Yinhong Qin
yq2021@nyu.edu

Homework 3

ECE 6143A-Machine Learning
October 21, 2021

Problem 1

1. First, prove Mercer's theorem in one direction:

   Proof. Consider an arbitrary vector $\mathbf{v} \in \mathbb{R}^n$ and its quadratic form $\mathbf{v}^T\mathbf{K}\mathbf{v}$ where $\mathbf{K}_{i,j} = k(x_i, x_j) = \boldsymbol{\phi}(x_i)^T\boldsymbol{\phi}(x_j) = \langle\boldsymbol{\phi}(x_i), \boldsymbol{\phi}(x_j)\rangle, x_i, x_j \in S, \quad k(.,.)$ is a valid Mercer kernel function and S is a finite sample.

$$\begin{aligned}
\mathbf{v}^T\mathbf{K}\mathbf{v} &= \sum_{i=1}^{n}\sum_{j=1}^{n} v_i v_j K_{i,j} \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} v_i v_j k(x_i, x_j) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} v_i v_j \boldsymbol{\phi}(x_i)^T\boldsymbol{\phi}(x_j) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} (v_i\boldsymbol{\phi}(x_i))^T(v_j\boldsymbol{\phi}(x_j)) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} \langle v_i\boldsymbol{\phi}(x_i), v_j\boldsymbol{\phi}(x_j)\rangle
\end{aligned}$$

   Since the distributive law holds for inner product, we can have:

$$\begin{aligned}
\mathbf{v}^T\mathbf{K}\mathbf{v} &= \sum_{i=1}^{n}\sum_{j=1}^{n} \langle v_i\boldsymbol{\phi}(x_i), v_j\boldsymbol{\phi}(x_j)\rangle \\
&= \langle\sum_{i=1}^{n} v_i\boldsymbol{\phi}(x_i), \sum_{j=1}^{n} v_j\boldsymbol{\phi}(x_j)\rangle \\
&= \left\|\sum_{i=1}^{n} v_i\boldsymbol{\phi}(x_i)\right\|^2 \geq 0
\end{aligned}$$

   □

   Therefore, for any Mercer kernel $k(.,.)$ and finite sample S, the kernel matrix $\mathbf{K}$ is positive semi-definite.

2. Proof of some lemmas:

Lemma 1. *Given two Mercer kernels, $k_1(.,.)$ and $k_2(.,.)$,*

$$k(.,.) = \alpha k_1(.,.) + \beta k_2(.,.), \ for \ \alpha, \beta \geq 0$$

is also a Mercer kernel.

Proof. To prove $k(.,.)$ is a Mercer kernel, we can prove the corresponding kernel matrix $\mathbf{K}$, $K_{i,j} = k(x_i, x_j)$ is positive semi-definite.

Consider an arbitrary vector $\mathbf{v} \in \mathbb{R}^n$ and the quadratic form

$$\begin{aligned}
\mathbf{v}^T\mathbf{K}\mathbf{v} &= \sum_{i=1}^{n}\sum_{j=1}^{n} v_i v_j K_{i,j} \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} v_i v_j k(x_i, x_j) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} v_i v_j (\alpha k_1(x_i, x_j) + \beta k_2(x_i, x_j)) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} v_i v_j (\alpha \boldsymbol{\phi}_1(x_i)^T \boldsymbol{\phi}_1(x_j) + \beta \boldsymbol{\phi}_2(x_i)^T \boldsymbol{\phi}_2(x_j)) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} (\alpha (v_i\boldsymbol{\phi}_1(x_i))^T (v_j\boldsymbol{\phi}_1(x_j)) + \beta (v_i\boldsymbol{\phi}_2(x_i))^T (v_j\boldsymbol{\phi}_2(x_j))) \\
&= \alpha \sum_{i=1}^{n}\sum_{j=1}^{n} \langle v_i\boldsymbol{\phi}_1(x_i), v_j\boldsymbol{\phi}_1(x_j)\rangle + \beta \sum_{i=1}^{n}\sum_{j=1}^{n} \langle v_i\boldsymbol{\phi}_2(x_i), v_j\boldsymbol{\phi}_2(x_j)\rangle \\
&= \alpha \langle \sum_{i=1}^{n} v_i\boldsymbol{\phi}_1(x_i), \sum_{j=1}^{n} v_j\boldsymbol{\phi}_1(x_j)\rangle + \beta \langle \sum_{i=1}^{n} v_i\boldsymbol{\phi}_2(x_i), \sum_{j=1}^{n} v_j\boldsymbol{\phi}_2(x_j)\rangle \\
&= \alpha \left\| \sum_{i=1}^{n} v_i\boldsymbol{\phi}_1(x_i) \right\|^2 + \beta \left\| \sum_{i=1}^{n} v_i\boldsymbol{\phi}_2(x_i) \right\|^2 \geq 0
\end{aligned}$$

□

Lemma 2. *Given a Mercer kernel $k_1(.,.)$ and $k_2(.,.)$,*

$$k(.,.) = k_1(.,.) \times k_2(.,.)$$

is also a Mercer kernel.

Proof. Consider for two arbitrary samples x_i and x_j,

$$\begin{aligned}
k(x_i, x_j) &= k_1(x_i, x_j)k_2(x_i, x_j) \\
&= (\boldsymbol{\phi}_1(x_i)^T \boldsymbol{\phi}_1(x_j))(\boldsymbol{\phi}_2(x_i)^T \boldsymbol{\phi}_2(x_j)) \\
&= (\sum_{p=0}^{\infty} \boldsymbol{\phi}_1(x_i)_p \boldsymbol{\phi}_1(x_j)_p)(\sum_{q=0}^{\infty} \boldsymbol{\phi}_2(x_i)_q \boldsymbol{\phi}_2(x_j)_q) \\
&= \sum_{p=0}^{\infty}\sum_{q=0}^{\infty} \boldsymbol{\phi}_1(x_i)_p \boldsymbol{\phi}_1(x_j)_p \boldsymbol{\phi}_2(x_i)_q \boldsymbol{\phi}_2(x_j)_q \\
&= \sum_{p=0}^{\infty}\sum_{q=0}^{\infty} (\boldsymbol{\phi}_1(x_i)_p \boldsymbol{\phi}_2(x_i)_q)(\boldsymbol{\phi}_1(x_j)_p \boldsymbol{\phi}_2(x_j)_q)
\end{aligned}$$

Since $\boldsymbol{\phi} : \mathbb{R}^n \to \mathbb{R}^\infty$, let $\boldsymbol{\phi}_3 = \boldsymbol{\phi}_1 \cdot \boldsymbol{\phi}_2$, whose component is the product of each component from x_i and x_j, then we can have:

$$\begin{aligned}
k(x_i, x_j) &= \sum_{p=0}^{\infty}\sum_{q=0}^{\infty} (\boldsymbol{\phi}_1(x_i)_p \boldsymbol{\phi}_2(x_i)_q)(\boldsymbol{\phi}_1(x_j)_p \boldsymbol{\phi}_2(x_j)_q) \\
&= \boldsymbol{\phi}_3(x_i)\boldsymbol{\phi}_3(x_j)
\end{aligned}$$

According to the Mercer's theorem, for an arbitrary vector $\mathbf{v} \in \mathbb{R}^n$ and its quadratic form $\mathbf{v}^T\mathbf{K}\mathbf{v}$, where $\mathbf{K}_{i,j} = k(x_i, x_j) = \boldsymbol{\phi}_3(x_i)^T\boldsymbol{\phi}_3(x_j)$,

$$\mathbf{v}^T\mathbf{K}\mathbf{v} \geq 0$$

□

Lemma 3. *Given a Mercer Kernel $k(.,.)$, $f(k(.,.))$ is also a Mercer Kernel, where f is a polynomial function with positive coefficients.*

Proof.

$$f(k(x,y)) = (k(x,y)+a)^b = (\phi(x)^T\phi(y)+a)^b,\ a>0, b>0$$

According to the Binomial theorem,

$$\begin{aligned} f(k(x,y)) &= (k(x,y)+a)^b \\ &= \sum_{i=0}^{b} \binom{b}{i} k^i(x,y)a^{b-i} \end{aligned}$$

According to Lemma 2, each term $k^i(x,y)$ is a Mercel kernel.

We can also notice that the coefficients $\binom{b}{i}$ and a^{b-i} in each term is greater or equal to 0. According to Lemma 1, the sum of each term

$$f(k(x,y)) = (k(x,y)+a)^b = \sum_{i=0}^{b} \binom{b}{i} k^i(x,y)a^{b-i}$$

is therefore also a Mercel kernel. □

Lemma 4. *Given a Mercer Kernel $k(.,.)$, $\exp(k(.,.))$ is also a Mercer Kernel.*

Proof. The Taylor series of $\exp(k(.,.))$ is:

$$\exp(k(.,.)) = \sum_{i=0}^{\infty} \frac{k^i(x,y)}{i!}$$

Similarly to the proof of Lemma 3, we can know that each $k^i(x,y)$ is a Mercel kernel and $\frac{1}{i!} \geq 0$ always holds. Thus, $\exp(k(.,.))$ is also a Mercer kernel. □

3. For the kernel

$$K(x,y) = \exp(-\frac{1}{2}\left\|x-y\right\|^2) = \phi(x)\phi(y)$$

To find the explicit form of $\phi(x)$ and $\phi(y)$, we can use the Taylor series of exponential function.

$$\begin{aligned}
K(x,y) &= \exp(-\frac{1}{2}\left\|x-y\right\|^2) \\
&= \exp(-\frac{1}{2}(x-y)^T(x-y)) \\
&= \exp(-\frac{1}{2}(\left\|x\right\|^2 - 2x^Ty + \left\|y\right\|^2)) \\
&= \exp(-\frac{1}{2}\left\|x\right\|^2)\exp(-\frac{1}{2}\left\|y\right\|^2)\exp(x^Ty) \\
&= \exp(-\frac{1}{2}\left\|x\right\|^2)\exp(-\frac{1}{2}\left\|y\right\|^2)\sum_{i=0}^{\infty}\frac{(x^Ty)^i}{i!}
\end{aligned}$$

According to the multinomial theorem, we can know that:

$$\begin{aligned}
((x^T)y)^i &= (\sum_{k=1}^{n} x_k y_k)^i \\
&= \sum \frac{i!}{t_1!t_2!\cdots t_n!}(x_1y_1)_1^{t_1}(x_2y_2)^{t_2}\cdots(x_ny_n)^{t_n},\ t_1+t_2+\cdots+t_n = i
\end{aligned}$$

Thus, plug this expansion to the kernel, we can have:

$$\begin{aligned}
K(x,y) &= \exp(-\frac{1}{2}\left\|x\right\|^2)\exp(-\frac{1}{2}\left\|y\right\|^2)\sum_{i=0}^{\infty}(\frac{(x^Ty)^i}{i!}) \\
&= \exp(-\frac{1}{2}\left\|x\right\|^2)\exp(-\frac{1}{2}\left\|y\right\|^2)\sum_{i=0}^{\infty}\sum_{t_1+t_2+\cdots+t_n=i}\frac{1}{t_1!t_2!\cdots t_n!}(x_1y_1)^{t_1}(x_2y_2)^{t_2}\cdots(x_ny_n)^{t_n} \\
&= \exp(-\frac{1}{2}\left\|x\right\|^2)\exp(-\frac{1}{2}\left\|y\right\|^2)\sum_{i=0}^{\infty}\sum_{t_1+t_2+\cdots+t_n=i}\frac{x_1^{t_1}x_2^{t_2}\cdots x_n^{t_n}}{\sqrt{t_1!t_2!\cdots t_n!}}\cdot\frac{y_1^{t_1}y_2^{t_2}\cdots y_n^{t_n}}{\sqrt{t_1!t_2!\cdots t_n!}} \\
&= \exp(-\frac{1}{2}\left\|x\right\|^2\sum_{i=0}^{\infty}\sum_{t_1+t_2+\cdots+t_n=i}\frac{x_1^{t_1}x_2^{t_2}\cdots x_n^{t_n}}{\sqrt{t_1!t_2!\cdots t_n!}})\cdot \\
&\quad \exp(-\frac{1}{2}\left\|y\right\|^2\sum_{i=0}^{\infty}\sum_{t_1+t_2+\cdots+t_n=i}\frac{y_1^{t_1}y_2^{t_2}\cdots y_n^{t_n}}{\sqrt{t_1!t_2!\cdots t_n!}})
\end{aligned}$$

Therefore, the explicit form of ϕ is:

$$\phi(x) = \exp(-\frac{1}{2}\left\|x\right\|^2)(1, \sum_{t_1+t_2+\cdots+t_n=1}\frac{x_1^{t_1}x_2^{t_2}\cdots x_n^{t_n}}{\sqrt{t_1!t_2!\cdots t_n!}}, \cdots, \sum_{t_1+t_2+\cdots+t_n=m}\frac{x_1^{t_1}x_2^{t_2}\cdots x_n^{t_n}}{\sqrt{t_1!t_2!\cdots t_n!}}, \cdots))$$

Problem 2

For this problem, I changed the negative label from 0 to -1. Below are my several testing results using linear kernel, polynomial kernel and RBF kernel with different C values.

1. When using linear kernel, the results are:

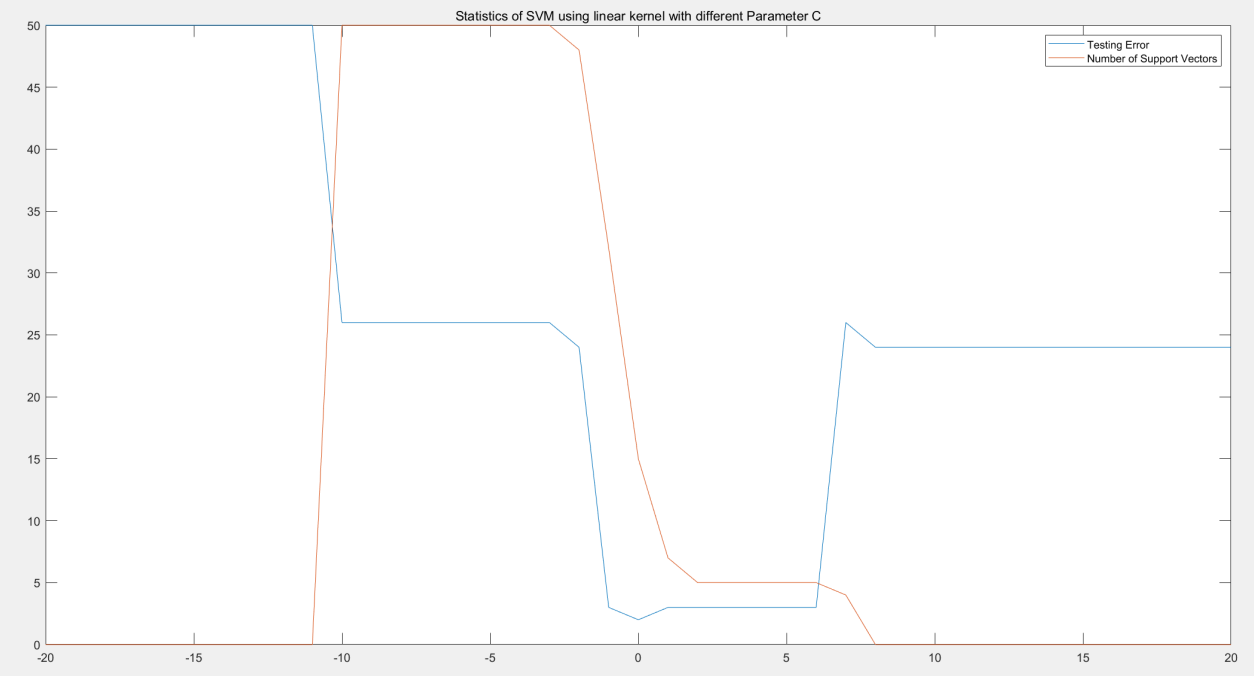

Figure 2.1 Testing results using linear kernel

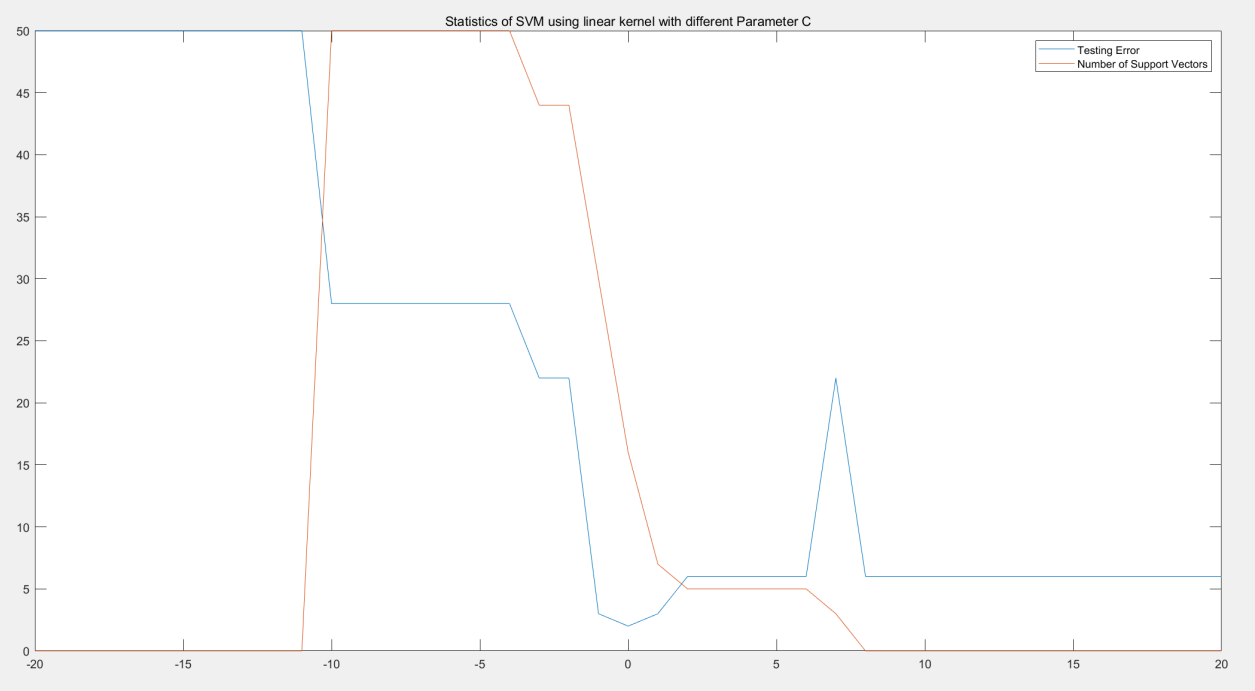

Figure 2.2 Testing results using linear kernel

When using linear kernel, the only parameters we can adjust is the "C". Typically, when C increases the model are more impacted by misclassification error on training set. Therefore the SVM model are more likely to be overfitting and have a relatively smaller margin. On the other hand, if the parameter C decreases, the model become less sensitive to misclassification error and then it may become underfitting. It also may have a larger margin.

We can notice that, for this problem, when $C \in [10^{-1}, 10^{6}]$, the testing error is relative lower than other cases. We should also notice that the number of support vectors are not as much as other cases. It does make sense because when the SVM model correctly classify the testing data, there is no need that all the points should be a support vector. When $C \in [10^{-11}, 10^{-3}]$, the SVM model has relatively larger margin so more points are in the margin region.

This conclusion is also valid when we using the polynomial kernel and RBF kernel.

2. When using polynomial kernel, the results are:
   Note 1: The range of parameter C is $[10^{-20}, 10^{20}]$, the order of polynomial varies in the range $[1, 10]$.
   Note 2: The number in each cell is the error number of each case.

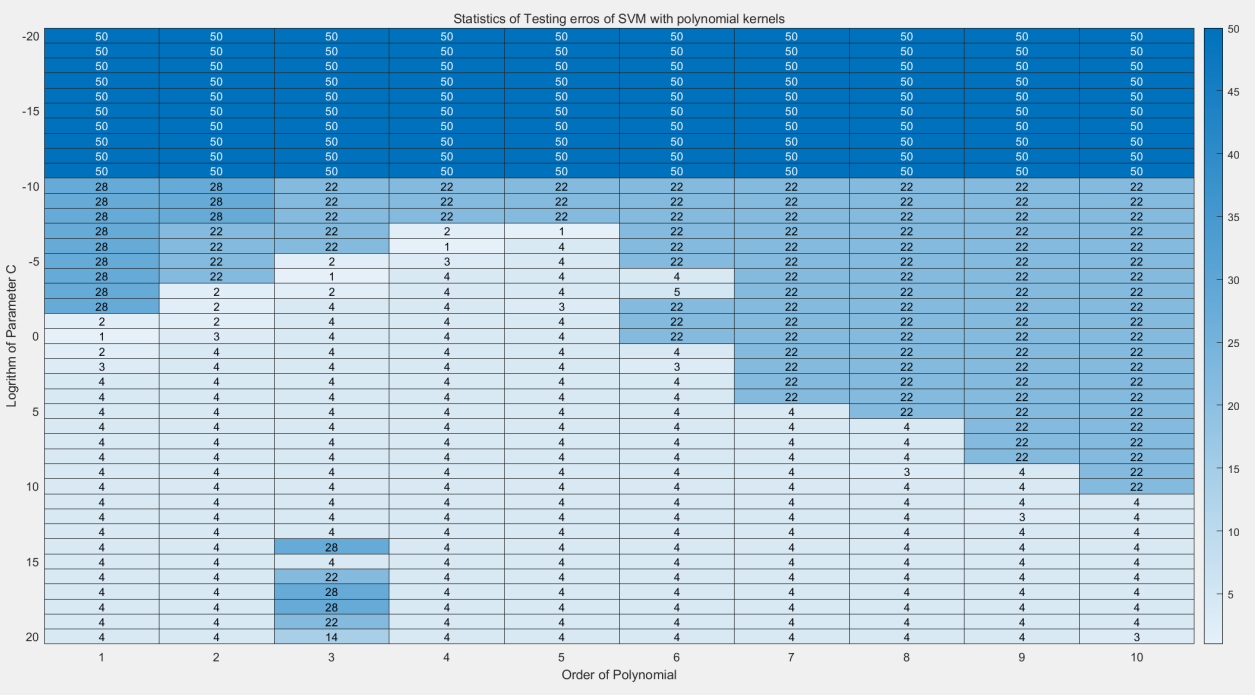

Figure 2.3 Testing results using polynomial kernel

Note: The number in each cell is the number of support vectors data each case.

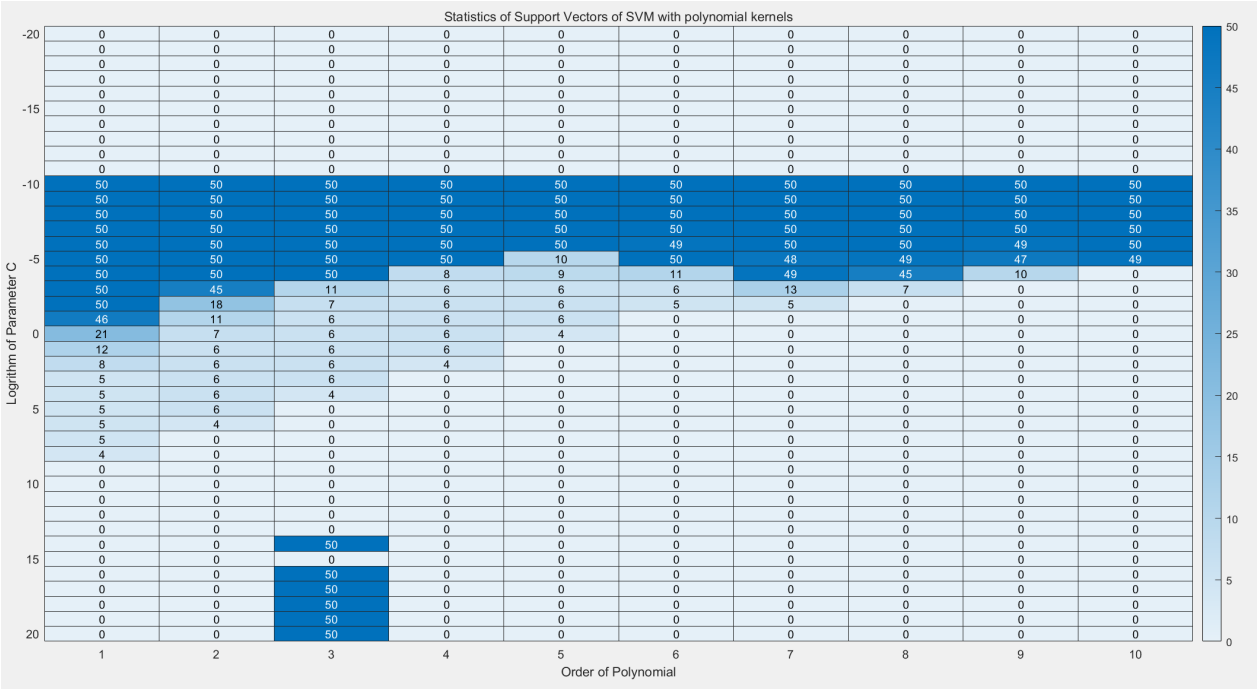

Figure 2.4 Testing results using polynomial kernel

When using polynomial kernel, we can notice that in most cases, the model performs better with larger parameter C value. For a certain order number, the performance of SVM model improves when C increases. The distribution of number of support vectors acts similarly to the model with linear kernel.

3. When using RBF kernels, the results are:
   Note: The number in each cell is the error number of each case.

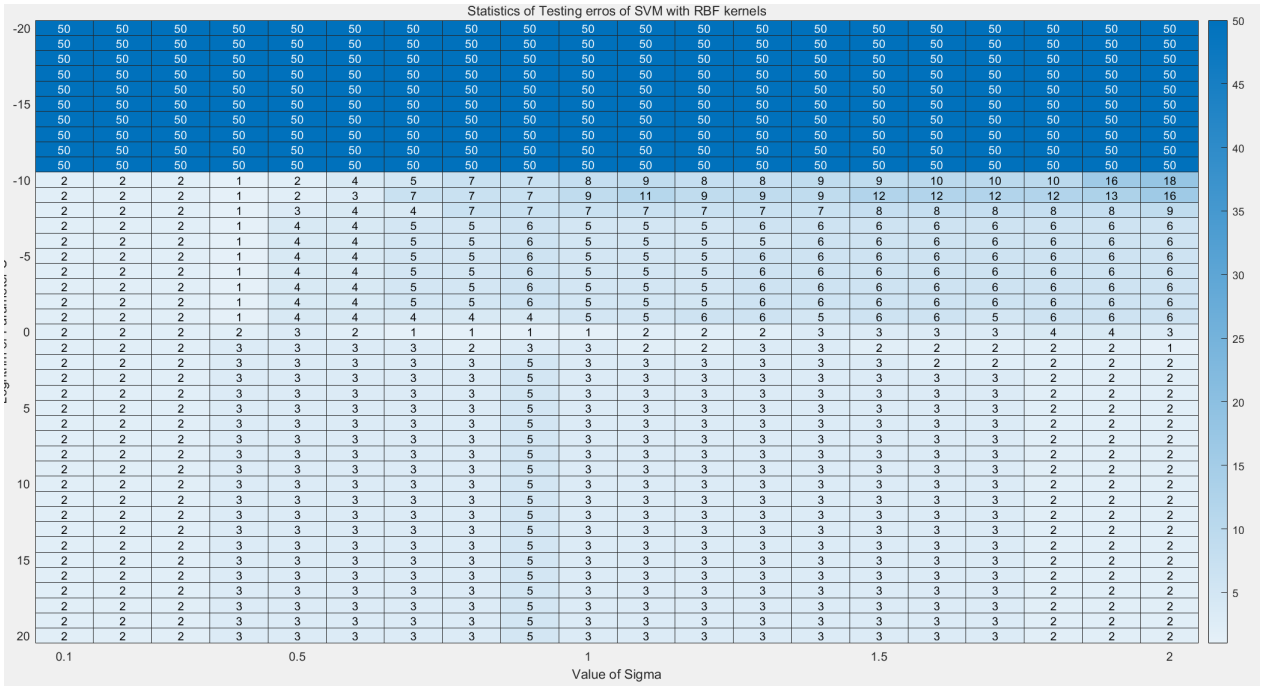

Figure 2.5 Testing results using RBF kernel

Note: The number in each cell is the number of support vectors data each case.

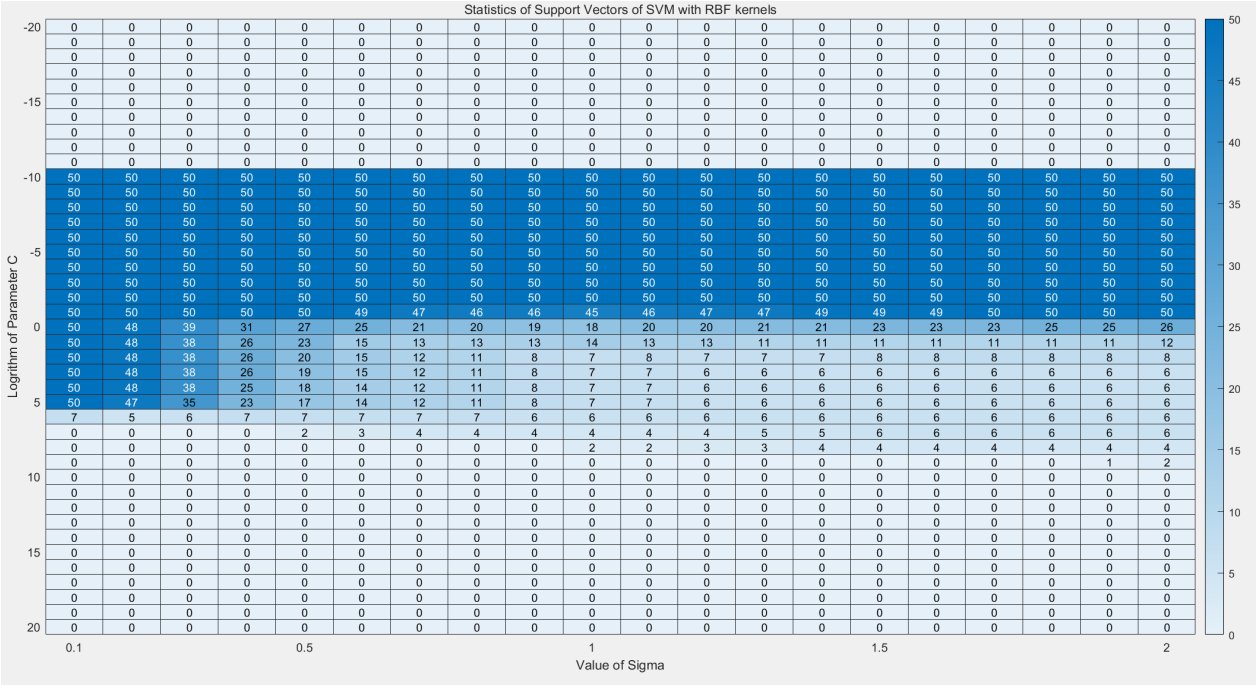

Figure 2.6 Testing results using RBF kernel

When using RBF kernels, the model is not good enough when the parameter C is too small. The increase of sigma and C will both improve the model's performance. The distribution of number of support vectors acts similarly to the model with linear kernel.

Problem 3

The pdf function is:

$$\begin{aligned} f(x|\alpha) &= \alpha^{-2} x \exp(-\frac{x}{\alpha}) \\ &= \frac{x}{\alpha^2 \cdot \exp(\frac{x}{\alpha})} \end{aligned}$$

which $x > 0, \alpha > 0$.

For samples $x_1 = 0.25, x_2 = 0.75, x_3 = 1.5, x_4 = 2.5, x_5 = 2.0$, the likelihood function $L(x, \alpha)$ is:

$$L(x, \alpha) = \prod_{i=1}^{5} f(x_i|\alpha)$$

The log likelihood function $l(x, \alpha)$ is:

$$\begin{aligned} l(x, \alpha) &= \ln L(x, \alpha) \\ &= \sum_{i=1}^{5} \ln f(x_i|\alpha) \end{aligned}$$

To find the maximum likelihood estimator for α, we need to compute the partial derivative of $l(x, \alpha)$ with respect to α first. We can notice that:

$$\begin{aligned} \frac{\partial \ln f(x_i|\alpha)}{\partial \alpha} &= \frac{\mathrm{d} \ln f(x_i|\alpha)}{\mathrm{d} f(x_i|\alpha)} \cdot \frac{\partial f(x_i|\alpha)}{\partial \alpha} \\ &= \frac{1}{f(x_i|\alpha)} \cdot \frac{\partial \frac{x_i}{\alpha^2 \exp(\frac{x_i}{\alpha})}}{\partial \alpha} \\ &= \frac{1}{f(x_i|\alpha)} \cdot \frac{\mathrm{d} \frac{x_i}{\alpha^2 \exp(\frac{x_i}{\alpha})}}{\mathrm{d}(\alpha^2 \exp(\frac{x_i}{\alpha}))} \cdot \frac{\partial(\alpha^2 \exp(\frac{x_i}{\alpha}))}{\partial \alpha} \\ &= \frac{1}{f(x_i|\alpha)} \cdot -\frac{x_i}{(\alpha^2 \exp(\frac{x_i}{\alpha}))^2} \cdot (2\alpha \exp(\frac{x_i}{\alpha}) + \alpha^2 \exp(\frac{x_i}{\alpha}) x_i \frac{-1}{\alpha^2}) \\ &= -\frac{1}{\alpha \exp(\frac{x_i}{\alpha})} \cdot (2\alpha - x_i) \exp(\frac{x_i}{\alpha}) \\ &= -\frac{(2\alpha - x_i)}{\alpha} = \frac{x_i}{\alpha} - 2 \end{aligned}$$

Thus, we can have:

$$\begin{aligned} \frac{\partial l(x, \alpha)}{\partial \alpha} &= \sum_{i=1}^{5} \frac{\partial \ln f(x_i|\alpha)}{\partial \alpha} \\ &= \sum_{i=1}^{5} (\frac{x_i}{\alpha} - 2) \end{aligned}$$

Solve the equation:

$$\begin{aligned}
\frac{\partial l(x,\alpha)}{\partial \alpha} &= 0 \\
\sum_{i=1}^{5}(\frac{x_i}{\alpha} - 2) &= 0 \\
\frac{0.25+0.75+1.5+2.5+2.0}{\alpha} - 10 &= 0 \\
10\alpha &= 7 \\
\alpha &= 0.7
\end{aligned}$$

Therefore, the maximum likelihood estimator for α is 0.7.