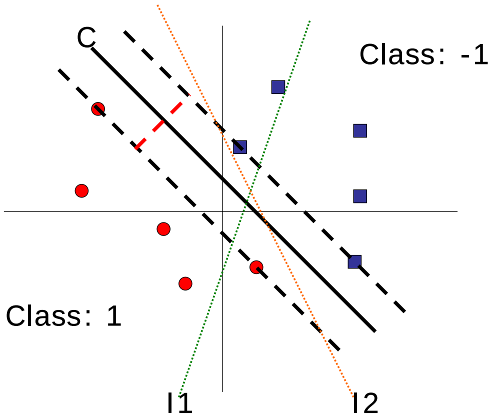
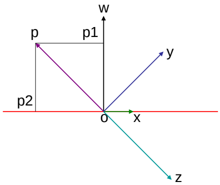
Support vector machines

October 16, 2009

Idea: Given a training sample, the support vector machine constructs a hyperplane as the decision surface in such a way that the margin of separation between a positive and negative examples is maximized.



Basic geometry:



Equation of the hyperplane

$$g(\mathbf{x}) = \mathbf{w}^\top(\mathbf{x} - \mathbf{x}_0) = \mathbf{w}^\top\mathbf{x} + b = 0$$

Sides of the hyperplane

$$\mathbf{w}^\top\mathbf{y} + b > 0 \text{ and } \mathbf{w}\mathbf{z} + b < 0$$

Projection on the hyperplane

$$\mathbf{p} = \mathbf{p}_1 + \mathbf{p}_2 \text{ where } \mathbf{p}_1 = r\frac{\mathbf{w}}{||\mathbf{w}||}$$

Classification Given samples and corresponding class labels i.e. $\{\mathbf{x}_i, d_i\}_{i=1}^n$

$$d_i = 1 \text{ if } \mathbf{w}^\top \mathbf{x}_i + b \geq 1$$
$$d_i = -1 \text{ if } \mathbf{w}^\top \mathbf{x}_i + b \leq -1$$

i.e.

$$d_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1$$

Margin

$$\frac{2}{||\mathbf{w}||}$$

Problem

$$\min_{\mathbf{w}, b} \frac{1}{2}\mathbf{w}^\top \mathbf{w} \text{ such that } d_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 \, \forall \, i$$

More math!!!

Use Lagrange multipliers

$$\max_{(\alpha_1,\ldots,\alpha_n)} \min_{\mathbf{w},b} \frac{1}{2}\mathbf{w}^\top \mathbf{w} - \sum_{i=1}^{n} \alpha_i [d_i(\mathbf{w}^\top \mathbf{x}_i + b) - 1] \text{ such that } \alpha_i \geq 0$$

- Derivative with respect to **w** is zero i.e.

$$\frac{\partial J}{\partial \mathbf{w}} = 0 \Rightarrow \mathbf{w} = \sum_{i=1}^{n} \alpha_i d_i \mathbf{x}_i$$

  w is in the span of the samples

- Derivative with respect to b is zero i.e.

$$\frac{\partial J}{\partial b} = 0 \Rightarrow \sum_{i=1}^{n} \alpha_i d_i = 0$$

KKT condition

$$\alpha_i[d_i(\mathbf{w}^\top \mathbf{x}_i + b) - 1] = 0$$

Only α_i's with $d_i(\mathbf{w}^\top \mathbf{x}_i + b) = 1$ can take nonzero values

Sparsity! Support vectors!

Dual problem

$$\max_{(\alpha_i,\ldots,\alpha_n)} \sum_{i=1}^n \alpha_i - \frac{1}{2}\sum_{i=1}^n \sum_{j=1}^n \alpha_i\alpha_j d_i d_j \mathbf{x}_i^\top \mathbf{x}_j \text{ such that } \sum_{i=1}^n \alpha_i d_i = 0, \alpha_i \geq 0$$

Quadratic optimization problem

Nonlinear decision surface Use similar ideas as in RBF.

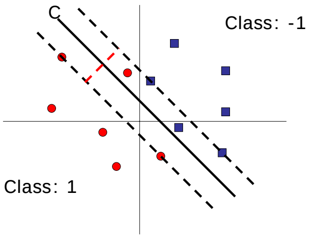
$$\mathbf{w} = \sum_{i=1}^{n} \alpha_i d_i \varphi(\mathbf{x})$$

But wait! Note that, α depends on $\mathbf{x}$ through the inner product $< \mathbf{x}|\mathbf{y} >_1 = \mathbf{x}^\top \mathbf{y}$.

Specify the inner product without specifying the nonlinear functions explicitly. For example,

$$< \varphi(\mathbf{x})|\varphi(\mathbf{y}) >_2 = (\mathbf{x}^\top \mathbf{y})^2$$

For this example if $\mathbf{x} = [x_1, x_2]$ then $\varphi(\mathbf{x}) = [x_1^2, x_2^2, \sqrt{2}x_1 x_2]$



Slack variable

$$d_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \xi_i, \xi_i \geq 0$$

$0 < \xi_i \leq 1$:Correct classification but inside margin

$\xi_i > 1$:On the wrong side!

Primal problem

$$\min_{\mathbf{w}, b, \xi_1, \ldots, \xi_n} \frac{1}{2}\mathbf{w}^\top \mathbf{w} + C \sum_{i=1}^{n} \xi_i \text{ such that } d_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 \, \forall \, i$$

C acts as regularizer.

Dual problem

$$\max_{(\alpha_1, \ldots, \alpha_n)} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j d_i d_j \mathbf{x}_i^\top \mathbf{x}_j \text{ s.t. } \sum_{i=1}^{n} \alpha_i d_i = 0, 0 \leq \alpha_i \leq C$$

Isn't it awesome?