SUPPORT VECTOR MACHINES

(Webb & Copsey 2011) pp. 249 –

Can we do better than the separating hyperplane
<SLIDE> // which is better H_1, H_2 or H_3

Lets use the linear classifier again:

$$w_1 x_1 + w_2 x_2 + b = \begin{cases} > 0 \Rightarrow \text{class } \omega_1 : \text{output } y_i = +1 \\ < 0 \Rightarrow \text{class } \omega_2 : \text{output } y_i = -1 \end{cases}$$

omega

$b = w_0$ // for notational consistency

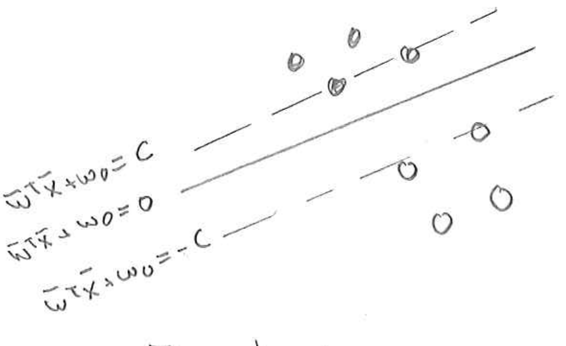
Lets use the matrix forms

$$\bar{w} = \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_N \end{pmatrix} \quad \bar{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{pmatrix}$$

Now, for all correctly classified points:

$$y_i(\bar{w}^T \bar{x} + w_0) > 0$$

We wish to maximise the margin



The larger we can push C the better is the margin.
Distance of two parallel lines (wikipedia) $ax + bx + c_1 = 0$ and $ax + bx + c_2 = 0$:

$$\text{dist}(l_1, l_2) = \frac{|c_2 - c_1|}{\sqrt{a^2 + b^2}}$$

Remember that as lines $ax+bx+c=0$ and e.g. $2ax+2bx+2c=0$ are equivalent and we may thus fix $c=1$

$$\Rightarrow dist(l_1, l_2) = \frac{|1-(-1)|}{\sqrt{w_1^2+w_2^2}} = \underline{\underline{\frac{2}{\sqrt{\bar{w}^T \bar{w}}}}}$$

to maximise that we need to minimise $\sqrt{\bar{w}^T\bar{w}}$ which is that we minimise $\bar{w}^T\bar{w}$

Our maximum margin problem is

$$\min \quad \bar{w}^T \bar{w} \qquad // \underset{\text{(quadratic)}}{\text{2. asteen}} \text{ polynomi}$$

subject to

$$\left.\begin{aligned} y_1(\bar{w}^T\bar{x}_1 + w_0) &\geq 1 \\ y_2(\bar{w}^T\bar{x}_2 + w_0) &\geq 1 \\ &\vdots \\ y_N(\bar{w}^T\bar{x}_2 + w_0) &\geq 1 \end{aligned}\right\} \text{// lineaariset rajoitteet}$$

The quadratic function is convex and therefore our problem is a <u>convex optimization problem with linear inequalities</u> (check wikipedia).

OPTIMISATION: <u>Next note</u>

Back: we run quadratic programming solver to above and those restrictions that turn to equalities ($y_i(\bar{w}^T\bar{x}_i + w_0) = 1$) are called the support vectors (active restrictions)

OPTIMIZATION

1. Linear programming

$\max \quad w_1 x_1 + w_2 x_2$

subject to $x_1 + x_2 \leq B$

$x_1 \geq 0, x_2 \geq 0$

e.g. max total price of selling gold (price 100 w_1) and silver (10 w_2) if you can carry max B kg.

=> e.g. the Simplex method

2. Unconstrained problems

$\max / \min \quad f(x)$

e.g. minimise the fitting error (MSE) of a linear function or polynomial

=> gradient descent

3. Constrained problems

e.g. SVM learning

$\min \quad f(x)$

subj. to $h_i(x) = 0, \; i = 1, 2, \ldots, m$

$g_j(x) \leq 0, \; j = 1, 2, \ldots, n$

=> Methods combining ideas from 1. and 2.

4. Discrete optimization
   * combinatorial (graphs) (e.g. the traveling salesman problem)
   * integer programming - many ways equivalent to combinatorial

=> Often NP-hard, but for many special cases fast and effective approximation algorithms exist (ideologies: beam search, branch-and-bound, etc.)

Effective scientific approach: formulate your "learning" as a function to minimise or maximise - see a proper template from above and run existing solvers (e.g. quadprog() in Matlab)

SVM with linearly non-separable data

we introduce "slack variables" ξ_i <SLIDE>

$$y_i(\bar{w}^T\bar{x}_i + w_0) \geq 1 - \xi_i \quad i=1,\ldots,n$$

$$\xi_i \geq 0$$

This allow some points to be on the wrong side of the margin and in the minimisation we give penalty on that

$$\min \; \bar{w}^T\bar{w} + C\sum_i \xi_i$$

$\Rightarrow$ Run quadratic prog. solver for the new cost function!

Nonlinear SVM