Support Vector Machines

Yihui Saw

Massachusetts Institute of Technology

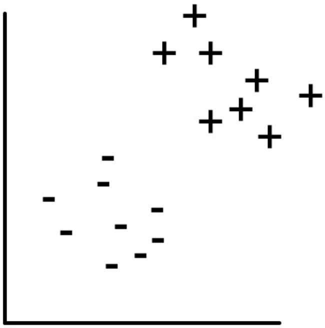
yihui@mit.edu

April 30, 2013

Overview

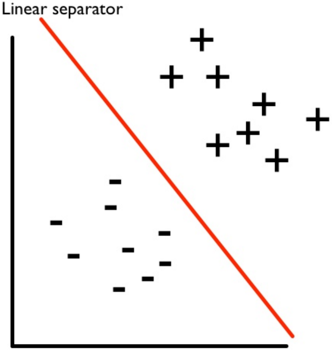
Background/Motivation

Given 2 classes, and a new data point we want to be able to classify this new data point as one of the 2 classes.



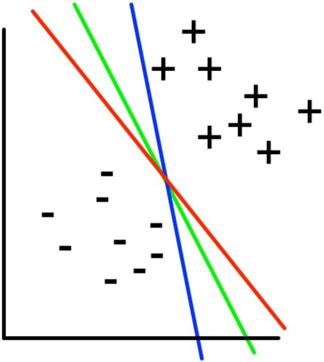
Background/Motivation

Goal: Find a linear separator.



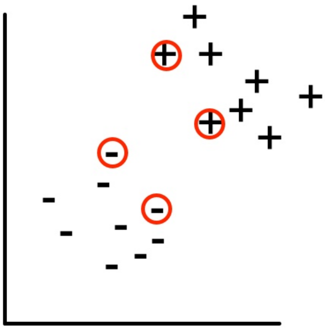
Background/Motivation

But there are many possible linear separators.



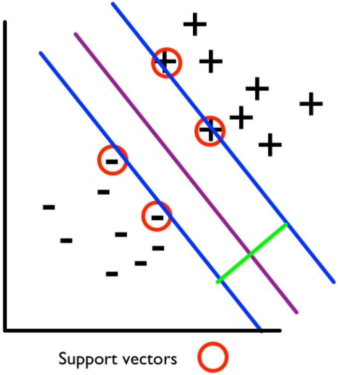
The SVM Model

But there are many possible linear separators.



The SVM Model

Maximize the gap between support vectors.



Formal Definition

Input: Set of samples S where each sample x_i is a vector of d variables $x \in R^d$ and y_i is its class $y \in \{+1, -1\}$.

Goal: Find Θ, Θ_0 such that $y_i(\Theta \cdot x_i + \Theta_0) \geq 1$ while maximizing the gap $\frac{1}{|\Theta|}$

The quadratic program

Primal

min $\frac{1}{2}|\Theta|^2$ subject to $y_i(\Theta \cdot x_i + \Theta_0) \geq 1$ where $i = 1, ..., n$

The dual problem

Dual

max $\sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j)$ subject to
$\alpha_i \geq 0, i = 1, 2, ..., n$

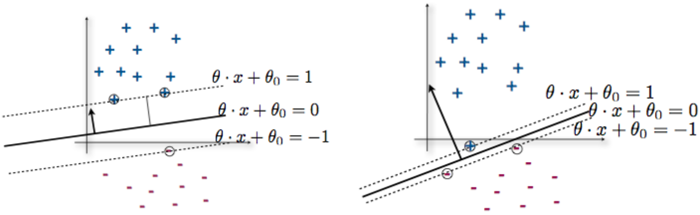
The solution satisfies:

(support vector) $\alpha_i > 0 : \ y_i (\sum_{j=1}^{n} \alpha_j y_j x_j) \cdot x_i = 1$

(non-support vector) $\alpha_i = 0 : \ y_i (\sum_{j=1}^{n} \alpha_j y_j x_j) \cdot x_i \geq 1$

Support Vector Machines with Errors

Sometimes our examples contain errors



Solution: Introduce "slack" variables $\xi_i \geq 0$ to our optimization problem

Support Vector Machines with Errors

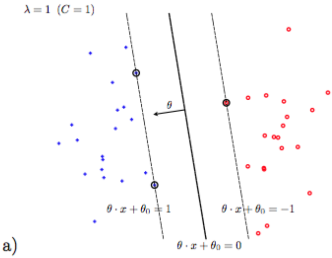
(primal) min $\frac{\lambda}{2}|\Theta|^2 + \frac{1}{n}\sum_{i=1}^{n}\xi_i$

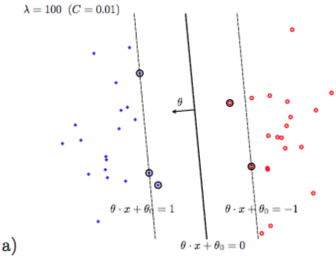
subject to $y_i(\Theta \cdot x_i + \Theta_0) \geq 1 - \xi_i, \xi_i \geq 0, i = 1, ..., n$

λ is the regularization parameter. It balances how much we favor increasing the margin over satisfying the classification constraints.

Support Vector Machines with Errors

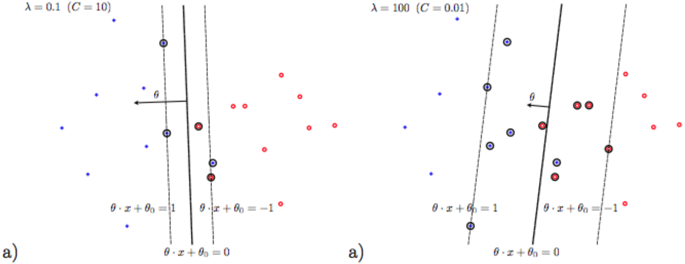
The effect of slack when examples are still linearly separable





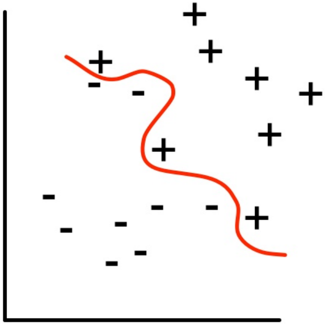
Support Vector Machines with Errors

The effect of slack when examples are no longer linearly separable



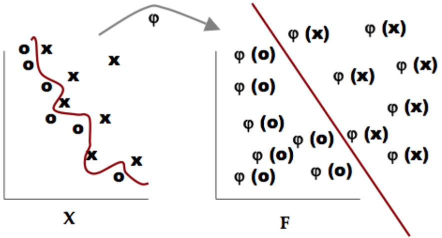
Non-linear problems

Problems that are not linearly separable



Non-linear problems

The idea is to gain linearly separation by mapping the data to a higher dimensional space



Non-linear problems

Recall the dual of the problem:

max $\sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j)$

In the non-linear case, we replace $x_i \cdot x_j$ with $\phi(x_i) \cdot \phi(x_j)$.
So we don't need to know what Φ is explicitly. Calculate Kernel function K instead where

$$K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j)$$

Kernels

With Kernels, we can implicitly work with very high (or even infinite) dimensional feature vectors.
Example the radial basis kernel

$$K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j) = e^{\frac{-|x_i - x_j|^2}{2}}$$

is infinite dimensional.

Kernel function

A kernel function is valid if and only if there exists some map $\phi(x)$ such that

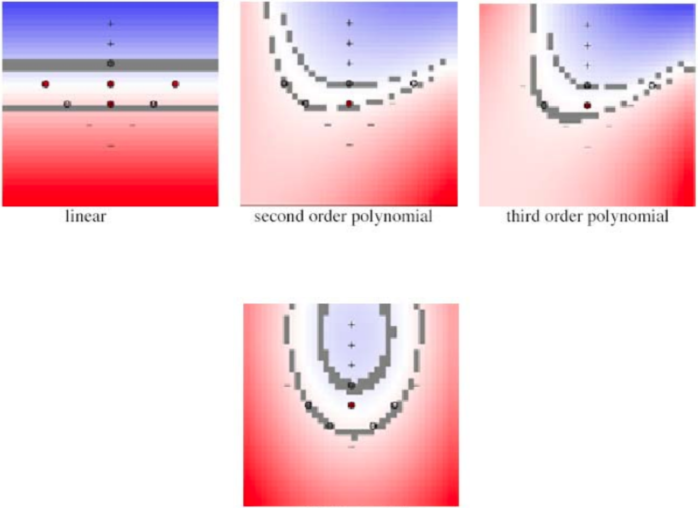
$$K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j)$$

Kernels

Rules

1. $K(x_i, x_j) = 1$ is a kernel function.
2. Let $f : R^d \rightarrow R$ be any real valued function of x. Then, if $K(x_i, x_j)$ is a kernel, function, then so is $\tilde{K}(x_i, x_j) = f(x_i)K(x_i, x_j)f(x_j)$.
3. If $K_1(x_i, x_j)$ and $K_2(x_i, x_j)$ are kernels, then so is their sum.
4. If $K_1(x_i, x_j)$ and $K_2(x_i, x_j)$ are kernels, then so is their product.

Kernels



linear

second order polynomial

third order polynomial

radial basis, 2.0

Applications

- Text (and hypertext) categorization
- Image classification
- Bioinformatics (Protein classification, Cancer classification)
- etc ...

The End