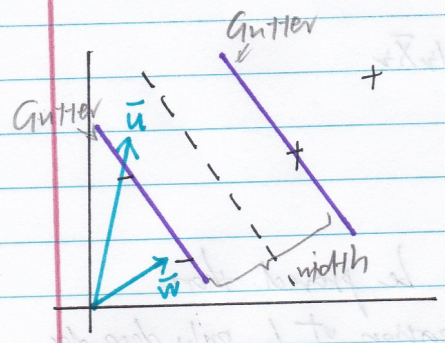
Support Vector Machines



$\bar{w}$ is a vector perpendicular to the middle line

$\bar{u}$ represent the unknown to be classified.

$\bar{w} \cdot \bar{u} \geq c \qquad c = -b$

① $\bar{w} \cdot \bar{u} + b \geq 0$ then the unknown is a "+"
this is our decision rule

Although we defined the decision rule, we yet have to find out what $\bar{w}$ and b is, the goal is to maximize the width

$\bar{w} \cdot \bar{x}_+ + b \geq 1$ ($\bar{x}_+$ represent a known "+" sample)

$\bar{w} \cdot \bar{x}_- + b \leq -1$

for our convenience, we introduce a new variable y_i such that $y_i = \begin{cases} 1 & \text{for "+" samples} \\ -1 & \text{for "-" samples} \end{cases}$

we multiply y_i to both sides of the two inequalities

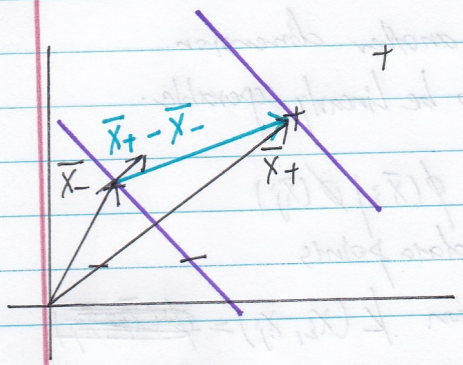
$\Rightarrow \quad y_i(\bar{w} \cdot \bar{x}_+ + b) \geq 1 \cdot 1$

$y_i(\bar{w} \cdot \bar{x}_- + b) \geq (-1)(-1)$ (since multiple negative values to both sides flips the sign)

the two inequalities above are actually the same.

$y_i(\bar{w} \cdot \bar{x}_i + b) - 1 \geq 0$

② $y_i(\bar{x}_i \cdot \bar{w} + b) - 1 = 0$ for x_i in gutter



$\text{width} = (\bar{x}_+ - \bar{x}_-) \cdot \frac{\bar{w}}{\|\bar{w}\|}$ this is a unit vector of the same direction of $\bar{w}$

$1-b \qquad 1+b$

③ $\Rightarrow \text{width} = (1-b+1+b) \cdot \frac{\bar{w}}{\|\bar{w}\|} = \frac{2}{\|\bar{w}\|}$

So to maximize $\frac{2}{\|\bar{w}\|}$, we want to minimize $\|\bar{w}\|$

for our convience, we will instead find the min of $\frac{1}{2}\|\bar{w}\|^2$

we use Larange multipliers here to find the minimum of a term with constraints (the constraint is $y_i(\vec{x}_i \cdot \vec{w} + b) - 1 = 0$

$$L = \frac{1}{2}\|\vec{w}\|^2 - \sum \alpha_i [y_i(\vec{w} \cdot \vec{x}_i + b) - 1]$$

$$\frac{\partial L}{\partial \vec{w}} = \vec{w} - \sum \alpha_i y_i x_i = 0 \Rightarrow \vec{w} = \sum \alpha_i y_i \vec{x}_i$$

$$\frac{\partial L}{\partial b} = -\sum \alpha_i y_i = 0 \Rightarrow \sum \alpha_i y_i = 0$$

plug in $\vec{w} = \sum \alpha_i y_i \vec{x}_i$ in L

~~$L = \frac{1}{2}(\sum \alpha_i y_i \vec{x}_i) \cdot (\sum \alpha_j$~~

we will eventually get $L = \sum \alpha_i - \frac{1}{2}\sum_i \sum_j \alpha_i \alpha_j y_i y_j (\vec{x}_i \cdot \vec{x}_j)$

It can be proved that the minization of L only depends on this dot product!!

plug $\vec{w} = \sum \alpha_i y_i \vec{x}_i$ in the decision rule

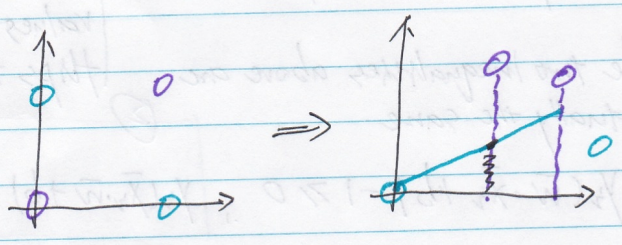
if $\sum \alpha_i y_i (\vec{x}_i \cdot \vec{u}) + b \geq 0$ then classify $\vec{u}$ as "+"

The decision rule depends on the dot product too

SVM only works for linear classification.

but if we transform the non-linear data from another perspective.

Eg:



we can transform a data point x_i to another dimension using function $\phi(\vec{x})$ for the data to be linearly seperable.

then the decision rule will dependent on $\phi(\vec{x}_i) \phi(\vec{x}_j)$ the dot product of two transformed data points.

to simplify, we can have a kernal function $K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j)$

There's two popular choice of the kernal function

① $(\vec{u} \cdot \vec{v} + 1)^n$ ② $e^{-\frac{\|x_i - x_j\|}{\sigma}}$