Theoretical questions

1. Let us look on the defined multiclass SVM problem:

$$f(\boldsymbol{w}_1, \dots, \boldsymbol{w}_K) = \frac{1}{n}\sum_{i=1}^{n} \ell(\boldsymbol{w}_1, \dots, \boldsymbol{w}_K, \boldsymbol{x}_i, y_i)$$

Where our loss function is defined as:

$$\ell(\boldsymbol{w}_1, \dots, \boldsymbol{w}_K, \boldsymbol{x}_i, y_i) =$$
$$\max_{j\in[K]}\left(\boldsymbol{w}_j \cdot \boldsymbol{x}_i - \boldsymbol{w}_{y_i}\boldsymbol{x}_i + \mathbb{1}_{\{j\neq y_i\}}\right) =$$
$$\max_{j\in[K]}\left(\left(\boldsymbol{w}_j - \boldsymbol{w}_{y_i}\right) \cdot \boldsymbol{x}_i + \mathbb{1}_{\{j\neq y_i\}}\right)$$

Given $(\boldsymbol{x}_i, y_i)$ sample, we notice that $\ell(\boldsymbol{w}_1, \dots, \boldsymbol{w}_K, \boldsymbol{x}_i, y_i) \geq 0$, because taking $j = y_i$ satisfies that $\left(\boldsymbol{w}_j - \boldsymbol{w}_{y_i}\right) \cdot \boldsymbol{x}_i + \mathbb{1}_{\{j\neq y_i\}} = 0$, and therefore $\max_{j\in[K]}\left(\left(\boldsymbol{w}_j - \boldsymbol{w}_{y_i}\right) \cdot \boldsymbol{x}_i + \mathbb{1}_{\{j\neq y_i\}}\right) \geq 0$.

We also notice that given $\boldsymbol{w}_1, \dots, \boldsymbol{w}_K$ s.t $f(\boldsymbol{w}_1, \dots, \boldsymbol{w}_K) = 0$, our data is linearly separable.

Proof:
By contradiction, we assume that our data isn't linearly separable.
This means that there is $(\boldsymbol{x}_i, y_i)$ s.t $y_i \neq argmax_y \boldsymbol{w}_y \cdot \boldsymbol{x}_i$.
We denote the argmax with y^*, this means that

$$\boldsymbol{w}_{y^*} \cdot \boldsymbol{x}_i - \boldsymbol{w}_{y_i} \cdot \boldsymbol{x}_i > 0 \Rightarrow$$
$$\left(\boldsymbol{w}_{y^*} - \boldsymbol{w}_{y_i}\right) \cdot \boldsymbol{x}_i + \mathbb{1}_{\{j\neq y_i\}} > 0 \Rightarrow$$
$$\ell(\boldsymbol{w}_1, \dots, \boldsymbol{w}_K, \boldsymbol{x}_i, y_i) > 0$$

This means that $f(\boldsymbol{w}_1, \dots, \boldsymbol{w}_K) > 0$, contradiction to our assumption.

Thus, our strategy is to find $\boldsymbol{w}_1, \dots, \boldsymbol{w}_K$ s.t $f(\boldsymbol{w}_1, \dots, \boldsymbol{w}_K) = 0$, this implies that also for the minimizer f is 0, thus the minimizer linearly separates our data.

We assume that there exist $\boldsymbol{w}_1^*, \dots, \boldsymbol{w}_K^*$ that linearly separate our data.
If $f(\boldsymbol{w}_1^*, \dots, \boldsymbol{w}_K^*) = 0$, we're finished, otherwise this means that there is $i \in [K]$ s.t $\ell(\boldsymbol{w}_1^*, \dots, \boldsymbol{w}_K^*, \boldsymbol{x}_i, y_i) = \max_{j\in[K]}\left(\left(\boldsymbol{w}_j^* - \boldsymbol{w}_{y_i}^*\right) \cdot \boldsymbol{x}_i + \mathbb{1}_{\{j\neq y_i\}}\right) > 0$
Let's mark the set of all $i's$ that satisfies the above with I.
Given $i \in I$, there is $j \neq y_i$ s.t $\left(\boldsymbol{w}_j^* - \boldsymbol{w}_{y_i}^*\right) \cdot \boldsymbol{x}_i + 1 > 0 \Leftrightarrow \left(\boldsymbol{w}_j - \boldsymbol{w}_{y_i}\right) \cdot \boldsymbol{x}_i > -1$
From reliability we know that $\boldsymbol{w}_{y_i}^* \cdot \boldsymbol{x}_i > \boldsymbol{w}_j^* \cdot \boldsymbol{x}_i \Leftrightarrow \left(\boldsymbol{w}_j^* - \boldsymbol{w}_{y_i}^*\right) \cdot \boldsymbol{x}_i < 0$.

If we define $C_{i,j} = \left|\frac{1}{\left(w_j^* - w_{y_i}^*\right)\cdot x_i}\right|$ we get that

$$\left(C_{i,j}\boldsymbol{w}_j^* - C_{i,j}\boldsymbol{w}_{y_i}^*\right) \cdot \boldsymbol{x}_i = C_{i,j}\left(\boldsymbol{w}_j^* - \boldsymbol{w}_{y_i}^*\right) \cdot \boldsymbol{x}_i = -1$$

This means that

$$\left(C_{i,j}\boldsymbol{w}_j^* - C_{i,j}\boldsymbol{w}_{y_i}^*\right) \cdot \boldsymbol{x}_i + 1 = C_{i,j}\left(\boldsymbol{w}_j^* - \boldsymbol{w}_{y_i}^*\right) \cdot \boldsymbol{x}_i + 1 = -1 + 1 = 0$$

Let's define $C = \max_{i \in I} \max_{j \in [K] \backslash \{y_i\}} C_{i,j}$.
Because of our assumption that $(\boldsymbol{w}_1^*, \ldots, \boldsymbol{w}_K^*) > 0,\ \delta > 0$.
Let's mark $\boldsymbol{w}_i = C\boldsymbol{w}_{\boldsymbol{i}}^*$**.**

Given $(\boldsymbol{x}_i, y_i)$ s.t $i \in I$ and $j \in [K] \backslash \{y_i\}$:

$$\left(\boldsymbol{w}_j - \boldsymbol{w}_{y_i}\right) \cdot \boldsymbol{x}_i + \mathbb{1}_{\{j \neq y_i\}} = \left(C\boldsymbol{w}_j^* - \boldsymbol{C}\boldsymbol{w}_{y_i}^*\right) \cdot \boldsymbol{x}_i + 1 \leq -1 + 1 = 0$$

And for $i \notin \mathrm{I}, j \in [K]$ $\left(\boldsymbol{w}_j^* - \boldsymbol{w}_{y_i}^*\right) \cdot \boldsymbol{x}_i + \mathbb{1}_{\{j \neq y_i\}} \leq 0$, which implies that

$$\left(\boldsymbol{w}_j - \boldsymbol{w}_{y_i}\right) \cdot \boldsymbol{x}_i + \mathbb{1}_{\{j \neq y_i\}} = \left(\boldsymbol{C}\boldsymbol{w}_j^* - \boldsymbol{C}\boldsymbol{w}_{y_i}^*\right) \cdot \boldsymbol{x}_i + \mathbb{1}_{\{j \neq y_i\}} \leq 0$$

This means that for any $(\boldsymbol{x}_i, y_i)$, $\ell(\boldsymbol{w}_1, \ldots, \boldsymbol{w}_K, \boldsymbol{x}_i, y_i) = 0$.
Therefore, $f(\boldsymbol{w}_1, \ldots, \boldsymbol{w}_K) = 0$.

2. Let us solve SVM problem for two distinct $x_1, x_2 \in \mathbb{R}^d$ s.t $y_1 = 1, y_2 = -1$.
   We want to solve the following objective

$$\min_{w} 0.5\|w\|_2^2$$
$$s.t\ y_i(w \ \cdot x_i) \geq 1\ \forall i = 1{,}2$$

Let consider the dual problem obtained in class:

$$\max_{\alpha} \sum_i \alpha_i - 0.5 \sum_{i,j} \alpha_i \alpha_j y_i y_j x_i \cdot x_j$$
$$s.t \sum_i \alpha_i y_i = 0$$
$$\alpha_i \geq 0$$

In our case the objective we want to maximize is:

$$\alpha_1 + \alpha_2 - 0.5(\alpha_1^2\|x_1\|_2^2 - 2\alpha_1\alpha_2 x_1 \cdot x_2 + \alpha_2^2\|x_2\|_2^2)$$
$$= \alpha_1 + \alpha_2 - 0.5\|\alpha_1 x_1 - \alpha_2 x_2\|_2^2$$

And our conditions turn to

$$\alpha_1 - \alpha_2 = 0$$

If we plug this into the objective, we receive:

$$2\alpha_1 - 0.5\alpha_1^2\|x_1 - x_2\|_2^2$$

The maximum of this obtained when $\alpha_1 = -\frac{2}{2*(-0.5)\|x_1-x_2\|_2^2} = \frac{2}{\|x_1-x_2\|_2^2} > 0$.
Because x_1, x_2 are two distinct points we divide by the norm of the differences.
Thus, we got that $\alpha_1 = \alpha_2 = \frac{2}{\|x_1-x_2\|_2^2} > 0$.

From KKT we can obtain that

- $w^* = \sum_i \alpha_1 y_i x_i = \frac{2}{\|x_1-x_2\|_2^2} x_1 - \frac{2}{\|x_1-x_2\|_2^2} x_2 = \frac{2}{\|x_1-x_2\|_2^2}(x_1 - x_2)$
- For any support vector $1 = y_i(w^* \cdot x_i + b^*)$
  In our case, we got 2 support vectors.
  Using (x_1, y_1) we get that

$$b^* = 1 - w^* \cdot x_1 = 1 - \frac{2}{\|x_1 - x_2\|_2^2}(x_1 - x_2) \cdot x_1$$

For conclusion, the results of SVM are:

$$w^* = \frac{2}{\|x_1 - x_2\|_2^2}(x_1 - x_2)$$
$$b^* = 1 - \frac{2}{\|x_1 - x_2\|_2^2}(x_1 - x_2) \cdot x_1$$

3.

a. Let consider the problem

$$\min_{\boldsymbol{w},b,\boldsymbol{\xi}} 0.5\|w\|_2^2 + \frac{C}{2}\sum_{i=1}^{n} \xi_i^2$$
$$s.t.\ y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \geq 1 - \xi_i \quad \forall 1 \leq i \leq n$$

With $(*)$, and opt^* the optimal value for $(*)$

And the problem

$$\min_{\boldsymbol{w},b,\boldsymbol{\xi}} 0.5\|w\|_2^2 + \frac{C}{2}\sum_{i=1}^{n} \xi_i^2$$
$$s.t.\ y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \geq 1 - \xi_i \quad \forall 1 \leq i \leq n$$
$$\xi_i \geq 0 \quad \forall 1 \leq i \leq n$$

With $(**)$, and opt^{**} the optimal value for $(**)$

We prove that $opt^* = opt^{**}$ by showing that:

a) $opt^* \leq opt^{**}$

b) $opt^{**} \leq opt^*$

Proving a:

Because in $(*)$ and in $(**)$ we want to minimize the same objective, but $(**)$ has more constraints, it follows that $opt^* \leq opt^{**}$.

Proving b:

Let there be $\boldsymbol{w}^*, b^*, \boldsymbol{\xi}^*$ that accomplish opt^*.

We will define a solution to $(**)$ using them:

$$\boldsymbol{w}^{**} = \boldsymbol{w}^*, b^{**} = b^*, \boldsymbol{\xi}_i^{**} = |\boldsymbol{\xi}_i^*|$$

Let us show this is a solution to $(**)$:

1. From the definition of $\boldsymbol{\xi}^{**}, \boldsymbol{\xi}_i^{**} \geq 0 \ \ \forall 1 \leq i \leq n$
2. We notice that $1 - \boldsymbol{\xi}_i^* \geq 1 - \boldsymbol{\xi}_i^{**} \quad \forall 1 \leq i \leq n$
   therefore, $y_i\left(\boldsymbol{w}^{**T}\boldsymbol{x}_i + b^{**}\right) = y_i\left(\boldsymbol{w}^{*T}\boldsymbol{x}_i + b^*\right) \geq 1 - \boldsymbol{\xi}_i^* \geq 1 - \boldsymbol{\xi}_i^{**}$

This means that $\boldsymbol{w}^{**}, b^{**}, \boldsymbol{\xi}^{**}$ is a solution to $(**)$.

And we can see that $0.5\|\boldsymbol{w}^{**}\|_2^2 + \frac{C}{2}\sum_{i=1}^{n} \boldsymbol{\xi}^{**2}_i = 0.5\|\boldsymbol{w}^*\|_2^2 + \frac{C}{2}\sum_{i=1}^{n} \boldsymbol{\xi}^{*2}_i = opt^*$

This means that $opt^{**} \leq opt^*$.

b.

We arrange our constraints to the form: $0 \geq 1 - \xi_i - y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \quad \forall 1 \leq i \leq n$

And we obtain that the Lagrangian is:

$$\mathcal{L}(\boldsymbol{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}) = 0.5\|\boldsymbol{w}\|_2^2 + \frac{C}{2}\sum_{i=1}^{n} \xi_i^2 + \sum_{i=1}^{n} \alpha_i(1 - \xi_i - y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b))$$

c. We compute the partial derivatives with respect to $\boldsymbol{w}, b, \boldsymbol{\xi}$:

$$\nabla_w \mathcal{L}(\boldsymbol{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}) = \boldsymbol{w} - \sum_{i=1}^{n} \boldsymbol{\alpha}_i y_i \boldsymbol{x}_i = 0 \Rightarrow \boldsymbol{w} = \sum_{i=1}^{n} \boldsymbol{\alpha}_i y_i \boldsymbol{x}_i$$

$$\nabla_\beta \mathcal{L}(\boldsymbol{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}) = \sum_{i=1}^{n} \boldsymbol{\alpha}_i y_i = 0$$

$$\frac{\partial}{\partial \boldsymbol{\xi}_i} \mathcal{L}(\boldsymbol{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}) = C\boldsymbol{\xi}_i - \boldsymbol{\alpha}_i = 0 \Rightarrow \boldsymbol{\xi}_i = \frac{\boldsymbol{\alpha}_i}{C}$$

(If $C = 0$ then $\boldsymbol{\alpha} = \boldsymbol{0}$ and the dual problem is to maximize a const function with value $\boldsymbol{0}$)

If we plug this into the Lagrangian we get that

$$g(\alpha) = \min_{\boldsymbol{w}, b, \boldsymbol{\xi}} \mathcal{L}(\boldsymbol{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}) =$$

$$0.5\|\boldsymbol{w}\|_2^2 + \frac{C}{2}\sum_{i=1}^{n} \xi_i^2 - \sum_{i=1}^{n} \boldsymbol{\alpha}_i \xi_i + \sum_{i=1}^{n} \boldsymbol{\alpha}_i - \boldsymbol{w}^T \sum_{i=1}^{n} \boldsymbol{\alpha}_i y_i \, \boldsymbol{x}_i =$$

$$\sum_{i=1}^{n} \boldsymbol{\alpha}_i + \frac{C}{2}\sum_{i=1}^{n} \frac{\boldsymbol{\alpha}_i^2}{C^2} - \sum_{i=1}^{n} \frac{\boldsymbol{\alpha}_i^2}{C} + 0.5\|\boldsymbol{w}\|_2^2 - \boldsymbol{w}^T\boldsymbol{w} =$$

$$\sum_{i=1}^{n} \boldsymbol{\alpha}_i - \frac{1}{2C}\sum_{i=1}^{n} \boldsymbol{\alpha}_i^2 - 0.5 \sum_{1 \leq i < j \leq n} \boldsymbol{\alpha}_i \boldsymbol{\alpha}_j y_i y_j \boldsymbol{x}_i \cdot \boldsymbol{x}_j$$

d. The dual problem is:

$$\max_{\alpha} g(\alpha) = \max_{\alpha} \sum_{i=1}^{n} \boldsymbol{\alpha}_i - \frac{1}{2C}\sum_{i=1}^{n} \boldsymbol{\alpha}_i^2 - 0.5 \sum_{1 \leq i, j \leq n} \boldsymbol{\alpha}_i \boldsymbol{\alpha}_j y_i y_j \boldsymbol{x}_i \cdot \boldsymbol{x}_j$$

$$s.t. \sum_{i=1}^{n} \boldsymbol{\alpha}_i y_i = 0$$

$$\boldsymbol{\alpha}_i \geq 0$$

4. Let there be $\boldsymbol{w}^*$ the solution for hard SVM and let there be $\boldsymbol{w}', \xi'$ the solution for soft SVM.
   We assume that $\|\boldsymbol{w}^*\|^2 \leq C$.

   From the definition of $\boldsymbol{w}^*$, $\forall i \quad y_i \boldsymbol{w}^* \cdot \boldsymbol{x}_i \geq 1$.
   If we define $\xi^* = \mathbf{0}$, we see that $\boldsymbol{w}^*, \xi^*$ satisfies the soft SVM constraints.

   Therefore,

$$0.5\|\boldsymbol{w}'\|^2 + C\sum_{i=1}^{n} \xi_i' \leq 0.5\|\boldsymbol{w}^*\|^2$$

   And because $\|\boldsymbol{w}^*\|^2 \leq C$ we get that

$$0.5\|\boldsymbol{w}'\|^2 + C\sum_{i=1}^{n} \xi_i' \geq 0.5\|\boldsymbol{w}'\|^2 + \|\boldsymbol{w}^*\|^2\sum_{i=1}^{n} \xi_i'$$

   To conclude, we got that

$$0.5\|\boldsymbol{w}'\|^2 + \|\boldsymbol{w}^*\|^2\sum_{i=1}^{n} \xi_i' \leq 0.5\|\boldsymbol{w}^*\|^2 \Leftrightarrow$$

$$0.5\|\boldsymbol{w}'\|^2 \leq 0.5\|\boldsymbol{w}^*\|^2 - \|\boldsymbol{w}^*\|^2\sum_{i=1}^{n} \xi_i' \Leftrightarrow$$

$$0.5\|\boldsymbol{w}'\|^2 \leq \|\boldsymbol{w}^*\|^2\left(0.5 - \sum_{i=1}^{n} \xi_i'\right) \Leftrightarrow$$

$$\frac{1}{2} - \sum_{i=1}^{n} \xi_i' \geq \frac{1}{2}\frac{\|\boldsymbol{w}'\|^2}{\|\boldsymbol{w}^*\|^2} \Leftrightarrow$$

$$\sum_{i=1}^{n} \xi_i' \leq \frac{1}{2}\left(1 - \frac{\|\boldsymbol{w}'\|^2}{\|\boldsymbol{w}^*\|^2}\right) < 1$$

   And this implies that $\forall i \;\; 0 \leq \xi_i' < 1$.
   And because of that, $\forall i \;\; y_i \mathbf{w}' \cdot \boldsymbol{x}_i \geq 1 - \xi_i' > 0$

   And this means that $\boldsymbol{w}'$ separates the data.

5. Let there be $S=\{x_1,\dots,x_n\}\subset\mathbb{R}$ distinct real numbers, and $q\geq n$.

   Given x_i, x_j

$$(K_S)_{i,j}=K(x_i,x_j)=(1+x_ix_j)^q=\sum_{k=0}^{q}\binom{q}{k}(x_ix_j)^k=$$
$$\sum_{k=0}^{q}\binom{q}{k}(x_ix_j)^k=$$
$$\sum_{k=0}^{q}\binom{q}{k}^{\frac{1}{2}}x_i^k\binom{q}{k}^{\frac{1}{2}}x_j^k$$

   This means that if we define $\phi(x)_k=\binom{q}{k-1}^{\frac{1}{2}}x^{k-1}$ for $k=\ 1,\dots,q+1$ we get that $K(x_i,x_j)=\phi(x_i)\phi(x_j)$.

   If we examine $\Phi_S=\begin{pmatrix}\phi(x_1)^T\\ \vdots\\ \phi(x_n)^T\end{pmatrix}$, we see that

$$\Phi_S=\begin{pmatrix}1 & \binom{q}{1}^{\frac{1}{2}}x_1 & \binom{q}{2}^{\frac{1}{2}}x_1^2 & \cdots & \binom{q}{q}^{\frac{1}{2}}x_1^q\\ \vdots & \vdots & \vdots & \vdots & \vdots\\ 1 & \binom{q}{1}^{\frac{1}{2}}x_n & \binom{q}{2}^{\frac{1}{2}}x_n^2 & \cdots & \binom{q}{q}^{\frac{1}{2}}x_n^q\end{pmatrix}$$

   The matrix $X_S=\begin{pmatrix}1 & x_1 & x_1^2 & \cdots & x_1^q\\ \vdots & \vdots & \vdots & \vdots & \vdots\\ 1 & x_n & x_n^2 & \cdots & x_n^q\end{pmatrix}$ is Vandermonde matrix, and therefore its rank is n.

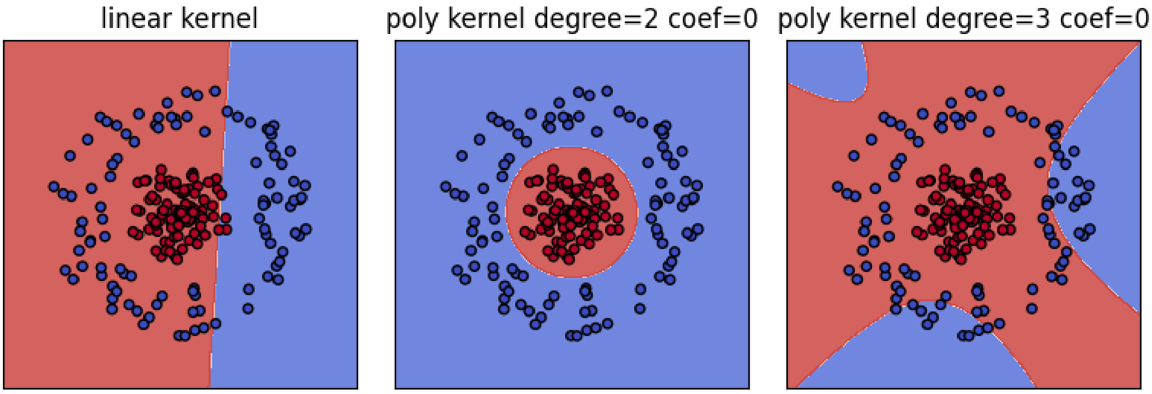
   We notice that Φ_S is obtained from X_S by multiplying the i-th column by $\binom{q}{i-1}^{\frac{1}{2}}$. And we remember that column operations don't change the rank of the matrix, therefore $rank(\Phi_S)=n$.

   And according to the hint, this imply that hard SVM achieves zero training error.

Programming questions

1.

a.

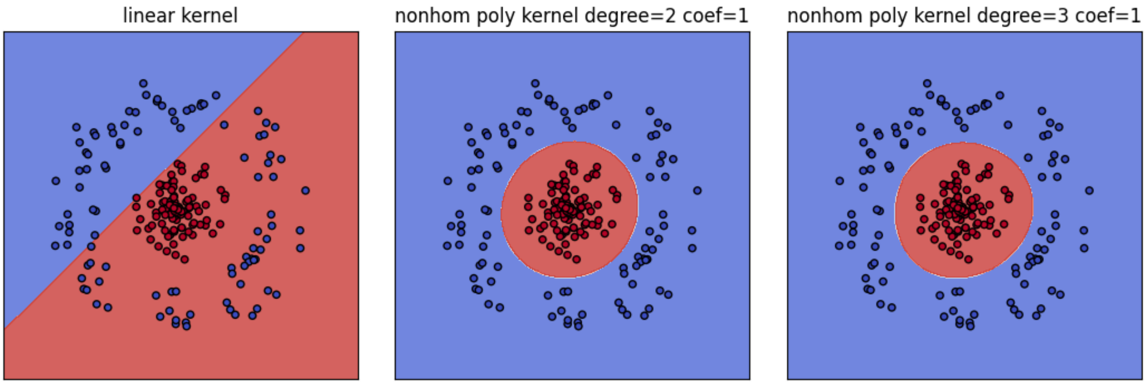


This graph was plotted using regularization parameter $C = 10$.
We can see that using homogenous polynomial kernel with degree 2 separated our data perfectly.
This happened because in the feature space that the polynomial kernel with degree 2 induces, our data is linearly separable.
For example, assuming our data is in $\mathbb{R}^2$.

$$K(x, x') = (x \cdot x')^2 = (x_1 x_1' + x_2 x_2')^2 = x_1^2 {x_1'}^2 + 2x_1 x_2 x_1' x_2' + x_2^2 x_2'^2 = \begin{pmatrix} x_1^2 \\ \sqrt{2} x_1 x_2 \\ x_2^2 \end{pmatrix} \cdot \begin{pmatrix} x'^2_1 \\ \sqrt{2} x'_1 x'_2 \\ x'^2_2 \end{pmatrix}$$

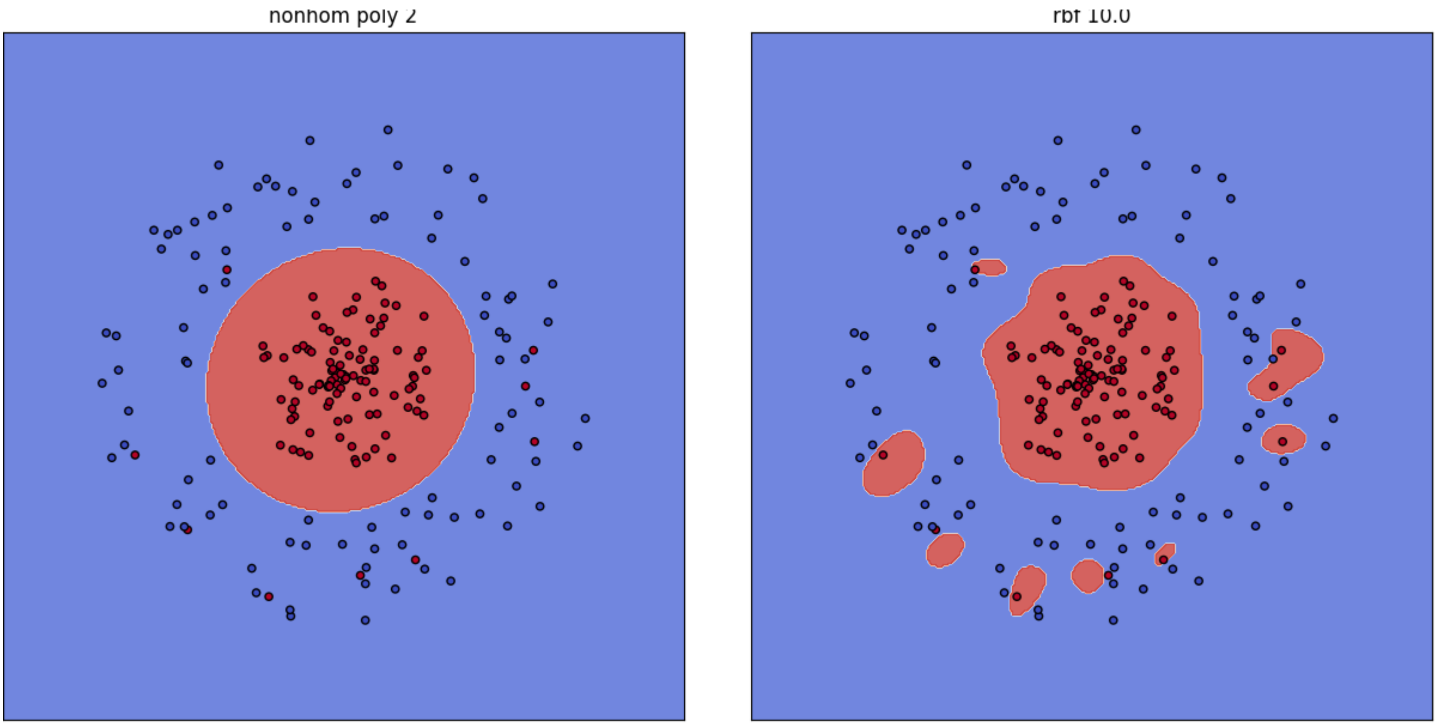
Because the data is labeled by a circle, the polynomial kernel with degree 2 mapping managed to include that "data" in the features, and because of that SVM managed to linearly separate our data in $\mathbb{R}^3$ using the new features.
This property does not hold with the linear kernel and homogenous polynomial kernel with degree 3.

b.



This graph was plotted using regularization parameter $C = 10$.
Now also the polynomial kernel with degree 3 managed to separate our data.
This because the feature space that the non-homogenous polynomial kernel with degree 3 our data is linearly separable.

c.



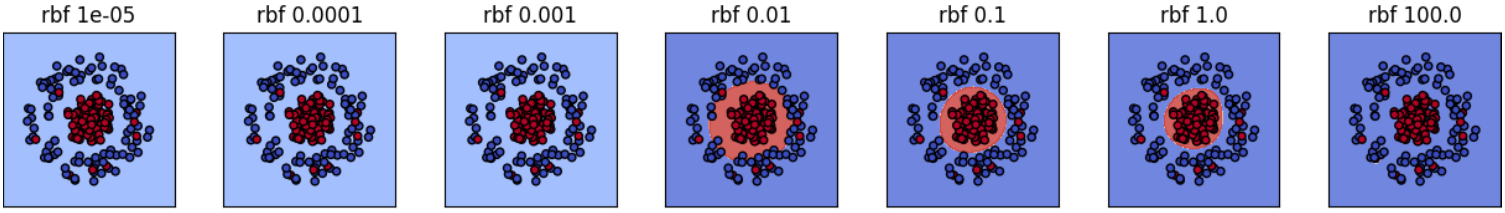
We can see that both kernels managed to fit out data correctly, with the extra noise we added.

The main difference is that the rbf classifier will classify new samples that close to the “noise”, with the label of the “noise”.

Because of that I think that the polynomial kernel is better than the rbf kernel with $\gamma = 10$,because when we want to predict a label of a new sample, the rbf classifier may label it falsely.

When testing different values of γ, we see that larger/smaller values of it result a classifier that overfitting our data.

Using $\gamma = 0.1, 1$ gave a similar result to the one of the non-homogeneous polynomial kernel with degree 2.


rbf 1e-05
rbf 0.0001
rbf 0.001
rbf 0.01
rbf 0.1
rbf 1.0
rbf 100.0