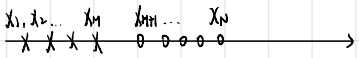
1.



Let x be $y=-1$, o be $y=1$

We know margin $= \min_{n=1,\dots N}$ distance(x_n, w)

Hence, in order to separate $y=-1$, $y=1$, and also achieve max margin, we pick w be the middle of x_{M+1}, x_M, $w = \frac{x_M + x_{M+1}}{2}$, the margin will be $\frac{x_{M+1} - x_M}{2}$

2.

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

subject to $(w^T x_n + b) \geq 1$ for $y_n = 1$ $\quad \Rightarrow \quad 0 \geq 1 - (w^T x_n + b)$ for $y_n = 1$

$-(w^T x_n + b) \geq 1$ for $y_n = -1$ $\quad \Rightarrow \quad 0 \geq 1 + (w^T x_n + b)$ for $y_n = -1$

First, we find the lagrange function, we suppose $y_1 \sim y_M = 1$, $y_{M+1} \sim y_N = -1$

$$\mathcal{L}(b, w, \alpha) = \frac{1}{2} w^T w + \sum_{n=1}^{M} \alpha_n (1 - y_n (w^T x_n + b)) + \sum_{n=M+1}^{N} \alpha_n (1 - y_n (w^T x_n + b))$$

Then, we solve the Lagrange duel.

$$\max_{\text{all } \alpha \geq 0} \left[\min_{b,w} \frac{1}{2} w^T w + \sum_{n=1}^{M} \alpha_n (1 - y_n (w^T x_n + b)) + \sum_{n=M+1}^{N} \alpha_n (1 - y_n (w^T x_n + b)) \right]$$

Inner problem "unconstrained" at optimal: $\frac{\partial \mathcal{L}(b, w, \alpha)}{\partial b} = 0 = -\sum_{n=1}^{N} y_n \alpha_n$

then it becomes $\max_{\text{all } \alpha \geq 0, \sum_{n=1}^{N} y_n \alpha_n = 0}$ $\left[\min_{b,w} \frac{1}{2} w^T w + \sum_{n=1}^{M} \alpha_n (1 - y_n w^T x_n) + \sum_{n=M+1}^{N} \alpha_n (1 - y_n w^T x_n) \right]$

Inner problem "unconstrained" at optimal: $\frac{\partial \mathcal{L}(b, w, \alpha)}{\partial w_i} = 0 = w_i - \sum_{n=1}^{N} \alpha_n y_n x_{n,i} \Rightarrow w = \sum_{n=1}^{N} \alpha_n y_n x_n$

then it becomes $\max_{\text{all } \alpha \geq 0, \sum y_n \alpha_n = 0, w = \sum \alpha_n y_n x_n} -\frac{1}{2} \left|\left| \sum_{n=1}^{N} \alpha_n y_n x_n \right|\right|^2 + \sum_{n=1}^{M} \alpha_n + \sum_{n=M+1}^{N} \alpha_n$

3.

We know that the points on the hyperplane constructed by (w_f, b_f) will have the following property

$$\begin{cases} w_f^T x_n + b_f = 1 & \text{if } y_n = 1 \\ w_f^T x_n + b_f = -f & \text{if } y_n = -1 \end{cases}$$

we can rewrite into the following equation.

$$w_f^T x_n + b_f = \frac{1+y_n}{2} - \frac{1-y_n}{2} f = \frac{1}{2}(1-f) + \frac{1}{2}(1+f) y_n \quad \cdots ①$$

Moreover, we know the hyperplane obtained by $f = 1$ has the property that

$$w_1^T x_n + b_1 = y_n \quad \cdots ②$$

we insert ② into ①

$$w_f^T x_n + b_f = \frac{1}{2}(1-f) + \frac{1}{2}(1+f)(w_1^T x_n + b_1) = \frac{1}{2}(1+f)(w_1^T x_n) + \frac{1}{2}(1-f) + \frac{1}{2}(1+f) b_1$$

Hence we have $\begin{cases} w_f^T x_n = \frac{1}{2}(1+f)(w_1^T x_n) \\ b_f = \frac{1}{2}(1-f) + \frac{1}{2}(1+f) b_1 = \frac{1+b_1}{2} - \frac{1-b_1}{2} f \end{cases}$

Since $w = \sum \alpha y_n x_n$, w and α are linear relationship.

$\alpha_f = \alpha_1 (\frac{1}{2} + \frac{1}{2} f)$, hence w_f is obtained correctly.

for $b_f = \frac{1+y_n}{2} - \frac{1-y_n}{2} f - w_f^T x_n$ with any $sv(x_n, y_n)$, which the same sv as $f = 1$, since we only change the weight.

From ① & ②, b_f can be obtained correctly.

Hence, $\begin{cases} w_f = (\frac{1}{2} + \frac{1}{2} f) w_1 \\ b_f = \frac{1+b_1}{2} - \frac{1-b_1}{2} f \end{cases}$

We know $\begin{cases} w_{11+b} = (\frac{1}{2} + 5b3) w_1 \\ b_{11+b} = \frac{1+b_1}{2} - \frac{1-b_1}{2} x_{11+b} \end{cases}$ ✱

4. From Q3, we derive the value of $\alpha_p = \alpha_1 (\frac{1}{2} + \frac{1}{2}\rho)$

Hence, we can see that α_1 and α_p have different optimal langrange multiplier.

Therefore, meaning that α_p is the only optimal solution for the uneven margin SVM.

αn^* is not an optimal solution of the uneven SVM.

5. $K(x, x') = K_1(x, x') K_2(x, x')$

$$= \phi_1(x)^T \phi_1(x') \phi_2(x)^T \phi_2(x')$$

$$= \left(\sum_{i=1}^{d} \phi_{1i}(x) \phi_{1i}(x')\right)\left(\sum_{j=1}^{d} \phi_{2j}(x) \phi_{2j}(x')\right)$$

$$= \sum_{i=1}^{d}\sum_{j=1}^{d} \left(\phi_{1i}(x) \phi_{2j}(x)\right)\left(\phi_{1i}(x') \phi_{2j}(x')\right)$$

$$= \sum_{i=1}^{d}\sum_{j=1}^{d} \phi_{ij}(x) \phi_{ij}(x')$$

$$= \phi(x)^T \cdot \phi(x')$$

$$\phi_1(x) = \begin{pmatrix} \phi_{11}(x) \\ \phi_{12}(x) \\ \vdots \\ \phi_{1d}(x) \end{pmatrix}$$

$$\phi_2(x) = \begin{pmatrix} \phi_{21}(x) \\ \phi_{22}(x) \\ \vdots \\ \phi_{2d}(x) \end{pmatrix}$$

where $\phi(x)$ is a $d \times d$ vector, s.t. $\phi_{ij}(x) = \phi_{1i}(x) \phi_{2j}(x)$

Now, we see that we can write $K(x, x')$ as an inner product using $\phi(x)$, we know $K(x, x')$ is a valid Kernel.

b.

$$\|\phi(x)\|^2 = \phi(x)^T \phi(x) = K(x, x)$$

$$\|\phi(x)\| = \sqrt{K(x,x)}$$

$$\|\phi(x) - \phi(x')\|^2 = \|\phi(x)\|^2 + \|\phi(x')\|^2 - 2\phi(x)^T\phi(x')$$

$$= K(x,x) + K(x',x') - 2K(x,x')$$

$$\|\phi(x) - \phi(x')\| = \sqrt{K(x,x) + K(x',x') - 2K(x,x')}$$

⇒ Since x & x' are unit vectors, $K(x,x)$ & $K(x',x')$ are both constant.

largest possible distance yields when $-2k(x,x')$ is as large as possible;

$-2k(x,x') = -2(1 + x^T x')^2$ achieve its largest value when $x^T x' = -1$

which is x and x' point towards the opposite direction,

suppose $x = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$, $x' = \begin{bmatrix} -1 \\ 0 \end{bmatrix}$, then we have

$$\|\phi(x) - \phi(x')\| = \sqrt{4 + 4 - 0} = 2\sqrt{2} \quad ※$$

Smallest distance yields when $-2k(x,x')$ is as small as possible

$-2k(x,x') = -2(1 + x^T x')^2$ achieve its smallest value when $x^T x' = 1$

which means x and x' are parellel, pointing towards the same direction. Suppose $x = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$, $x' = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$, then we have

$$\|\phi(x) - \phi(x')\| = \sqrt{4 + 4 - 8} = 0 \quad ※$$

7.

$$\|\tilde{\phi}(x)\| = \sqrt{\sum_{i=0}^{\infty} \sqrt{\frac{2^i}{i!}} (x^i)^2} = \sqrt{\sum_{i=0}^{\infty} \frac{2^i}{i!} x^{2i}} = \sqrt{\sum_{i=0}^{\infty} \frac{(2x^2)^i}{i!}} = \sqrt{e^{2x^2}} = e^{x^2} = \exp(x^2)$$

Taylor's series of e^{2x^2}

$$\Rightarrow \exp(-x^2) = \frac{1}{\|\tilde{\phi}(x)\|}$$ ※

8.

$$\cos(x, x') = \frac{x^T x'}{\|x\| \|x'\|} = \frac{1}{\|x^T\|} x^T \cdot \frac{1}{\|x'\|} x' = \phi(x)^T \phi(x')$$

$\phi(x) = \frac{1}{\|x\|} x$, we can write $\cos(x, x')$ as inner product of $\phi(x)$, hence $\cos(x, x')$ is a valid kernel. ※

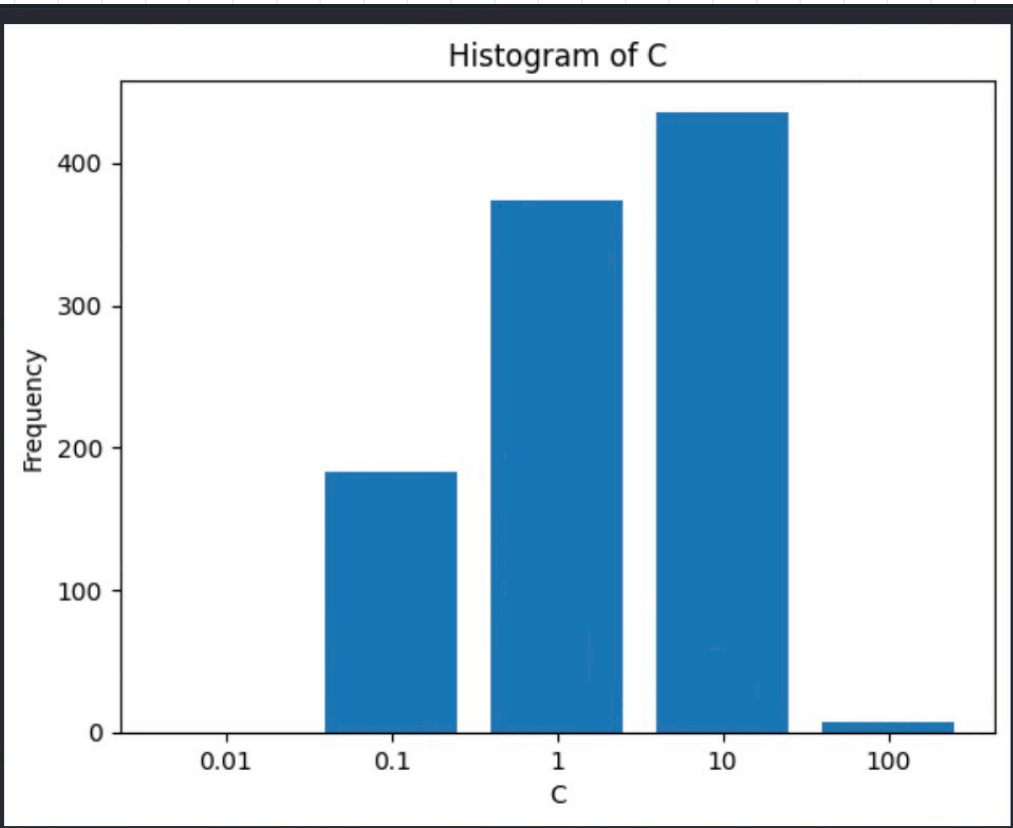
9

```
Minimum number of support vectors = 629, C = 10, Q = 4
```

10

```
Minimum Eout = 0.005000000000000044, C = 1
```

11



We can see that C = 10 has the highest frequency to have the best Eval.

Secondly, C = 1, then C = 0.1, lastly C = 100.

We know that C limits the value of a_n, when C is low it has higher margin, when C is high it will have lower margin.

This is the trade-off between low training error and low testing error.

We want C to be big enough that it achives good Ein but not overfitting, which results in bad Eval.

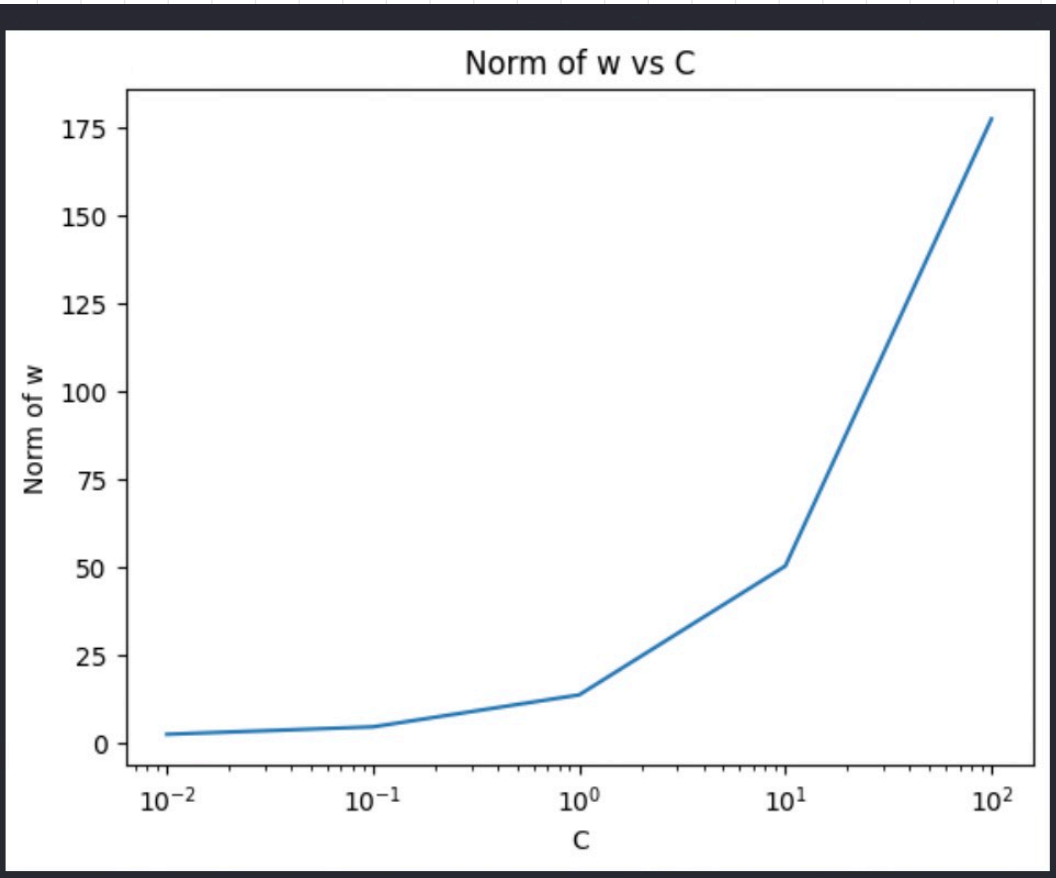
In this case that we see when C rises from 0.1, the frequency of each value increases.

Eventually, we reach C = 10, which is the best value of C, having highest frequency of best Eval.

Meaning when we train with C = 10, we have high chance to get both good Ein and Eval.

When we increase C to 100, the Ein might be better, but the model overfitted which result in bad Eval, we can see the frequency dropped significantly.

12



From the plot we can see that the norm of w have a bit of positive correlation to C.

With small |w| results in small C, large |w| results in large C.

Moreover, we know that small C will have larger margin, large C will have small margin.

Hence, we can conclude that large norm |w| will results in small margin, which reduces the Ein, however might cause Eval to be worse if we made it too big.

13.

Primal SVM: $\min_{w,b} \frac{1}{2} w^T w$ s.t. $y_n(w^T x_n + b) \geq 1$ for $n = 1, 2, \dots N$

$\mathcal{L}(b, w, \alpha) = \underbrace{\frac{1}{2} w^T w}_{\text{obj.}} + \sum_{n=1}^{N} \alpha_n \underbrace{(1 - y_n(w^T x_n + b))}_{\text{constrain.}}$

Lagrange problem for primal SVM: $\min_{b,w} \left(\max_{\text{all } \alpha_n \geq 0} \mathcal{L}(b, w, \alpha) \right)$

Lagrange dual problem: $\max_{\text{all } \alpha_n \geq 0} \left(\min_{b,w} \mathcal{L}(b, w, \alpha) \right)$

Dual SVM: $\min_{\alpha} \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} \alpha_n \alpha_m y_n y_m z_n^T z_m - \sum_{n=1}^{N} \alpha_n$

subject to $\sum_{n=1}^{N} y_n \alpha_n = 0$, $\alpha_n \geq 0$, for $n = 1, 2 \dots N$

$\mathcal{L}(\alpha, \lambda) = \underbrace{\frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} \alpha_n \alpha_m y_n y_m z_n^T z_m - \sum_{n=1}^{N} \alpha_n}_{\text{obj.}} \underbrace{- \sum_{n=1}^{N} \lambda_n \alpha_n}_{\text{constraint.}} = \frac{1}{2} \text{III} \dots - \sum_{n=1}^{N} (1 + \lambda_n) \alpha_n$

Lagrange problem for Dual SVM $\equiv \min_{\alpha} \left(\max_{\text{all } \lambda \geq 0} (\mathcal{L}(\alpha, \lambda)) \right)$

$= \min_{\alpha} \left(\infty \text{ if violate}, \boxed{\text{obj.}} \text{ if feasable} \right)$

violate: $\alpha_n < 0$ → λ_n goes to ∞ → $-(1+\lambda_n)\alpha_n$ is $+$

feasable: $\alpha_n \geq 0$ → $\lambda = 0$ → $-(1+\lambda_n)\alpha_n$ is $-$

$\min_{\alpha} \left(\max_{\text{all } \lambda \geq 0} (\mathcal{L}(\alpha, \lambda)) \right) \geq \min_{\alpha} \mathcal{L}(\alpha, \lambda')$ for any fix λ' with all $\lambda'_n \geq 0$

// max ≥ any

$\min_{\alpha} \left(\max_{\text{all } \lambda \geq 0} (\mathcal{L}(\alpha, \lambda)) \right) \geq \max_{\text{all } \lambda'_n \geq 0} \left(\min_{\alpha} \mathcal{L}(\alpha, \lambda') \right)$ for best λ' on RHS.

// best is one of any

Lagrange dual for dual SVM: $\max_{\text{all } \lambda_n \geq 0} \left(\min_{\alpha} \mathcal{L}(\alpha, \lambda) \right)$

Largrange for primal SVM: $\min_{b,w} \left(\max_{\text{all } \alpha_n \geq 0} \mathcal{L}(b, w, \alpha) \right)$

As we can see, the max and min are opposite, and they do optimization on different variables. From chatbPT, the dual of dual is not always the same as the primal problem can be confirm.