alpha_m y_n y_m\phi(\mathbf{x}_n)^T\phi(\mathbf{x}_m) - \frac{1}{4C}\sum_{n=1}^{N}\sum_{m=1}^{N}\llbracket n = m\rrbracket\alpha_n\alpha_m y_n y_m + \sum_{n=1}^{N}\alpha_n \\ &= -\frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}\alpha_n\alpha_m y_n y_m\left(\phi(\mathbf{x}_n)^T\phi(\mathbf{x}_m) + \frac{1}{2C}\llbracket n = m\rrbracket\right) + \sum_{n=1}^{N}\alpha_n\end{aligned}.$$
- The Lagrange function is try to maximize, therefore multiply a minus $(-)$ in front of the function to make it become a minimization problem:
  $$\min -\left(-\frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}\alpha_n\alpha_m y_n y_m\left(\phi(\mathbf{x}_n)^T\phi(\mathbf{x}_m) + \frac{1}{2C}\llbracket n = m\rrbracket\right) + \sum_{n=1}^{N}\alpha_n\right)$$
- $$\begin{aligned}&\longrightarrow \min\frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}\alpha_n\alpha_m y_n y_m\left(\phi(\mathbf{x}_n)^T\phi(\mathbf{x}_m) + \frac{1}{2C}\llbracket n = m\rrbracket\right) - \sum_{n=1}^{N}\alpha_n \\ &\longrightarrow \min\frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}\alpha_n\alpha_m y_n y_m\left(K(\mathbf{x}_n, \mathbf{x}_m) + \frac{1}{2C}\llbracket n = m\rrbracket\right) - \sum_{n=1}^{N}\alpha_n\end{aligned}.$$

14. **[e]** $\xi^* = \dfrac{1}{2C}\alpha^*$
    - As above, $\dfrac{\partial\mathcal{L}(\alpha, \mathbf{w}, b)}{\partial\xi_n} = 2C\xi_n - \alpha_n$, so $\xi_n = \dfrac{\alpha_n}{2C}$ (ξ_n and α_n is the n-th element in the vector ξ and α). The optimal ξ^* (vector) equals to $\frac{1}{2C}\alpha^*$ (vector).

---

15. **[d]** 8.5
    - `w_norm=8.459972213043049`
16. **[b]** "2" versus "not 2"

    -
      ```
      Problem 16
      1 versus not 1
      0/1 Error = 0.000676
      ---------------------------------------------
      2 versus not 2
      0/1 Error = 0.0
      ---------------------------------------------
      3 versus not 3
      0/1 Error = 0.022322
      ---------------------------------------------
      4 versus not 4
      0/1 Error = 0.040135
      ---------------------------------------------
      5 versus not 5
      0/1 Error = 0.006764
      ---------------------------------------------
      ```
17. **[c]** 700

    -
      ```
      Problem 17
      1 versus not 1
      number of support vectors: 145
      ---------------------------------------------
      2 versus not 2
      number of support vectors: 87
      ---------------------------------------------
      3 versus not 3
      number of support vectors: 433
      ---------------------------------------------
      4 versus not 4
      ```

```
number of support vectors: 711
--------------------------------------------
5 versus not 5
number of support vectors: 258
--------------------------------------------
```

18. **[d/e]** $C = 10 \vee C = 100$

-
  ```
  Problem 18
  6 versus not 6
  --------------------------------------------
  C = 0.01
  0/1 Error = 0.235
  --------------------------------------------
  C = 0.1
  0/1 Error = 0.1635
  --------------------------------------------
  C = 1
  0/1 Error = 0.1065
  --------------------------------------------
  C = 10
  0/1 Error = 0.097
  --------------------------------------------
  C = 100
  0/1 Error = 0.097
  --------------------------------------------
  ```

19. **[b]** $\gamma = 1$

-
  ```
  Problem 19
  6  versus not 6
  --------------------------------------------
  gamma = 0.1
  0/1 Error = 0.0985
  --------------------------------------------
  gamma = 1
  0/1 Error = 0.07
  --------------------------------------------
  gamma = 10
  0/1 Error = 0.1635
  --------------------------------------------
  gamma = 100
  0/1 Error = 0.235
  --------------------------------------------
  gamma = 1000
  0/1 Error = 0.235
  --------------------------------------------
  ```

20. **[b]** $\gamma = 1$

-
  ```
  Problem 20
  number of different gamma selected in 1000 iterations
  --------------------------------------------
  gamma=0.1: 124
  gamma=1: 876
  gamma=10: 0
  gamma=100: 0
  gamma=1000: 0
  ```