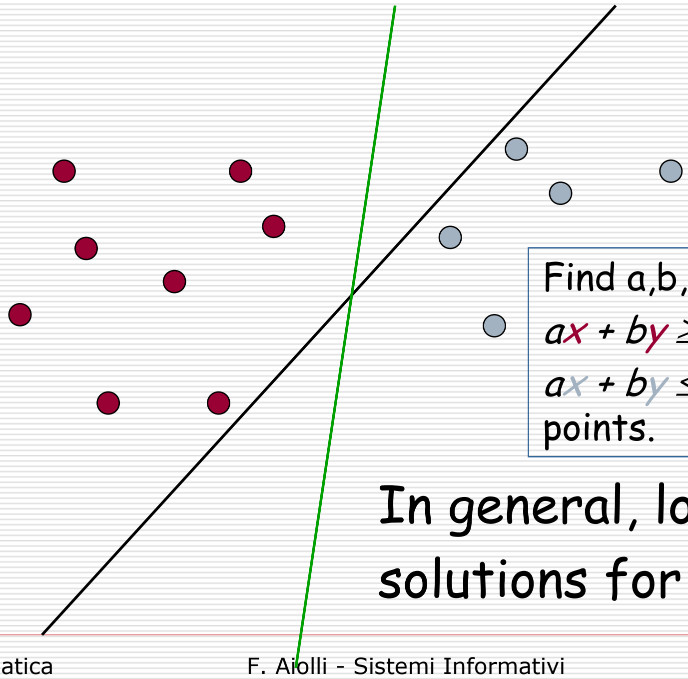
Which Hyperplane?



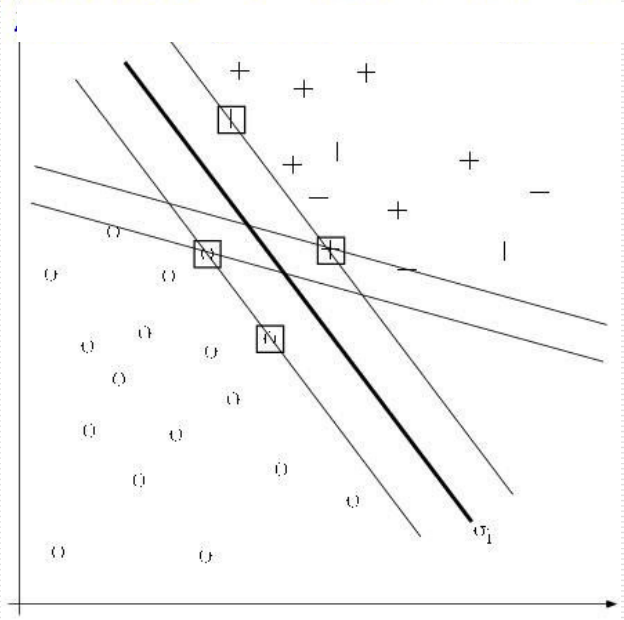
Find a,b,c, such that
$ax + by \geq c$ for red points
$ax + by \leq c$ for green points.

In general, lots of possible solutions for *a,b,c.*

The Support Vector Machine

- The support vector machine (SVM) method attempts to find, among all the decision surfaces $h_1, h_2, .., h_n$ in d-dimensional space, the one h_{svm} that does it by the widest possible margin
- This method applies the so called structural risk minimization principle, in contrast to the empirical minimization principle
- Learning a SVM is typically a quadratic problem

SVM



The Support Vector Machine

- The maximal margin hyperplane is also called optimal hyperplane
- Why it should be the best?
  - Keep training data far away from the classifier (fairly certain class. decisions)
  - The capacity of the model decreases as the separator become fatter. N.B. the bias has been fixed as we are looking for a linear separation in feature space

The Support Vector Machine

- The (functional) margin of data points is often used as a measure of confidence in the prediction of a classifier,

  $\rho_i = y_i (w x_i + b)$
- The geometric margin ρ of a classifier is the Euclidean distance between the hyperplane and the closest point
- It can be shown that $\rho = 2/||w||$

The SVM problem

- Find an hyperplane, consistent with the labels of the points, that maximizes the geometric margin, i.e.

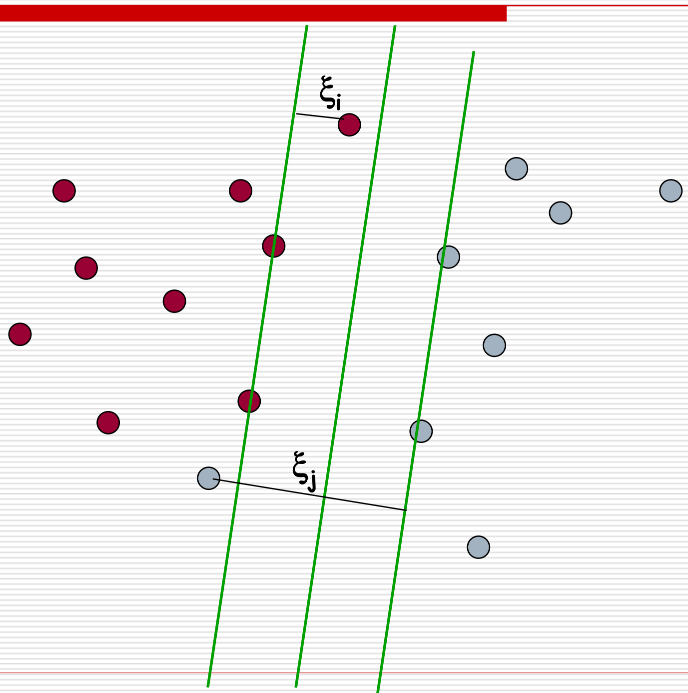
$$\text{Min } 1\backslash 2\ ||w||^2$$
$$\text{And for all } \{(x_i,y_i),\ y_i(w\ x_i + b) \geq 1\}$$

- This is a (convex) constrained quadratic problem. Thus it guarantees a unique solution!
- Many QP algorithms exists to find the solution of this quadratic problem
- In SVM related literature, many algorithms have been devised ad hoc for this kind of quadratic problem (svmlight, bsvm, SMO, …)

Solving the optimization problem

- Typically, solving an SVM boils down into solving the dual problem where Lagrange multipliers $\alpha_i \geq 0$ are associated with every constraint in the primary problem
- The solution turns out to be in the form
  - $w = \sum_i y_i \alpha_i x_i$
  - $b = y_k - \langle w, x_k \rangle$ for any x_k s.t. $\alpha_k > 0$
- In the solution most of the α_i are zeros. Examples associated with non zero multipliers are called support vectors
- $h_{SVM}(x) = \text{sign}(w\, x + b) = \text{sign}(\sum_i y_i\, \alpha_i \langle x_i, x \rangle + b)$

Non-separable Datasets



Soft margin SVM

- Find an hyperplane, consistent with the labels of the points, that maximizes the function

$$\text{Min } 1\backslash 2 \ ||w||^2 + C \sum_i \xi_i$$
$$\text{And for all } \{(x_i,y_i),\ y_i(w\ x_i + b) \geq 1-\xi_i,\ \xi_i \geq 0\}$$

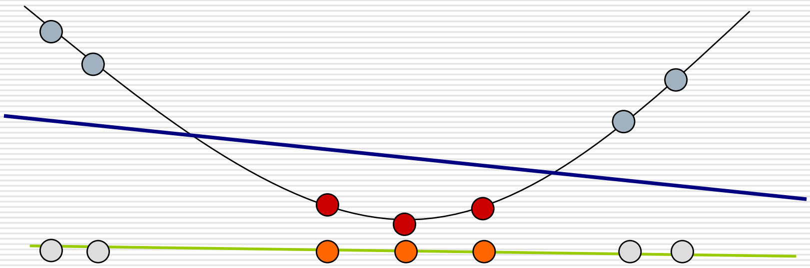
- The parameter C can be seen as a way to control overfitting.
- As C becomes larger it is unactractive to not respect the data at the cost of reducing the geometric margin.
- When C is small, larger margin is possible at the cost of increasing errors in training data
- Interestingly, the SVM solution is in the same form as in the hard margin case!

Non-separable Datasets

How can we separate these data?

Non-separable Datasets

Projecting them into a higher dimensional space



$$\Phi : \mathbf{x} \rightarrow \phi(\mathbf{x})$$

Solving the optimization problem

- $h_{SVM}(x) = \text{sign}(w\ \phi(x) + b)$
  $= \text{sign}(\sum_i y_i\ \alpha_i \langle\phi(x_i), \phi(x)\rangle + b)$
  $= \text{sign}(\sum_i y_i\ \alpha_i\ K(x_i,x) + b)$
- Where $K(x_i,x)$ is the kernel function such that $K(x_i,x) = \langle\phi(x_i), \phi(x)\rangle$

Advantages of SVM

- SVMs have important advantages for TC
  - The 'best' decision surface is determined by only a small set of training examples, called the support vector (in the linear case, everything can be compacted in one vector only)
  - Different kernel functions can be plugged in, corresponding to different ways of computing the similarity of document
  - The method is applicable also to the case in which the sample is not separable
  - No term selection is usually needed, as SVMs are fairly robust to overfitting and can scale up to high dimensionalities
- SVM has been shown among the top performing systems in a number of experiments [Dumais+98,Joachims98,Yang&Liu99]