# Theoretical questions

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

Where our loss function is defined as:

Given sample, we notice that , because taking satisfies that , and therefore

.

We also notice that given s.t , our data is linearly separable.

Proof:

By contradiction, we assume that our data isn’t linearly separable.

This means that there is s.t .

We denote the argmax with , this means that

This means that , contradiction to our assumption.

Thus, our strategy is to find s.t , this implies that also for the minimizer is 0, thus the minimizer linearly separates our data.

We assume that there exist that linearly separate our data.

If , we’re finished, otherwise this means that there is s.t

Let’s mark the set of all that satisfies the above with .

Given , there is s.t

From reliability we know that .

If we define we get that

This means that

Let’s define .

Because of our assumption that

Let’s mark **.**

Given s.t and

And for , which implies that

This means that for any ,

Therefore, .

1. Let us solve SVM problem for two distinct s.t   
   We want to solve the following objective

Let consider the dual problem obtained in class:

In our case the objective we want to maximize is:

And our conditions turn to

If we plug this into the objective, we receive:

The maximum of this obtained when .

Because are two distinct points we divide by the norm of the differences.

Thus, we got that

From KKT we can obtain that

* For any support vector

In our case, we got 2 support vectors.

Using we get that

For conclusion, the results of SVM are:

* 1. Let consider the problem

With , and the optimal value for

And the problem

With , and the optimal value for

We prove that by showing that:

Proving a:

Because in and in we want to minimize the same objective, but has more constraints, it follows that .

Proving b:

Let there be that accomplish

We will define a solution to using them:

Let us show this is a solution to :

1. From the definition of
2. We notice that   
   therefore,

This means that is a solution to .

And we can see that

This means that .

We arrange our constraints to the form:

And we obtain that the Lagrangian is:

* 1. We compute the partial derivatives with respect to :

(If then and the dual problem is to maximize a const function with value )

If we plug this into the Lagrangian we get that

* 1. The dual problem is:

1. Let there be the solution for hard SVM and let there be the solution for soft SVM.

We assume that .

From the definition of , .

If we define we see that satisfies the soft SVM constraints.

Therefore,

And because we get that

To conclude, we got that

And this implies that .

And because of that,

And this means that separates the data.

1. Let there be distinct real numbers, and .

Given

This means that if we define for we get that .

If we examine , we see that

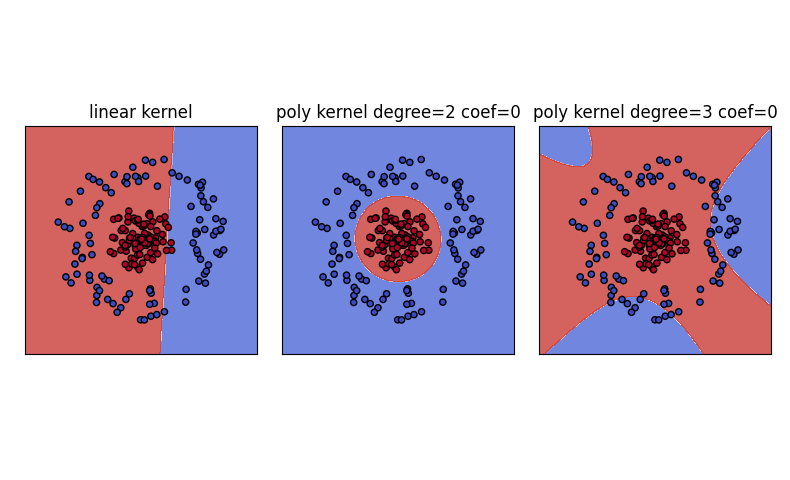
The matrix is Vandermonde matrix, and therefore its rank is .

We notice that is obtained from by multiplying the i-th column by .

And we remember that column operations don’t change the rank of the matrix, therefore .

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

# Programming questions



This graph was plotted using regularization parameter .

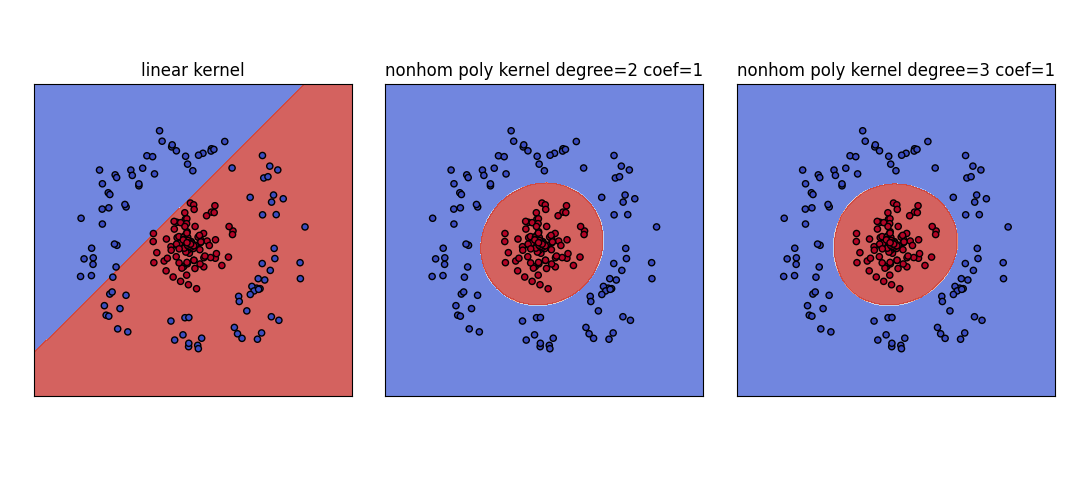
We can see that using homogenous polynomial kernel with degree 2 separated our data perfectly.

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 .

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 using the new features.

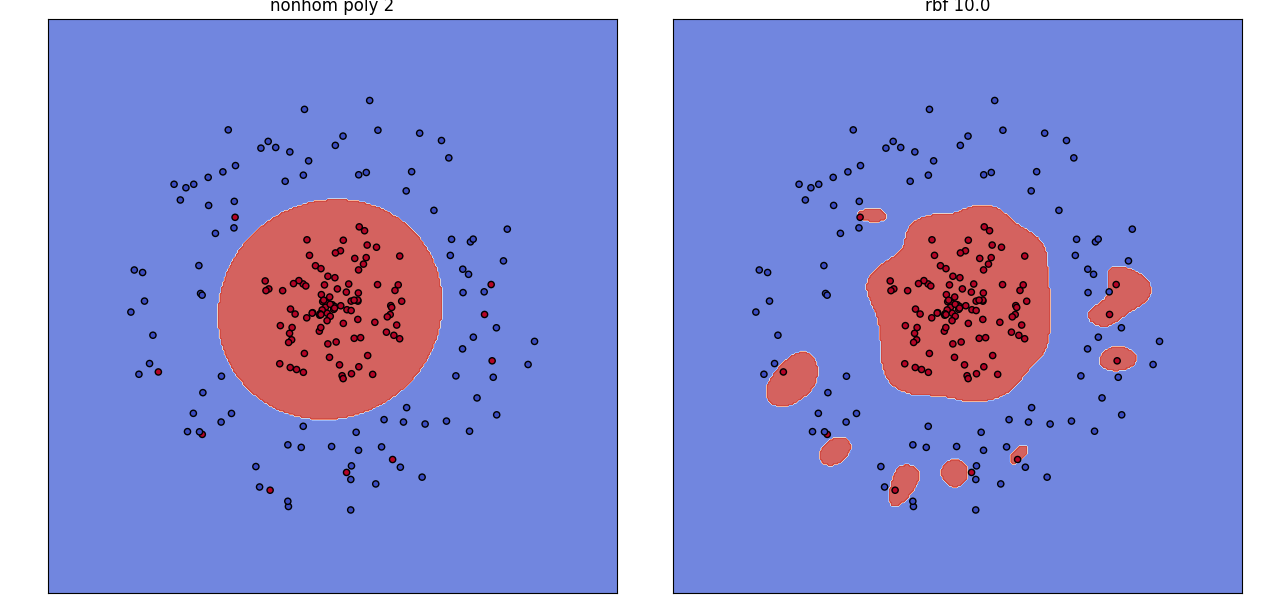
This property does not hold with the linear kernel and homogenous polynomial kernel with degree 3.



This graph was plotted using regularization parameter .

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 our data is linearly separable.

* 1. 

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 ,because when we want to predict a label of a new sample, the rbf classifier may label it falsely.

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

Using gave a similar result to the one of the non-homogeneous polynomial kernel with degree 2.

A picture containing qr code

Description automatically generated