भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

Kernels and Regularization

Assignment -5 - CS3390

Kartik Srinivas - ES20BTECH11015

November 13, 2022

Index

1 Section 5.1

1.1 Shalev 11.2

The ides is that searching over all the spaces of functions wihtout validation is meaningful only when we have a very small number of samples for training. In this case, reducng the estimation error is the main goal, and therefore the main objective is to increase the number of samples for training the models $\mathfrak{h}_i$. On the contrary, if we have too many examples, then learning them individually on less data and the cross validating will reduce the modelling error incurred by choosing only one type of model. In this case it is highly probable that the best model may not be the one with lowest training error. In orer to mitigate this we will set aside some examples(namely, αm) for validation. error

1.2 ESL 7.10

To be done.

2 5.2

2.1 Q1

The following table summarizes the results of both SVM and ligistic regression results with **best C after cross validation**

Run_{number}	C_{logi}	$Score_{logi-reg}$	$Score_{logi-unreg}$	C_{svm}	$Score_{svm}$
1	10	0.8412	0.79365	10	0.873
2	10	0.793	0.714	10	0.793
3	10	0.873	0.809	10	0.873
4	100	0.773	0.79	100	0.873
5	10	0.88	0.8412	10	0.873

2.2 Q2

The following table summarizes the results for Ridge regression, however the Scikit Support vector regression model does not work on multiple feature output, so only ridge regression results are presented with **best α after Cross validation**

Run_{number}	α_{ridge}	$Score_{ridge-reg}$	$Score_{linear-unreg}$
1	10	0.519	0.389
2	10	0.4875	0.2732
3	10	0.5039	0.327
4	100	0.5185	0.399
5	10	0.546	0.39

3 Section 5.3

3.1 16.3 Shalev and Schwartz

The objective is to solve the optimization problem in the representer theorem format. After substituting the value of $w = \alpha^T X$ the optimization problem becomes

$$\underset{\alpha}{min}\frac{\lambda}{2}\alpha^T G\alpha + \frac{1}{2m}\sum_{i=1}^{i=m}(\alpha^T G_i - y_i)^2$$

Now lets take the derivative of the above equation with respect to α and set it to zero.

$$\frac{\partial}{\partial\alpha}\frac{\lambda}{2}\alpha^T G\alpha + \frac{1}{2m}\sum_{i=1}^{i=m}(\alpha^T G_i - y_i)^2 = 0$$

$$\lambda G\alpha + \frac{1}{m}\sum_{i=1}^{i=m}(\alpha^T G_i\alpha - y_i)G_i = 0$$

On rearranging with $G_i^T\alpha = \alpha^T G_i$,

$$\lambda G\alpha + \frac{1}{m}\sum_{i=1}^{i=m}G_i G_i^T\alpha = \sum_{i=1}^{i=m}y_i G_i$$

Taking alpha common we get,

$$\alpha^* = \left(\lambda G + \frac{1}{m}\sum_{i=1}^{i=m}G_i G_i^T\right)^{-1}\sum_{i=1}^{i=m}y_i G_i$$

3.2 16.4 Shalev and Schwartz

We will find a mapping ψ that achieves this.

$$min(x,y) = \sum_{i=1}^{i=N} I_{\{x \leq i\}} I_{\{y \leq i\}}$$

The above function can be summarized using a dot product

$$min(x,y) = \begin{bmatrix} 1 & 1 & 1 & 1 & \dots & 0 & 0 \end{bmatrix} \times \begin{bmatrix} 1 \\ 1 \\ 1 \\ \\ . \\ . \\ 0 \\ 0 \end{bmatrix} = \psi(x)^T \psi(y)$$

$\psi(x)$ is a vector that has a set of ones until it reaches the x'th position in the series 1...n

3.2.1 Mohri 6.1

We know that $K(x,y) = \psi(x)^T \psi(y)$ for some $\psi : \chi \rightarrow H$. For the new K' set $\psi(l)' = \frac{\psi(l)}{\alpha(l)}$ $\alpha(x) > 0$ Then

$$K'(x,y) = \psi'(x)^T \psi'(y)$$

3.2.2 Mohri 6.2a

$$\begin{aligned} cos(x-y) &= cosxcosy + sinxsiny \quad (1) \\ &= \psi(x)^T \psi(y) + \phi(x)^T \phi(y) \quad (2) \\ &= k_1(x,y) + k_2(x,y) \quad (3) \end{aligned}$$

3.2.3 Part 6.2b

Use the exact same technique as [3.2.2] instead of $\psi(x) = cos(x)$ let $\psi(x) = cos(x^2)$

3.2.4 Part 6.2c

Here let us just **Break the sum into individual terms and the product into its basic unit**

$$cos(x_i^2 - y_i^2) = cos(\mathbf{1_i}^T\mathbf{x})cos(\mathbf{1_i}^T\mathbf{y}^2) + cos(\mathbf{1_i}^T\mathbf{x})cos(\mathbf{1_i}^T\mathbf{y}^2)$$

The iea sis that now it has been split into $\psi_i(x) = cos(\mathbf{1_i}^T\mathbf{x})$ and $\phi(x) = sin(\mathbf{1_i}^T\mathbf{x})$ And The sum and product of kernels remains a kernel This completes the proof.

3.2.5 Part 6.2d

The idea was given as a hint to use monotonicity, consider the function

$$f(\theta) = \sum_{i,j} \frac{c_i c_j \theta^{x_i + x_j}}{x_1 + x_j}$$

Note that $f(\theta) = c^T Gc$ Where G is my kernel Matrix that I must show is positive semidefinite.

$$\frac{\partial f(\theta)}{\partial \theta} = \sum_{i,j} \frac{c_i c_j (x_i + x_j)\theta^{x_i + x_j - 1}}{x_1 + x_j} = \sum_{i,j} c_i c_j a^{x_i} a^{x_j}$$

Now we can see that the function $f(\theta)$ is monotonic as the RHS boils down to $c^T Kc$ where $k(x, y) = a^x \times a^y$ which is a kernel.
Now, since the function is monotonic, $f(1) \geq f(0) = 0$. So we get that $G \succcurlyeq 0$
Hence proved.

3.2.6 Part 6.2e

$$cos\angle(x, x') = \frac{x^T x'}{\|x\|\|x'\|} = \psi(x)^T \psi(x')$$

where $\psi(x) = \frac{x}{\|x\|}$ Hence proved

3.2.7 Part 6.2f

Lests break the exponent down

$$exp(-\lambda(sin(x-y)^2)) = \exp\big(-\lambda(1 - cos^2(x - y))\big) = \exp\big(\lambda cos^2(x - y)\big)\, times \exp(-\lambda)$$

Now the second part is just a positive constant , let us look at the first part, it is of the form

$$\exp(\lambda k(x)) = 1 + \lambda k(x) + \frac{\lambda^2 k(x)^2}{2!} + \frac{\lambda^3 k(x)^3}{3!} + \ldots k_n(x)$$

We already proved that cos(x - y) is a kernel in 3.2.2 and we know that $\lambda > 0$. This is sufficient along with the **sum,product and limiting theorems** to prove that the first part is a kernel.

3.2.8 Part 6.2h

Using the integrals in the Hints you can see that the function is a product of the two integrals

$$k(x,y) = min(x,y) - xy = x(1-y) \ \text{ if } x \leq y \text{ and } y(1-x) \text{ otherwise}$$

Now consider the Integrals

$$\int_0^1 1_{t\leq x}1_{t\leq y}\ dt = \int_0^1 \psi(x)^T\psi(y)\ dt = \textbf{ Limiting sum of Kernels} \tag{4}$$

$$I_1 = min(x,y) = x\ or\ y \tag{5}$$

$$\int_0^1 1_{t\geq x}1_{t\geq y}\ dt = \int_0^1 \phi(x)^T\phi(y)\ dt = \textbf{Limiting sum of Kernels} \tag{6}$$

$$I_2 = 1 - max(x,y) = 1 - x\ or\ 1 - y \tag{7}$$

We can see that the function is a product of the two integrals which are both Kernels. Hence proved.

3.3 Q3

The plots are as follows

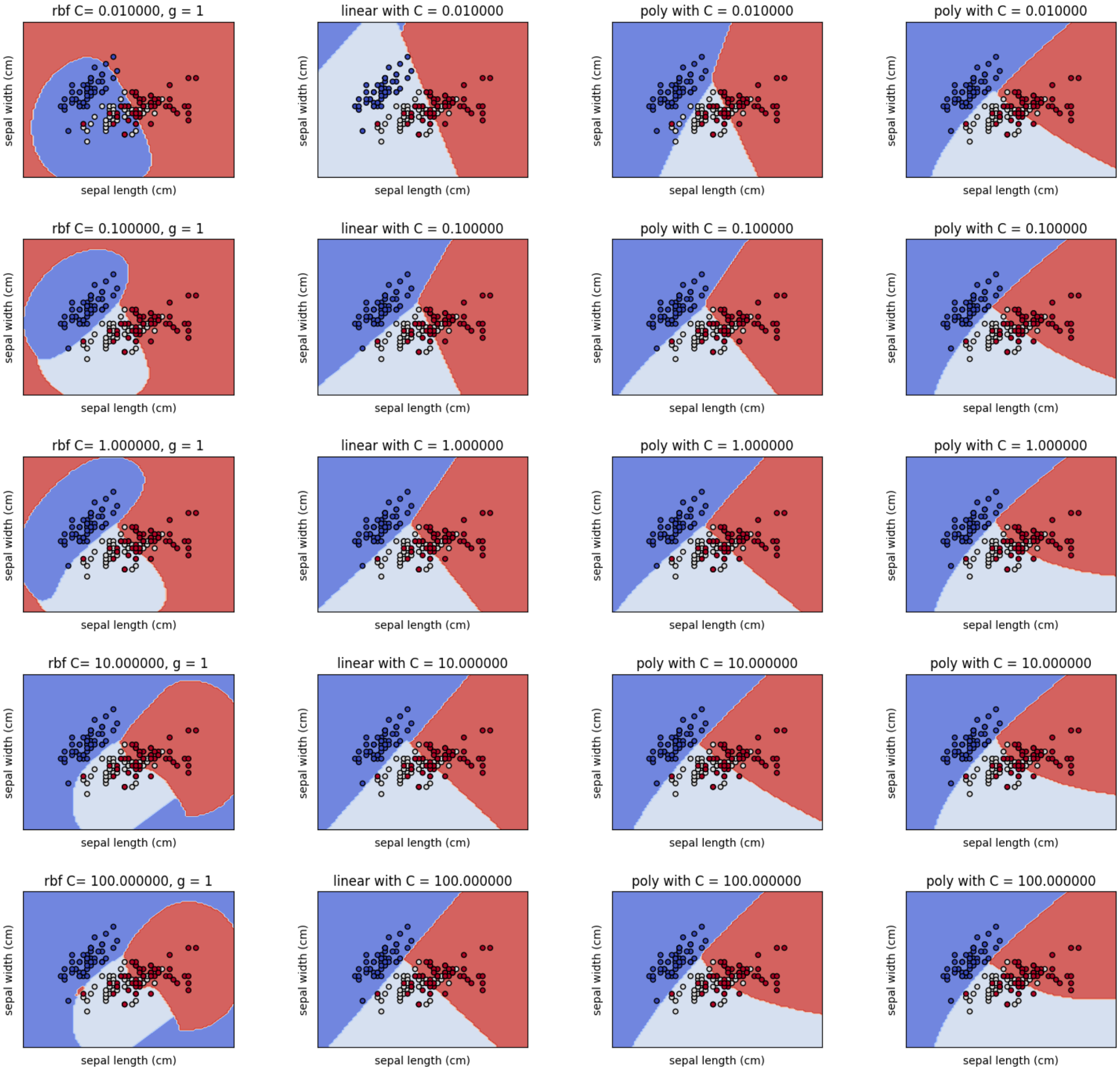
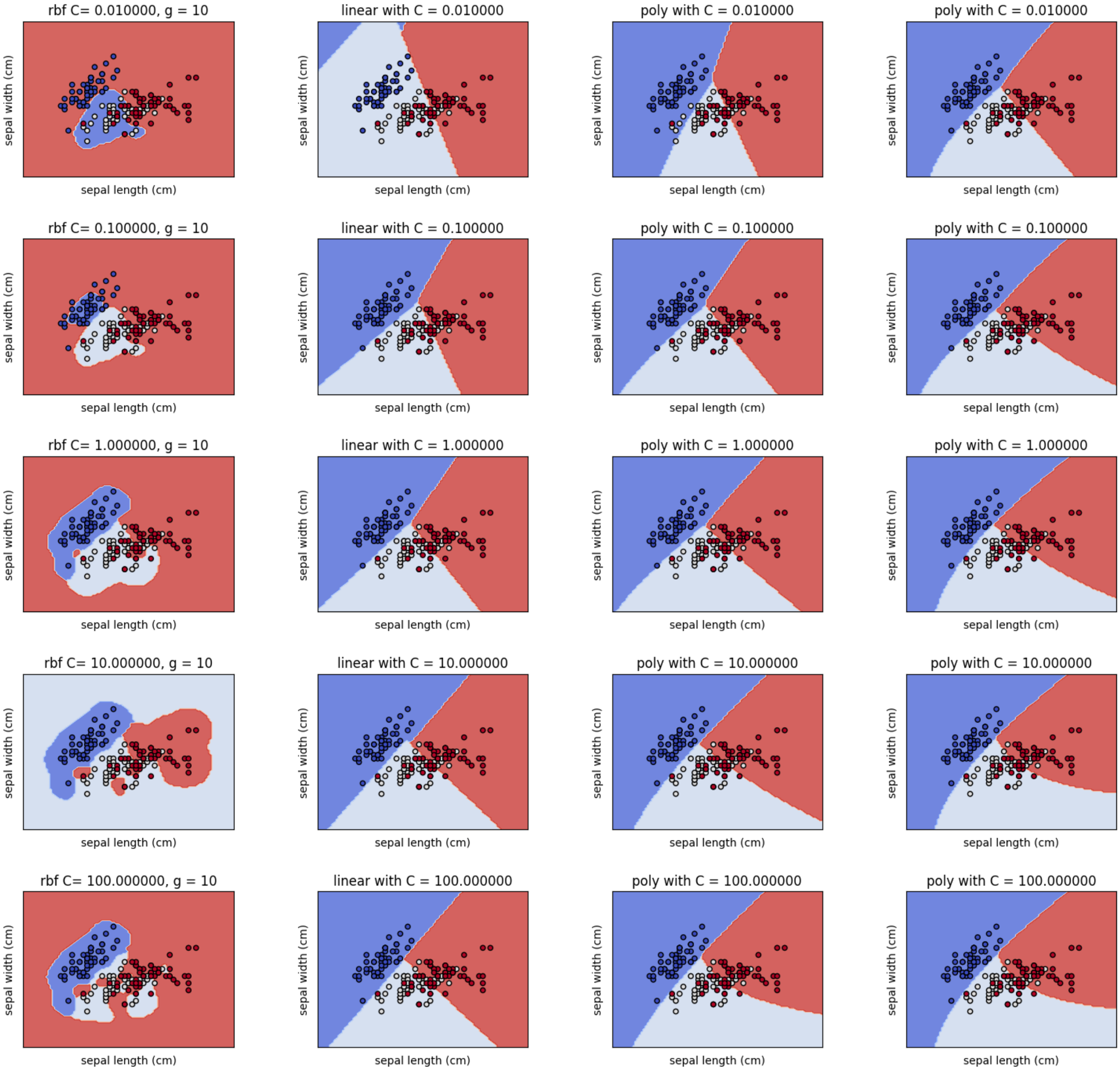


Figure 1: $\gamma = 1$



rbf C= 0.010000, g = 10

linear with C = 0.010000

poly with C = 0.010000

poly with C = 0.010000

rbf C= 0.100000, g = 10

linear with C = 0.100000

poly with C = 0.100000

poly with C = 0.100000

rbf C= 1.000000, g = 10

linear with C = 1.000000

poly with C = 1.000000

poly with C = 1.000000

rbf C= 10.000000, g = 10

linear with C = 10.000000

poly with C = 10.000000

poly with C = 10.000000

rbf C= 100.000000, g = 10

linear with C = 100.000000

poly with C = 100.000000

poly with C = 100.000000

Figure 2: $\gamma = 10$

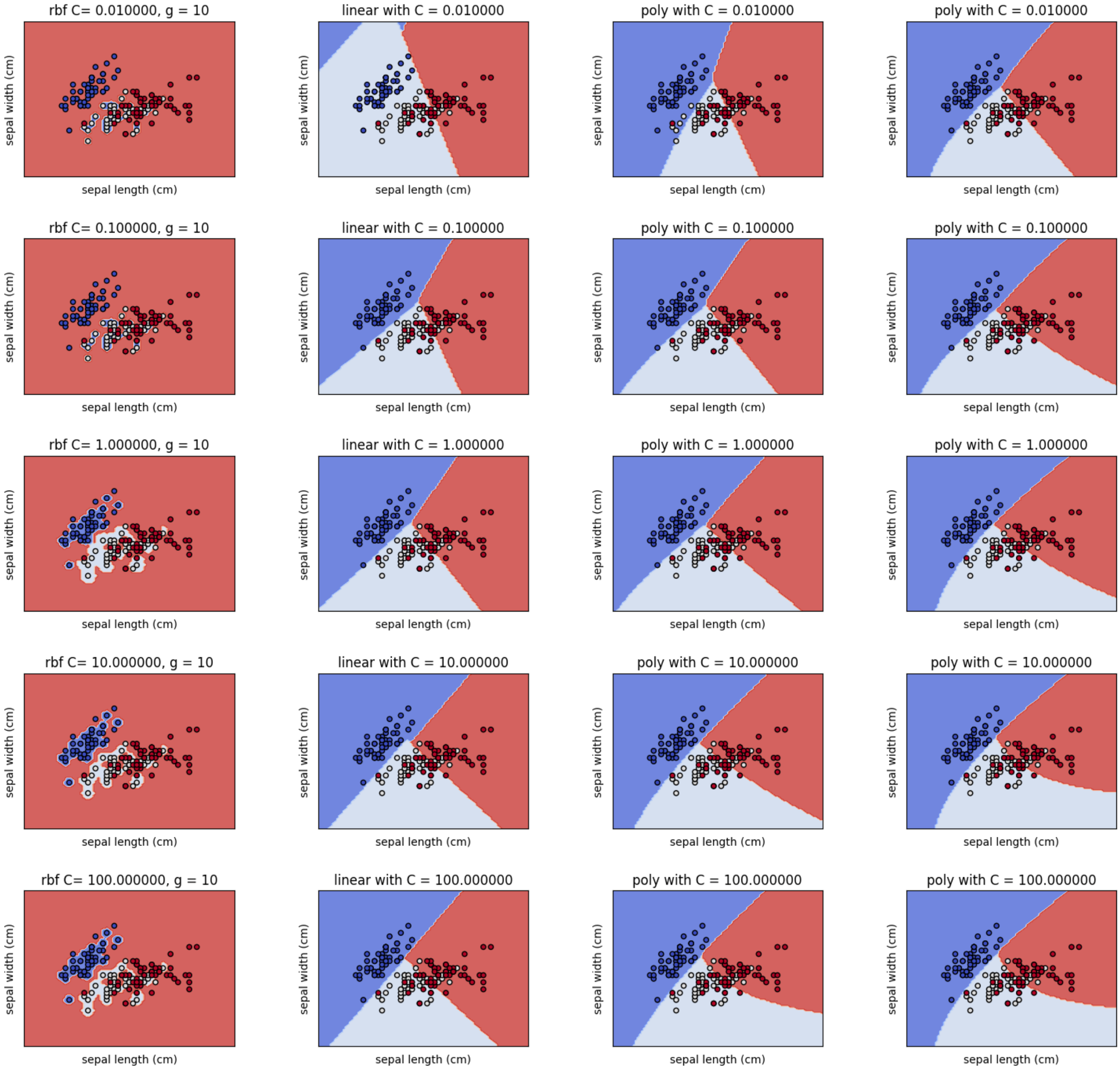


Figure 3: $\gamma = 100$

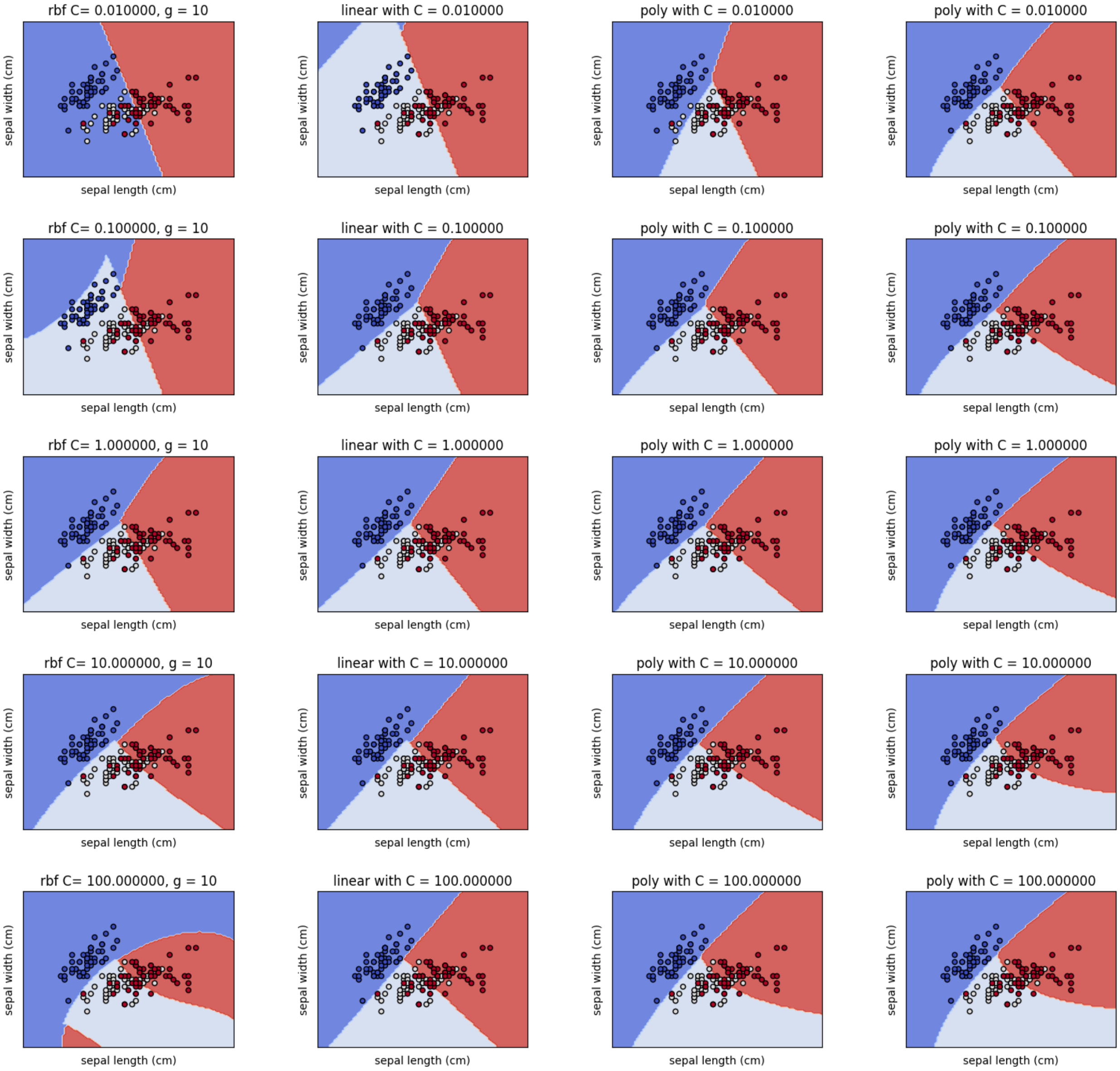


Figure 4: $\gamma = 0.1$