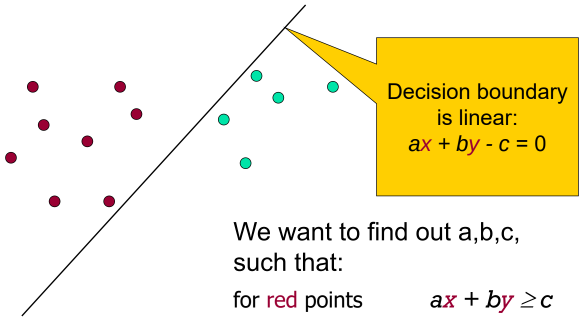
SVM (1): Support vector machine

Akito Sakurai

Linear discriminant function

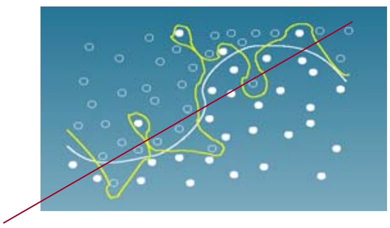


We want to find out a,b,c, such that:

for red points $ax + by \geq c$

for green pts. $ax + by \leq c$.

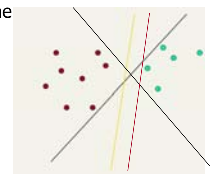
Complex boundaries



From Christopher Manning's slides

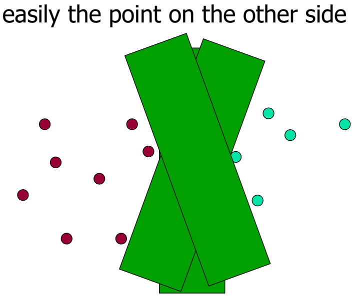
Which hyperplane is to be chosen

- *a,b,c* have infinite possibilities.
- Any one of which is the best [we have to define a standard to measure goodness]
  - Consider the measure for the perceptron learning algorithm if you know it
- SVM finds the "best" one.
  - Hyperplane that maximizes distance to the nearest "difficult point.
  - Intuitive interpretation: the further the points of the other classes are to the decision boundary, the less the uncertainty of decision is.



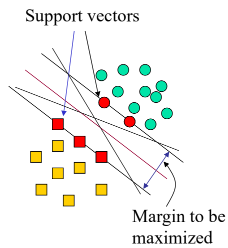
Another intuitive interpretation

- Replace a decision boundary by a strip with non-zero width. The narrower the width is, the more easily the point on the other side could jump in



Support vector machine (SVM)

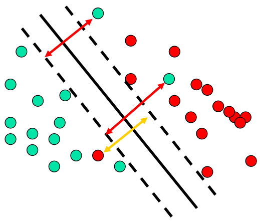
- SVM maximizes the margin around the separating hyperplane.
  - called "large margin classifier"
- Decision function is determined by its support vector which are in the training dataset.
- Formulated as quadratic programming
- Considered to work well for wide variety of problems



Large margin classifiers

If the dataset is not linearly separable,

- Allow errors, but
  - Have to pay penalty for the distance to the nearest allowable position
- While keeping the margin large



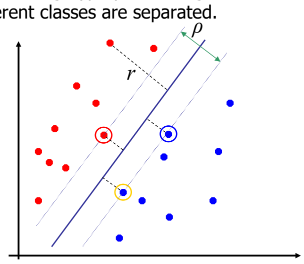
Margin: formulation

- w: normal vector to the decision boundarn
- x_i: i-th sample
- y_i: class to belong (+1 or -1) Note: not 1/0
- classifier: $\text{sign}(w^T x_i + b)$
- Functional Margin of x_i : $y_i (w^T x_i + b)$
  - Clearly when w gets longer, margin gets larger

(Functional margin of a dataset is the maximum of them)

Geometrical margin

- Distance from a sample to the hyperplane $r = \frac{\mathbf{w}^T \mathbf{x} + b}{\|\mathbf{w}\|}$
- Samples nearest to the hyperplane are support vectors.
- Margin ρ of separating hyperplane designates how far the support vectors of different classes are separated.



Linear SVM methematics

- Suppose that all the points are positioned further than hyperplane by function value 1. Then, the following two constraints are obtained from the training dataset $\{(\mathbf{x_i}, y_i)\}$:

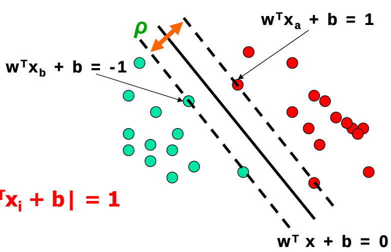
$$\mathbf{w}^\mathbf{T}\mathbf{x_i} + b \geq 1 \quad \text{if } y_i = 1$$
$$\mathbf{w}^\mathbf{T}\mathbf{x_i} + b \leq -1 \quad \text{if } y_i = -1$$

- For the support vectors, the above inequalities become equalities; Then, the margin is $\rho = 2/\|\mathbf{w}\|$ because the distance from each sample to the hyperplane is $r = \frac{\mathbf{w}^T \mathbf{x} + b}{\|\mathbf{w}\|}$

Assumption: The function (representing hyperplane) takes 1 and -1 on the marginal hyperplane

Linear SVM

- **Hyperplane**
  $w^T x + b = 0$
- **Constraints**:
  $\min_{i=1,\ldots,n} |w^T x_i + b| = 1$
- Rewritten to:
  $w^T(x_a - x_b) = 2$
  $\rho = \|x_a - x_b\|_2 = 2/\|w\|_2$



Linear SVM

- Formulated as the following quadratic programming:

Find **w** and *b* such that:
$\rho = \frac{2}{\|\mathbf{w}\|}$ is the maximal, and for all $\{(\mathbf{x_i}, y_i)\}$
$\mathbf{w}^\mathbf{T}\mathbf{x_i} + b \geq 1$ if $y_i = 1$; $\mathbf{w}^\mathbf{T}\mathbf{x_i} + b \leq -1$ if $y_i = -1$

- A better formulation (min $\|\mathbf{w}\|$ = max $1/\|\mathbf{w}\|$):

Find **w** and *b* such that:
$\mathbf{\Phi}(\mathbf{w}) = ½\, \mathbf{w}^\mathbf{T}\mathbf{w}$ is the minimal, and for all $\{(\mathbf{x_i}, y_i)\}$
$y_i (\mathbf{w}^\mathbf{T}\mathbf{x_i} + b) \geq 1$