SVM

Sunday, March 31, 2024 9:18 PM

Support Vector Machines

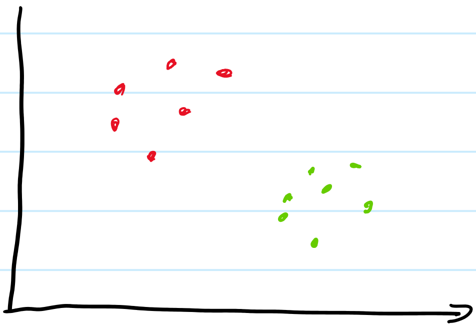
Easy Case:

Assume given is $\{(x_i, y_i)\}_{i \in \{1,\dots,N\}}$

$x_i \in \mathbb{R}^2$

$y_i \in \{\pm 1\}$ $\quad$ $+1$ is green

$\quad\quad\quad\quad$ -1 is red



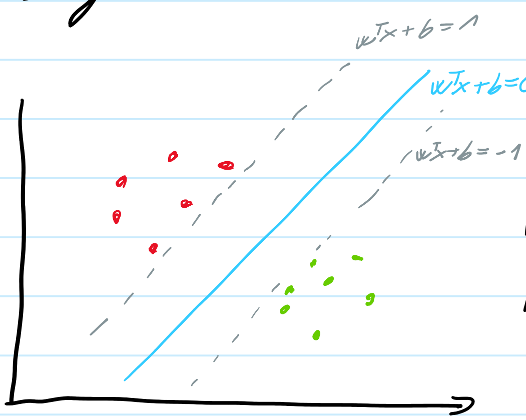
Assume it's seperated in such a simple way. (linearily separable)

We try to find $w \in \mathbb{R}^2, b \in \mathbb{R}$ that satisfy the following:

- $\|w\|$ as small as possible
- $w^T x_i + b \geq 1 \Rightarrow y_i = 1$
- $w^T x_i + b \leq -1 \Rightarrow y_i = -1$
- $\nexists$ no x_i : $w^T x_i + b \in (-1, 1)$

$\Bigg\} \Leftrightarrow y_i \cdot (w^T x_i + b) \geq 1$

Geometrically:



$\|w\|$ as small as possible

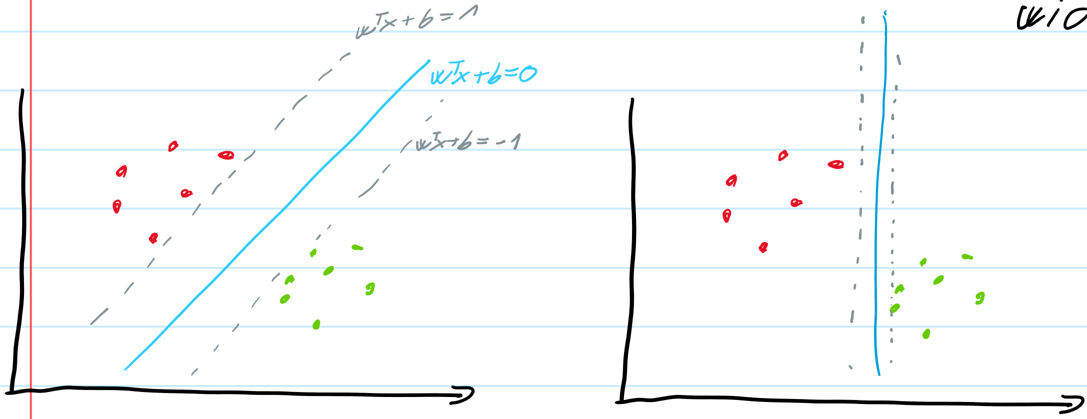
$\Leftrightarrow$

distance between grey lines is maximal

So we understand the intuition now! We want to find a seperating line, that is the best. There may be

So we understand the intuition now! We want to find a seperating line, that is the best. There may be uncountably many ones, but some better than other.
Ex:

Left is better than right, since the street is wider!



The Algorithm

We use loss function ($\lambda \in \mathbb{R}^+$ fixed)

$$L(w,b) = \lambda \|w\|^2 + \frac{1}{N} \sum_{i=1}^{N} \max(0, 1 - y_i(w^T x_i + b))$$

Notice: If $y_i(w^T x_i + b) \geq 1 \ \forall i \iff$ our line given by w, b splits the points

$\implies L(w,b) = \lambda \|w\|^2$

So the algorithm will try to make $\|w\|$ minimal, which will make the street wider, since the street has width $\frac{2}{\|w\|}$.

We use GRADIENT DESCENT!

$$w_{NEW} := w_{OLD} - \alpha \frac{\partial L}{\partial w}$$

$$b_{NEW} := b_{OLD} - \alpha \frac{\partial L}{\partial b}$$

If $y_i(w^T x_i + b) \geq 1$

$$\frac{\partial L}{\partial w_k} = 2\lambda w_k$$

$$\frac{\partial L}{\partial b} = 0$$

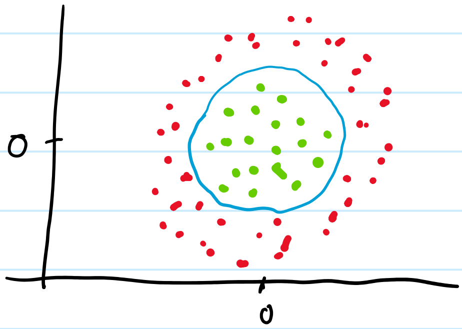
If $y_i(w^T x_i + b) < 1$

$$\frac{\partial L}{\partial w_k} = 2\lambda w_k - y_i (x_i)_k$$

$$\frac{\partial L}{\partial b} = -y_i$$

"Kernel Trick"

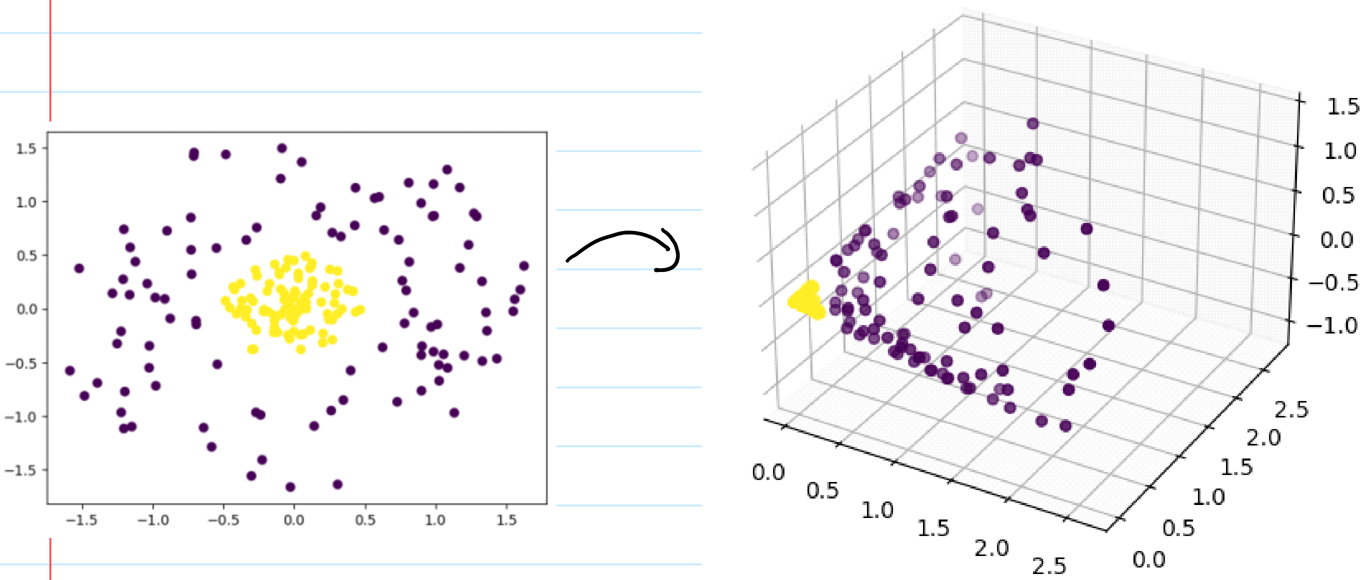
What if out data is not linearily separable? We need non-linear decision boundary. Ex:



We could use a transformation $\mathbb{R}^2 \xrightarrow{\varphi} \mathbb{R}^3$ given by

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \longmapsto \begin{pmatrix} x_1^2 \\ x_2^2 \\ x_1 x_2 \end{pmatrix}$$

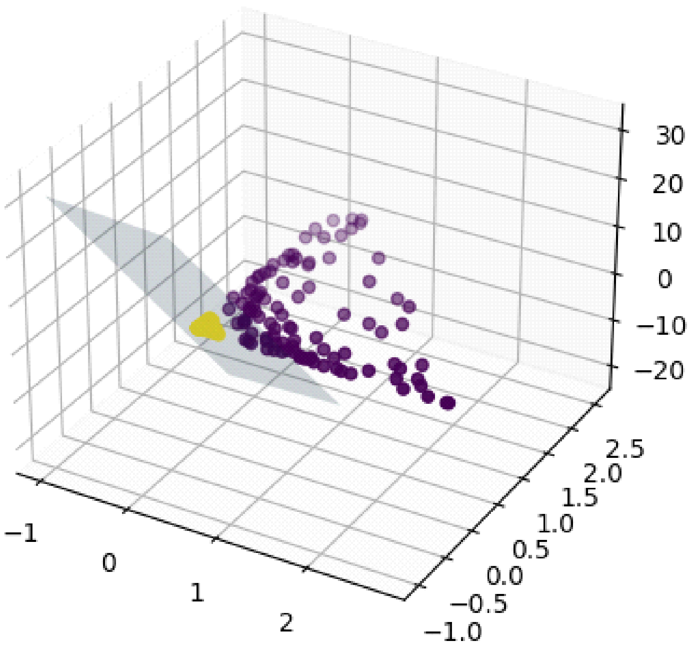
Then our data will look like



Now our data is linearily separable.
An interesting problem is finding a good φ.

Note! This is not actually the kernel trick, but is often confused with it.
By the kernel trick we usally mean a trick that saves computation cost for "good" φ ($\varphi(x)^T\varphi(x) = \varphi(x^Tx)$)

Applying SVM on our transformed data yields:


30
20
10
0
−10
−20
2.5
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1
0
1
2