CME 216, ME 343 - Winter 2020

Eric Darve, ICME

Overfitting and underfitting

The value of `C` can be optimized in different ways.

This is a broad topic and we will only cover the main ideas.

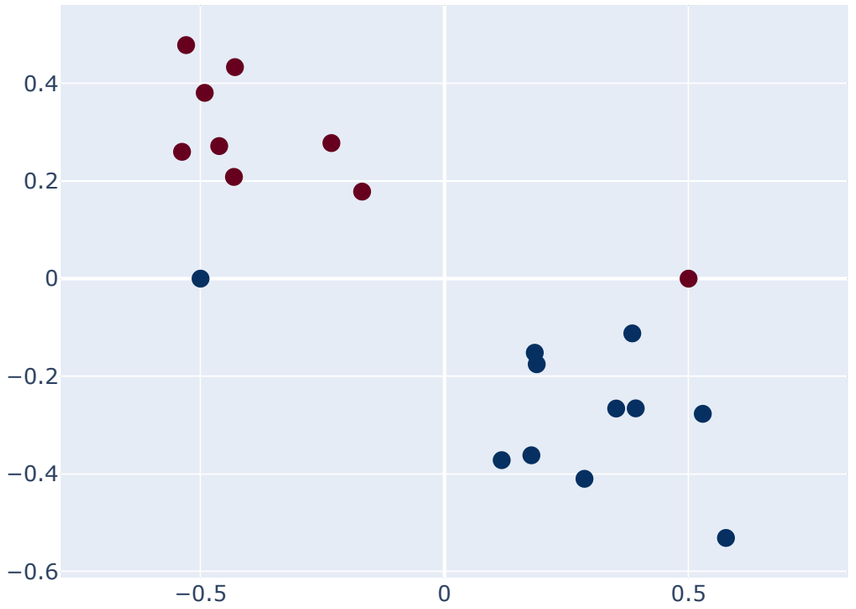
`C` must be tuned based on how we trust the data.

Generally speaking, if the data is very accurate (and a separating hyperplane exists) then `C` must be chosen very large.

But if the data is noisy (we do not trust it) then `C` must be small.

Let' s start by illustrating the effect of varying `C` in our method.

We consider the following problem.


0.4
0.2
0
−0.2
−0.4
−0.6
−0.5
0
0.5

We created two well-separated clusters with labels -1 and $+1$.

Then we added a blue point on the left and a red point on the right.

The real line of separation is $y = x$ as before.

So the outlier points can be considered as incorrect data here.

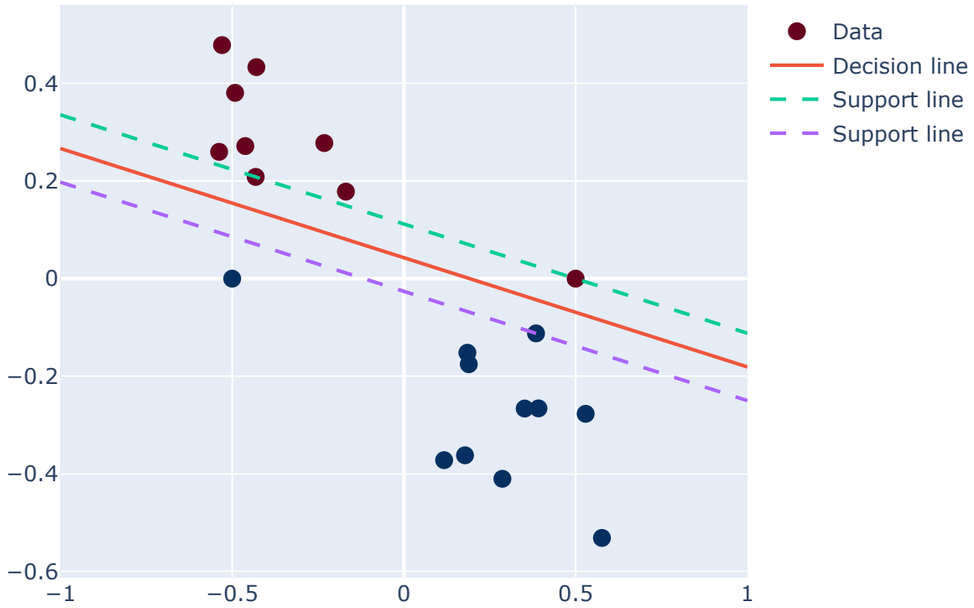
Either this data was entered incorrectly in our database, or there was some large error in the measurements.

Let's pick a large value of C

```
# fit the model
clf=svm.SVC(kernel="linear" , C=10^4)
clf.fit(X, y)
```

The SVM decision line has a negative slope as shown below.

C=10^4


Data
Decision line
Support line
Support line
0.4
0.2
0
−0.2
−0.4
−0.6
−1
−0.5
0
0.5
1

The red point on the right is classified with a label -1 (red-orange region).

And similarly for the blue point.

However, we know that these points are erroneous, and therefore the classification is wrong here.

This is a problem of *overfitting*.

We trust too much the data which leads to a large error.

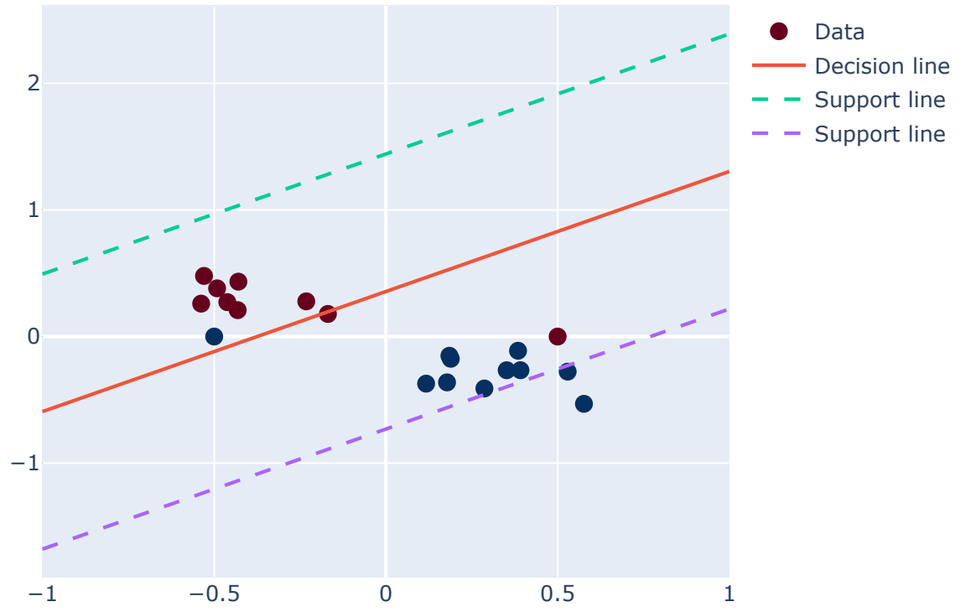
We can try again using a small `C`.

However, now the model believes that there is a large error in all the data.

As a result, the prediction is quite bad.

```
clf=svm.SVC(kernel="linear" , C=0.2)
clf.fit(X, y)
```

C=0.2



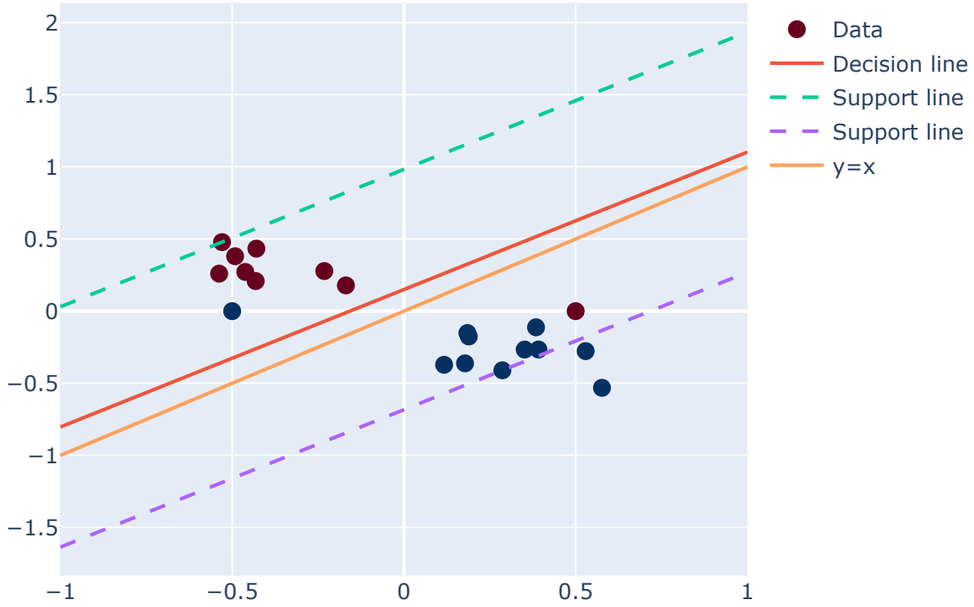
This case corresponds to a situation of *underfitting*.

That is we apply too much regularization by reducing `C` and do not trust enough the data.

If we pick `C=0.3`, we get a better fit in this case.

```
clf=svm.SVC(kernel="linear" , C=0.3)
clf.fit(X, y)
```

C=0.3


2
1.5
1
0.5
0
−0.5
−1
−1.5
−1
−0.5
0
0.5
1
Data
Decision line
Support line
Support line
y=x

This plot is intermediate between the previous plots.

We trust the outlier points but only to a moderate extent.

The solid orange line is the line $y = x$ but because of the outlier points, it is not possible in this case to recover that answer.

The SVC model is always biased by the outliers to some extent.