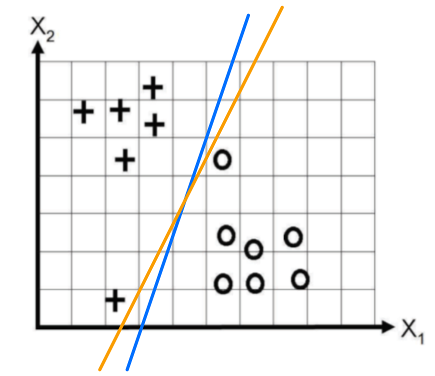
DDA3020 Assignment 2 Student ID: 121090429 Ou Ziyi 欧子旖

Written Problems

1. (1)

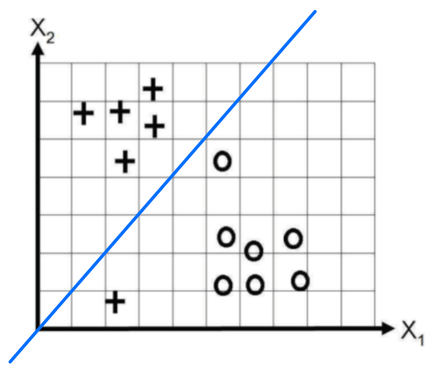


The decision boundary is sketched with blue line in the left figure.

My answer is not unique, the orange line is also a decision boundary.

The classification errors is zero.

(2)

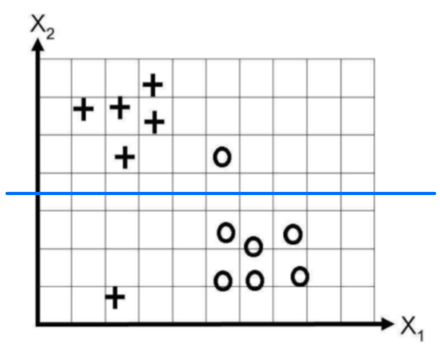


If $w_0=0$, then point $(0,0)$ must be on the decision boundary, since $\sigma(w_0+w_1x_1+w_2x_2)=\sigma(0)=\frac{1}{1+e^{-0}}=0.5$ at this point.

The decision boundary is sketched with blue line in the left figure.

The classification errors is one.

(3)



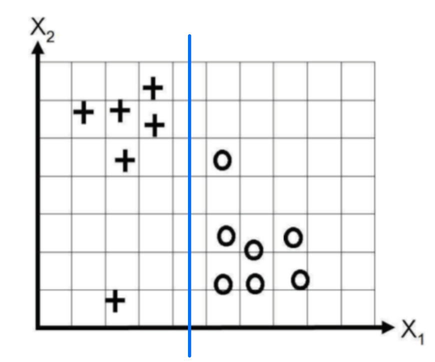
Only regularize w_1 parameter $\Rightarrow w_1=0$

$\sigma(w_0+w_1x_1+w_2x_2)=\sigma(w_0+w_2x_2)$

So the decision boundary should be a horizontal line, which is sketched with blue line in the left figure.

The classification errors is two.

(4)



Only regularize w_2 parameter $\Rightarrow w_2=0$

$\sigma(w_0+w_1x_1+w_2x_2)=\sigma(w_0+w_1x_1)$

So the decision boundary should be a vertical line, which is sketched with blue line in the left figure.

The classification errors is zero.

2. (1) $\phi(x_1)=[1,-1,1]^T$, $\phi(x_2)=[1,2,4]^T$

Let $w=[w_1,w_2,w_3]^T$

Since w is orthogonal to the decision boundary $\{\phi(x): w^T\phi(x)+w_0=0\}$,

if $v\in\{\phi(x): w^T\phi(x)+w_0=0\}$, $w^Tv=0$

Here $\phi(x_1)$ and $\phi(x_2)$ are the only two support vectors,

$\phi(x_1)-\phi(x_2)$ should be perpendicular to the decision boundary.

Therefore $\phi(x_1)-\phi(x_2)$ should be parallel to w

$$\phi(x_1)-\phi(x_2) = [0,-3,-3]^T$$

Therefore $v=[0,1,-1]^T$ should be perpendicular to W since

$$[0,-3,-3]\begin{bmatrix}0\\1\\-1\end{bmatrix}=0$$

(2) Here $\phi(x_1)$ and $\phi(x_2)$ are the only two support vectors,

so $\frac{1}{2}[\phi(x_1)+\phi(x_2)]$ must be on the decision boundary.

$$\frac{1}{2}[\phi(x_1)+\phi(x_2)] = [1,\tfrac{1}{2},\tfrac{5}{2}]$$

Given w perpendicular to decision boundary,

and $[0,-3,-3]^T$ is parallel to w,

here we suggest $W=[0,1,1]^T$.

$$w^T\begin{bmatrix}1\\\frac{1}{2}\\\frac{5}{2}\end{bmatrix}+w_0=0 \Rightarrow w_0=-3$$

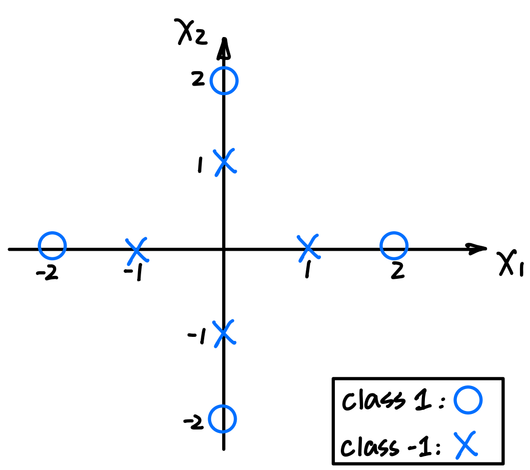
$$\begin{aligned}\|\phi(x_1)-\phi(x_2)\| &= \|[0,-3,-3]\|\\ &= \sqrt{0+9+9}\\ &= \sqrt{18}\\ &= 3\sqrt{2}.\end{aligned}$$

Here the margin is $\frac{1}{2}\|\phi(x_1)-\phi(x_2)\| = \frac{3}{2}\sqrt{2}$.

(3) Since margin $=\frac{1}{\|w\|}=\frac{3}{2}\sqrt{2}$, $\|w\|=\frac{2}{3\sqrt{2}}=\frac{\sqrt{2}}{3}$.

By leveraging it, we get exact $w=[0,\frac{1}{3},\frac{1}{3}]^T$ and exact $w_0=-1$.

3. (1)



Denote the class -1 as X, class +1 as ○ in the left figure.

From the plot we can see that we can't find a svm linear classifier for this data set without classification error.

To solve the problem, we transform each point into high-dimensional space where data becomes linearly separable by applying kernel tricks.

Polynomial kernel, Radial Basis Function (RBF) kernel, and sigmoidal kernel are three widely used kernels.

Here we apply RBF kernel to solve the problem.

Then the decision function will be: $f(x) = \sum_{i=1}^{N} \alpha_i y_i k(x, x_i) + b$

with x_i: the supported vectors; y_i: corresponding label (+1 or -1 here), α_i: the Lagrange multipliers.

(2) RBF kernel is defined as: $k(x, x_i) = \exp\left\{-\frac{\|x-x_i\|^2}{2\sigma^2}\right\}$, which measures the similarity of 2 data points and introduces nonlinearity by transforming input features into higher dimensional space.

Advantages of RBF kernel over linear SVM:

① Ability to handle non-linearly separable data by mapping it to high-dimensional space.

② Ability to capture complex patterns, especially the circular pattern of this dataset.

③ Flexibility in parameter tuning: controls the smoothness of the decision boundary by introducing a parameter σ.

④ Robustness to noise and outliers.

4. (1) By the stationary condition of lagrange function,

we get: $w = \sum_{n=1}^{N} \alpha_n y_i \phi(x_n)$

$$\|w\|^2 = \sum_{n=1}^{N} \sum_{m=1}^{N} \alpha_n \alpha_m y_n y_m \phi(x_n)^T \phi(x_m)$$

Know that margin $\gamma = \frac{1}{\|w\|}$,

therefore $\gamma = \frac{1}{\sqrt{\sum_{n=1}^{N} \sum_{m=1}^{N} \alpha_n \alpha_m y_n y_m \phi(x_n)^T \phi(x_m)}}$

Here we can see that the kernel is: $k(x_n, x_m) = \langle \phi(x_n), \phi(x_m) \rangle$

$\Rightarrow \begin{cases} \text{The decision function is: } f(x) = \sum_{n=1}^{N} \alpha_n y_n \langle \phi(x_n), \phi(x) \rangle + b \\ \text{The decision boundary is: } f(x) = 0 \end{cases}$

Since ϕ transforms input vectors to high-dimensional space, it affects the decision boundary by $\langle \phi(x_n), \phi(x) \rangle$, which measures the similarity between transformed feature vectors.

$\langle \phi(x_n), \phi(x_m) \rangle = \|\phi(x_n)\| \cdot \|\phi(x_m)\| \cdot \cos\theta$, where θ is the angle between transformed vector $\phi(x_m)$ and $\phi(x_n)$

For example, when $\langle \phi(X_n), \phi(X_m) \rangle$ increases, that means $\cos\theta$ becomes larger, so $\phi(X_n)$ and $\phi(X_m)$ are pointing nearly the same direction. So the supported vectors must be aligned, which will lead to a narrow margin and a bad testing performance.

The above situation coincide with the expression we derive for γ w.r.t $\{a_n\}$ ($\gamma = \frac{1}{\sqrt{\sum_{n=1}^{N}\sum_{m=1}^{N} a_n a_m y_n y_m \phi(X_n)^T \phi(X_m)}}$), when $\phi(X_n)^T \phi(X_m)$ increases, γ decreases.

(2) Proof: Under transformation. we know that

$$\gamma = \frac{1}{\sqrt{\sum_{n=1}^{N}\sum_{m=1}^{N} \alpha_n \alpha_m y_n y_m \phi(X_n)^T \phi(X_m)}},$$

$$\text{So } \frac{1}{\gamma^2} = \sum_{n=1}^{N}\sum_{m=1}^{N} \alpha_n \alpha_m y_n y_m \phi(X_n)^T \phi(X_n)$$

The objective function of primal problem is $\frac{1}{2}\|w\|^2$, and the objective function of dual problem is

$$\sum_{n=1}^{N} \alpha_n - \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} \alpha_n \alpha_m y_n y_m \phi(X_n)^T \phi(X_m)$$

By the optimial condition of strong duality and weak duality,

$$\frac{1}{2}\|w\|^2 = \sum_{n=1}^{N} \alpha_n - \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} \alpha_n \alpha_m y_n y_m \phi(X_n)^T \phi(X_m)$$

By the stationary condition of lagrange function, $w = \sum_{n=1}^{N} \alpha_n y_i \phi(X_n)$

$$\Rightarrow \|w\|^2 = \sum_{n=1}^{N}\sum_{m=1}^{N} \alpha_n \alpha_m y_n y_m \phi(X_n)^T \phi(X_m)$$

$$\text{Therefore } \|w\|^2 = \sum_{n=1}^{N} \alpha_n = \sum_{n=1}^{N}\sum_{m=1}^{N} \alpha_n \alpha_m y_n y_m \phi(X_n)^T \phi(X_m)$$

Since $\gamma^2 = \frac{1}{\|w\|^2}$, $\frac{1}{\gamma^2} = \sum_{n=1}^{N} a_n$ is true under the transformation ϕ.