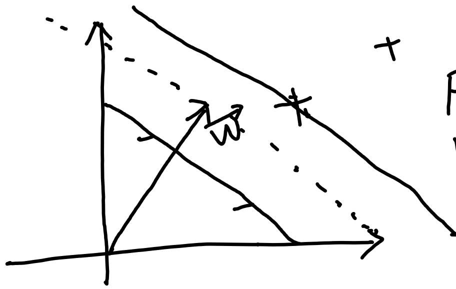
Support Vector Machines



Data points $\vec{x_i}$

Find a "gutter" of maximum width that separates the negative from the positive samples

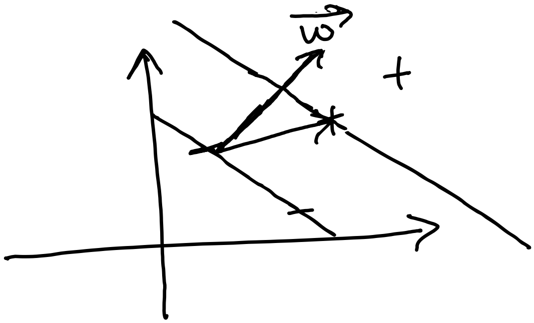
$\vec{w}$ is the normal vector to the street
When we "project" a sample onto $\vec{w}$, the length of the projection determines whether the sample is positive or negative

$$(\vec{w} \cdot \vec{x_+} + b) \geq 1 \quad \text{for positive samples (1)}$$
$$(\vec{w} \cdot \vec{x_-} + b) \leq -1 \quad \text{for negative samples (2)}$$

Invent variables $y_i = \begin{cases} 1 & \text{for positive samples} \\ -1 & \text{for negative samples} \end{cases}$

So (1) → $y_i(\vec{w} \cdot \vec{x_i} + b) \geq 1.$
(2) ↗ $y_i(\vec{w} \cdot \vec{x_i} + b) - 1 = 0$ for samples in the gutter

Decision rule: $(\vec{w} \cdot \vec{x} + b) \geq 0$ then +
else −



The width of the gutter is

$$\frac{\vec{w}}{\|w\|} \cdot (\vec{x}_+ - \vec{x}_-)$$

$$= \frac{(1-b-(-1-b))}{\|w\|} = \frac{2}{\|w\|}$$

To maximize $\frac{2}{\|w\|}$, we minimize $\|w\|$

or minimize $\frac{1}{2}\|w\|^2$

So the problem becomes:

Minimizing $\frac{1}{2}\|w\|^2$

Condition: $y_i(\vec{w} \cdot \vec{x_i} + b) - 1 \geq 0$

Use Lagrange multiplier:

Minimize:

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

$$\frac{\partial L}{\partial \vec{w}} = \vec{w} - \sum \alpha_i y_i \vec{x_i} = 0 \quad (4)$$

$$\frac{\partial L}{\partial \alpha_i} = y_i(\vec{w} \cdot \vec{x_i}) + b - 1 = 0 \quad (5)$$

$$\frac{\partial L}{\partial b} = -\sum \alpha_i y_i = 0 \quad (6)$$

Plug (4) into L:

$$L = \frac{1}{2}\left(\sum \alpha_i y_i \vec{x_i}\right) \cdot \left(\sum \alpha_i y_i \vec{x_i}\right) -$$

$$\sum(\alpha_i y_i \vec{x_i}) \cdot \left(\sum \alpha_i y_i \vec{x_i}\right) - \sum \alpha_i y_i b$$

$$+ \sum \alpha_i$$

$$= -\frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j \vec{x_i} \cdot \vec{x_j} + \sum \alpha_i \quad (7)$$

Plug $\vec{w}$ into the decision rule:

$$\sum \alpha_i y_i \vec{x_i} \cdot \vec{x} + b \geq 0 \text{ then } + \quad (8)$$

In (7), the α_i are the unknowns
(7) is a quadratic programmig problem and can be solve by standard quadratic programmig packages
It is also convex and can be solve with gradient descent

Look at (8):

$$\sum \alpha_i y_i \underbrace{\vec{x}_i \cdot \vec{x}} + b$$

(cosine distance)

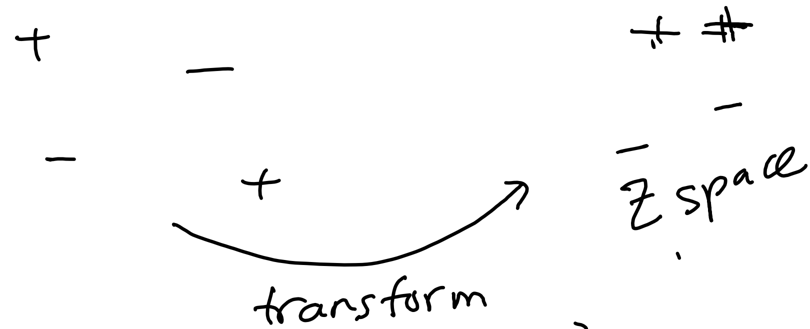
weighted

how "close" $\vec{x}$ is to $\vec{x_i}$

to decide whether $\vec{x}$ is + or −, we take a linear combination of the distances of $\vec{x}$ to all the $\vec{x}_i$'s.

Ultimately, only some of the α_i's will be non-zero, the rest are 0. (contribute nothing to the decision)
The $\vec{x}_i$'s with non zero α_i's are called the support vectors

In the linearly inseparable cases:



Call the transform ϕ: take $\vec{x}_i$'s to z space

Since in (7):

$$L = -\frac{1}{2} \sum \sum \alpha_i \alpha_j y_i y_j \vec{x}_i \cdot \vec{x}_j + \sum \alpha_i$$

L only depends on inner products, so we only need to define a function K

$$K(\vec{x}_i, \vec{x}_j) = \phi(\vec{x}_i) \cdot \phi(\vec{x}_j)$$

dot product in z space

Common K: $(\vec{x}_i \cdot \vec{x}_j + 1)^n$

<u>kernels</u> $\exp\left(\frac{\| x_i - x_j \|^2}{\lambda}\right)$

Example: in 2D

$$K(\vec{x},\vec{x}') = (\vec{x}\cdot\vec{x}'+1)^2 = (1+x_1x_1'+x_2x_2')^2$$

$$= 1+x_1^2x_1'^2+x_2^2x_2'^2+2x_1x_1'+2x_2x_2'+2x_1x_1'x_2x_2'$$

this is an inner product;

$$(1,\ x_1^2,\ x_2^2,\ \sqrt{2}\,x_1,\ \sqrt{2}\,x_2,\ \sqrt{2}\,x_1x_2)$$
$$(1,\ x_1'^2,\ x_2'^2,\ \sqrt{2}\,x_1',\ \sqrt{2}\,x_2',\ \sqrt{2}\,x_1'x_2')$$

additional dimensions

similar to the case of linear regression where a line is insufficient to model the data

inner product

needs higher order terms, or "features"

$x^2, x^3, x^4, \ldots$

Example: $K(\vec{x}, \vec{x}') = \exp(-\|\vec{x} - \vec{x}'\|^2)$

in the case of the dot product kernel:

$$K(\vec{x}, \vec{x}') = \vec{x} \cdot \vec{x}'$$

Consider the decision rule:

$$\sum \alpha_i y_i K(\vec{x}_i, \vec{x}) + b \geq 0$$

From the point of view of generating a field, consider α_i, y_i, b, and the $\vec{x}_i$'s fixed.

then, if $K(\vec{x}_i, \vec{x}) = \vec{x}_i \cdot \vec{x}$, this generates a linear field.

If $K(\vec{x}_i, \vec{x}) = (\vec{x}_i \cdot \vec{x} + 1)^2$, this generates a quadratic field.

Is $K(x, x') = \exp(-(x - x')^2)$
an inner product in some space?

$$K(x, x') = \exp(-(x - x')^2)$$

$$= \exp(-x^2)\exp(-x'^2)\exp(2xx')$$

$$= \exp(-x^2)\exp(-x'^2)\sum_{k=0}^{\infty} \frac{2^k (x)^k (x)^k}{k!}$$

Taylor series
Infinite dimensional space

Valid kernels:

- Is symmetric
  i.e. $K(\vec{x}, \vec{x}') = K(\vec{x}', \vec{x})$
  since the inner product is symmetric
- Matrix

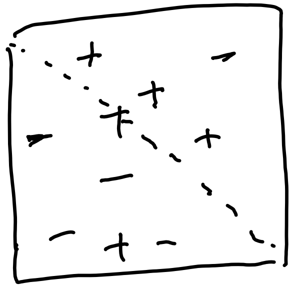
$$M = \begin{bmatrix} K(x_1, x_1) & K(x_1, x_2) & \cdots & K(x_1, x_N) \\ K(x_2, x_1) & K(x_2, x_2) & \cdots & K(x_2, x_N) \\ \cdots & \cdots & & \cdots \\ K(x_N, x_1) & K(x_N, x_2) & \cdots & K(x_N, x_N) \end{bmatrix}$$

is positive semi-definite

(≥ 0 in matrix lingo)

$x^T M x \geq 0 \ \forall x$

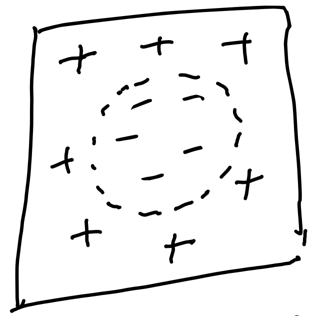
Soft SVM



some outliars

↓

don't want kernel to ovefit data

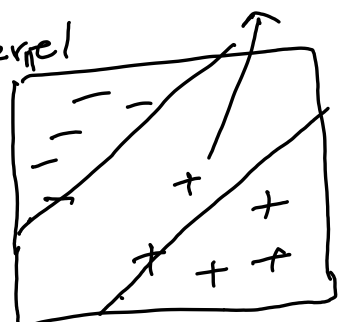


linearly non-separable

↓

needs kernel

violation



Allow some error:

$$y_i(\vec{w} \cdot x_i + b) \geq 1 - \varepsilon_i \quad (\varepsilon_i \geq 0)$$

before: $y_i(\vec{w} \cdot x_i + b) \geq 1$

We want to minimize the total violation: $\sum_{i=1}^{m} \varepsilon_i$

New optimization:

$$\frac{1}{2}\|\vec{w}\|^2 + C\sum \varepsilon_i \quad \text{subject to} \quad y_i(\vec{w} \cdot \vec{x}_i + b) \geq 1 - \varepsilon_i, \quad \varepsilon_i \geq 0$$

Lagrange multiplier:

$$L = \frac{1}{2}\|\vec{w}\|^2 + C\sum \varepsilon_i - \sum \alpha_i \left(y_i(\vec{w}\cdot\vec{x}_i) + b - 1 + \varepsilon_i\right) - \sum \beta_i \varepsilon_i$$

Minimize with respect to $\vec{w}$, b, ε_i

and maximize with respect to α_i and β_i

$$\frac{\partial L}{\partial w} = \vec{w} - \sum \alpha_i y_i x_i = 0$$ like before

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

$$\frac{\partial L}{\partial \varepsilon_i} = C - \alpha_i - \beta_i = 0$$

different from before

since $\beta_i \geq 0$, $\alpha_i \leq C$.

So, maximize

$$L(\vec{\alpha}) = \sum \alpha_i - \frac{1}{2}\sum\sum y_i y_j \alpha_i \alpha_j \vec{x}_i \cdot \vec{x}_j$$

subject to $0 \leq \alpha_i \leq C$