Kelly Technologies

An Introduction to Support Vector Machines

based on slides adapted from Pierre Dönnes' web site of SVM

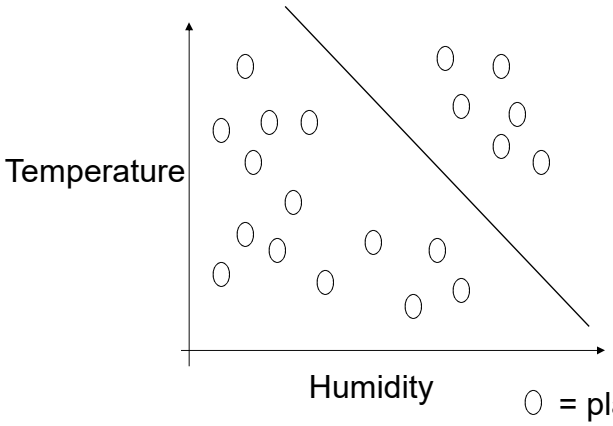
Kelly Technologies

Main Ideas

- Max-Margin Classifier
  - Formalize notion of the best linear separator
- Lagrangian Multipliers
  - Way to convert a constrained optimization problem to one that is easier to solve
- Kernels
  - Projecting data into higher-dimensional space makes it linearly separable
- Complexity
  - Depends only on the number of training examples, not on dimensionality of the kernel space!

Kelly Technologies

Tennis example



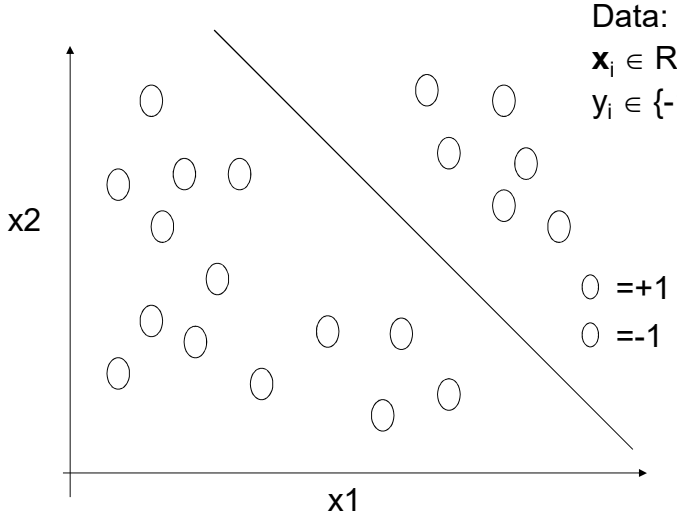
Kelly Technologies

Linear Support Vector Machines

Data: $<\mathbf{x}_i, y_i>$, i=1,..,l

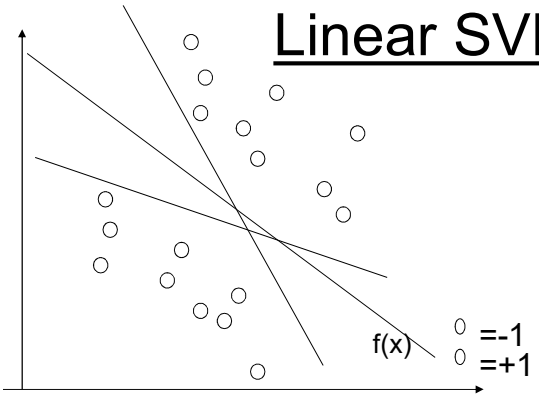
$\mathbf{x}_i \in R^d$

$y_i \in \{-1,+1\}$



Kelly Technologies

Linear SVM 2



Data: $<\mathbf{x}_i, y_i>$, i=1,..,l
$\mathbf{x}_i \in R^d$
$y_i \in \{-1,+1\}$

All hyperplanes in R^d are parameterize by a vector (**w**) and a constant b. Can be expressed as **w**•**x**+b=0 (remember the equation for a hyperplane from algebra!)

Our aim is to find such a hyperplane f(x)=sign(**w**•**x**+b), that correctly classify our data.

Kelly Technologies

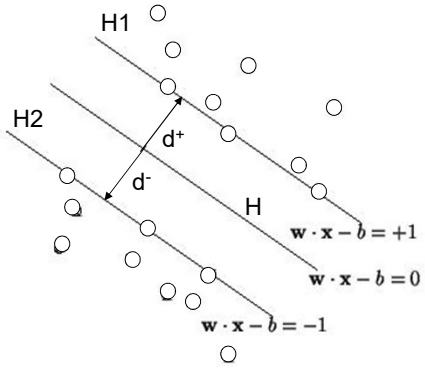
Definitions

Define the hyperplane H such that:

$\mathbf{x}_i$•**w**+b ≥ +1 when y_i =+1

$\mathbf{x}_i$•**w**+b ≤ -1 when y_i =-1

H1 and H2 are the planes:
H1: $\mathbf{x}_i$•**w**+b = +1
H2: $\mathbf{x}_i$•**w**+b = -1
The points on the planes H1 and H2 are the Support Vectors



d+ = the shortest distance to the closest positive point

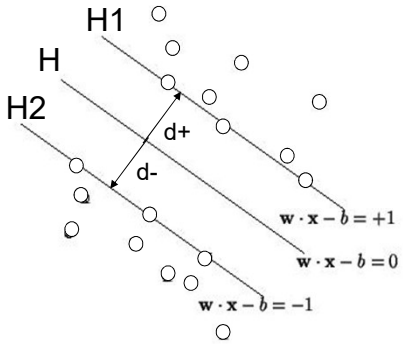
d- = the shortest distance to the closest negative point

The margin of a separating hyperplane is $d^+ + d^-$.

Kelly Technologies

Maximizing the margin

We want a classifier with as big margin as possible.



Recall the distance from a point(x_0,y_0) to a line: Ax+By+c = 0 is|A x_0 +B y_0 +c|/sqrt(A^2+B^2)

The distance between H and H1 is: |**w**•**x**+b|/||w||=1/||w||

The distance between H1 and H2 is: 2/||w||

In order to maximize the margin, we need to minimize ||w||. With the condition that there are no datapoints between H1 and H2:

$\mathbf{x}_i$•**w**+b ≥ +1 when y_i =+1

$\mathbf{x}_i$•**w**+b ≤ -1 when y_i =-1 } **Can be combined into yi($\mathbf{x}_i$•w) ≥ 1**

Kelly Technologies

Constrained Optimization Problem

Minimize $\| \mathbf{w} \| = \langle \mathbf{w} \cdot \mathbf{w} \rangle$ subject to $y_i(\langle \mathbf{x}_i \cdot \mathbf{w} \rangle + b) \geq 1$ for all i

Lagrangian method: maximize $\inf_{\mathbf{w}} L(\mathbf{w}, b, \alpha)$, where

$$L(\mathbf{w}, b, \alpha) = \frac{1}{2} \| \mathbf{w} \| - \sum_i \alpha_i [(y_i(\mathbf{x}_i \cdot \mathbf{w}) + b) - 1]$$

At the extremum, the partial derivative of L with respect both $\mathbf{w}$ and b must be 0. Taking the derivatives, setting them to 0, substituting back into L, and simplifying yields:

$$\text{Maximize } \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} y_i y_j \alpha_i \alpha_j \langle \mathbf{x}_i \cdot \mathbf{x}_j \rangle$$

$$\text{subject to } \sum_i y_i \alpha_i = 0 \text{ and } \alpha_i \geq 0$$

Kelly Technologies

Quadratic Programming

- Why is this reformulation a good thing?
- The problem

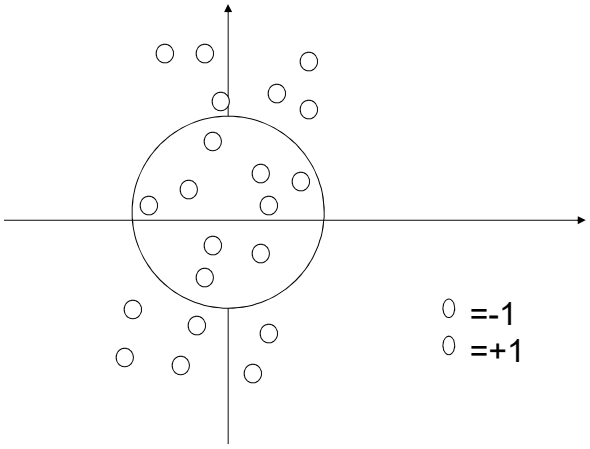
$$\text{Maximize} \sum_i \alpha_i - \frac{1}{2}\sum_{i,j} y_i y_j \alpha_i \alpha_j \left\langle \mathbf{x}_i \cdot \mathbf{x}_j \right\rangle$$

$$\text{subject to} \sum_i y_i \alpha_i = 0 \text{ and } \alpha_i \geq 0$$

  is an instance of what is called a positive, semi-definite programming problem
- For a fixed real-number accuracy, can be solved in $O(n \log n)$ time = $O(|D|^2 \log |D|^2)$

Kelly Technologies

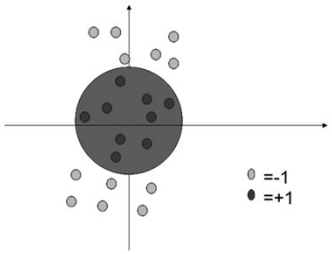
Problems with linear SVM

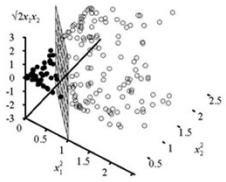


What if the decision function is not a linear?

Kelly es

Kernel Trick





Data points are linearly separable in the space $(x_1^2, x_2^2, \sqrt{2}x_1x_2)$

We want to maximize $\sum_i \alpha_i - \frac{1}{2}\sum_{i,j} y_i y_j \alpha_i \alpha_j \left\langle F(\mathbf{x}_i) \cdot F(\mathbf{x}_j) \right\rangle$

Define $K(\mathbf{x}_i, \mathbf{x}_j) = \left\langle F(\mathbf{x}_i) \cdot F(\mathbf{x}_j) \right\rangle$

Cool thing : K is often easy to compute directly! Here,

$K(\mathbf{x}_i, \mathbf{x}_j) = \left\langle \mathbf{x}_i \cdot \mathbf{x}_j \right\rangle^2$

Kelly Technologies

Other Kernels

The polynomial kernel
$K(x_i,x_j) = (x_i \bullet x_j + 1)^p$, where p is a tunable parameter.
Evaluating K only require one addition and one exponentiation more than the original dot product.

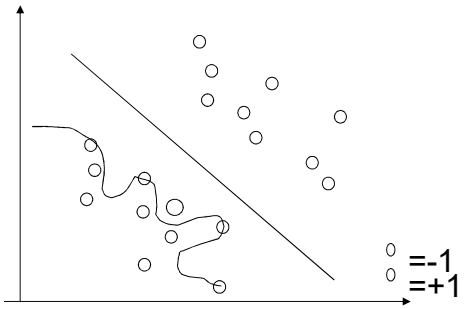
Gaussian kernels (also called radius basis functions)
$K(x_i,x_j) = \exp(||x_i - x_j||^2/2\sigma^2)$

Kelly Technologies

Overtraining/overfitting

A well known problem with machine learning methods is overtraining. This means that we have learned the training data very well, but we can not classify unseen examples correctly.

An example: A botanist really knowing trees.Everytime he sees a new tree, he claims it is not a tree.



Kelly Technologies

Overtraining/overfitting 2

A measure of the risk of overtraining with SVM (there are also other measures).

It can be shown that: The portion, n, of unseen data that will be missclassified is bounded by:

$n \leq$ Number of support vectors / number of training examples

Ockham´s razor principle: Simpler system are better than more complex ones. In SVM case: fewer support vectors mean a simpler representation of the hyperplane.

Example: Understanding a certain cancer if it can be described by one gene is easier than if we have to describe it with 5000.

Kelly Technologies

A Cautionary Example

Image classification of tanks. Autofire when an enemy tank is spotted.
Input data: Photos of own and enemy tanks.
Worked really good with the training set used.
In reality it failed completely.

Reason: All enemy tank photos taken in the morning. All own tanks in dawn.
The classifier could recognize dusk from dawn!!!!

Kelly Technologies

References

http://www.support-vector.net/

AN INTRODUCTION TO SUPPORT VECTOR MACHINES
(and other kernel-based learning methods)
N. Cristianini and J. Shawe-Taylor
Cambridge University Press
2000 ISBN: 0 521 78019 5