Support Vector Machines

Many of these slides are derived from Seyong Kim and Alex Thomo. Thanks!

Linear classifier (E.g., Perceptron)

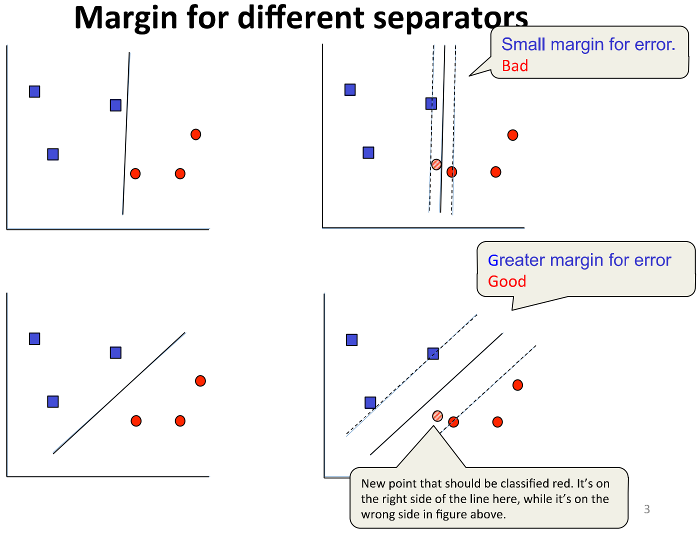
$$h(\mathbf{x}) = sign(\mathbf{w} \cdot \mathbf{x} + b)$$

Which outputs +1 or –1.

Say:
+1 corresponds to blue, and
-1 to red, or vice versa.

Many lines do the job.
Which one to choose?

Margin for different separators



Scale Invariance

$$h(\mathbf{x}) = sign(\mathbf{w} \cdot \mathbf{x} + b)$$

- We rescale **w** and *b* (without changing the line) such that:

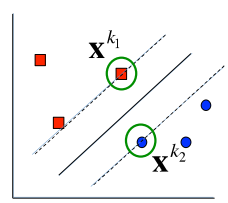
$$\mathbf{w} \cdot \mathbf{x}^{k_1} + b = 1$$

for the closest point(s) to the line on the +1 side, and

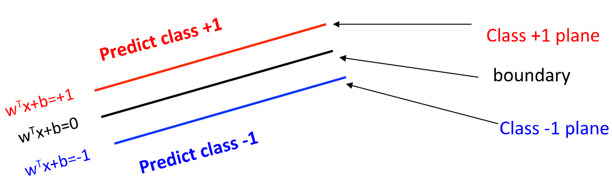
$$\mathbf{w} \cdot \mathbf{x}^{k_2} + b = -1$$

for the closest point(s) to the line on the -1 side.

Closest points are called "support vectors".

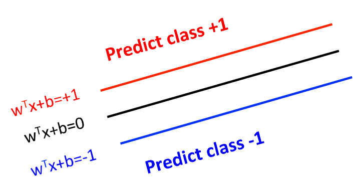


Specifying a max margin classifier



Classify as +1	if	$w^Tx+b \geq 1$
Classify as -1	if	$w^Tx+b \leq -1$
Undefined	if	$-1 < w^Tx+b < 1$

Specifying a max margin classifier

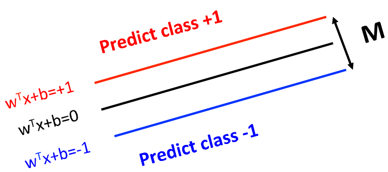


Is the linear separation assumption realistic?

We will deal with this shortly, but lets assume it for now

Classify as +1	if	$w^Tx+b \geq 1$
Classify as -1	if	$w^Tx+b \leq -1$
Undefined	if	$-1 < w^Tx+b < 1$

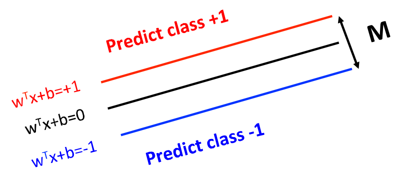
Maximizing the margin



Classify as +1 if $w^Tx+b \geq 1$
Classify as -1 if $w^Tx+b \leq -1$
Undefined if $-1 < w^Tx+b < 1$

- Lets define the width of the margin by M
- How can we encode our goal of maximizing M in terms of our parameters (w and b)?
- Lets start with a few observations

Maximizing the margin



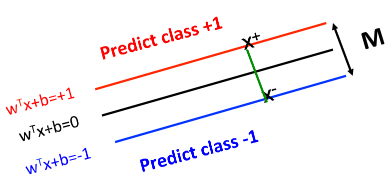
Classify as +1 if $w^Tx+b \geq 1$
Classify as -1 if $w^Tx+b \leq -1$
Undefined if $-1 < w^Tx+b < 1$

- Observation 1: the vector w is orthogonal to the +1 plane
- Why?

Let u and v be two points on the +1 plane, then for the vector defined by u and v we have $w^T(u-v) = 0$

Corollary: the vector w is orthogonal to the -1 plane

Maximizing the margin



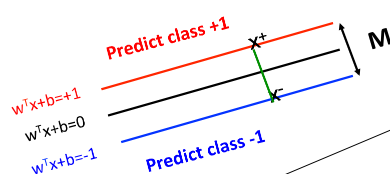
Classify as +1 if $w^Tx+b \geq 1$
Classify as -1 if $w^Tx+b \leq -1$
Undefined if $-1 < w^Tx+b < 1$

- Observation 1: the vector w is orthogonal to the +1 and -1 planes
- Observation 2: if x^+ is a point on the +1 plane and x^- is the closest point to x^+ on the -1 plane then

$$x^+ = \lambda w + x^-$$

Since w is orthogonal to both planes we need to 'travel' some distance along w to get from x^+ to x^-

Putting it together



- $w^T x^+ + b = +1$
- $w^T x^- + b = -1$
- $x^+ = \lambda w + x^-$
- $| x^+ - x^- | = M$

We can now define M in terms of w and b

$w^T x^+ + b = +1$

$\Rightarrow$

$w^T (\lambda w + x^-) + b = +1$

$\Rightarrow$

$w^Tx^- + b + \lambda w^Tw = +1$

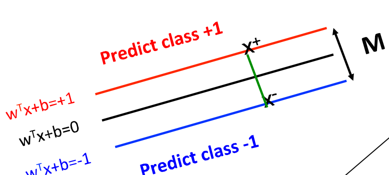
$\Rightarrow$

$-1 + \lambda w^Tw = +1$

$\Rightarrow$

$\lambda = 2/w^Tw$

Putting it together



- $w^T x^+ + b = +1$
- $w^T x^- + b = -1$
- $x^+ = \lambda w + x^-$
- $| x^+ - x^- | = M$
- $\lambda = 2/w^Tw$

We can now define M in terms of w and b

$M = |x^+ - x^-|$

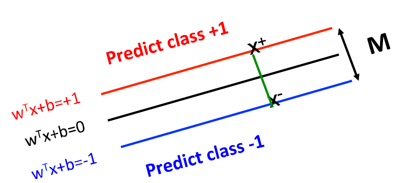
$\Rightarrow$

$$M = |\lambda w| = \lambda |w| = \lambda \sqrt{w^T w}$$

$\Rightarrow$

$$M = 2\frac{\sqrt{w^T w}}{w^T w} = \frac{2}{\sqrt{w^T w}}$$

Finding the optimal parameters



$$M = \frac{2}{\sqrt{w^T w}}$$

We can now search for the optimal parameters by finding a solution that:

1. Correctly classifies all points
2. Maximizes the margin (or equivalently minimizes w^Tw)

Several optimization methods can be used: Gradient descent, simulated annealing, EM etc.

Aside: Quadratic programming (QP)

Quadratic programming solves optimization problems of the following form:

$$\min_U \frac{u^T R u}{2} + d^T u + c$$

Quadratic term

subject to n inequality constraints:

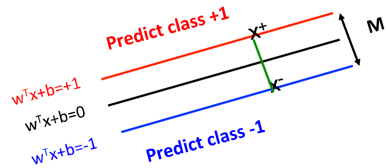
$$a_{11}u_1 + a_{12}u_2 + \ldots \le b_1$$
$$\vdots \quad \vdots \quad \vdots$$
$$a_{n1}u_1 + a_{n2}u_2 + \ldots \le b_n$$

and k equivalency constraints:

$$a_{n+1,1}u_1 + a_{n+1,2}u_2 + \ldots = b_{n+1}$$
$$\vdots \quad \vdots \quad \vdots$$
$$a_{n+k,1}u_1 + a_{n+k,2}u_2 + \ldots = b_{n+k}$$

When a problem can be specified as a QP problem we can use solvers that are better than gradient descent or simulated annealing

SVM as a QP problem



$$M = \frac{2}{\sqrt{w^T w}}$$

$$\min_U \frac{u^T R u}{2} + d^T u + c$$

subject to n inequality constraints:

$$a_{11}u_1 + a_{12}u_2 + \ldots \le b_1$$
$$\vdots \quad \vdots \quad \vdots$$
$$a_{n1}u_1 + a_{n2}u_2 + \ldots \le b_n$$

and k equivalency constraints:

$$a_{n+1,1}u_1 + a_{n+1,2}u_2 + \ldots = b_{n+1}$$
$$\vdots \quad \vdots \quad \vdots$$
$$a_{n+k,1}u_1 + a_{n+k,2}u_2 + \ldots = b_{n+k}$$

Min $(w^Tw)/2$

subject to the following inequality constraints:

For all x in class + 1

$w^Tx+b \ge 1$

For all x in class - 1

$w^Tx+b \le -1$

A total of n constraints if we have n input samples

SVM as a QP problem: a simplication

Min $(w^Tw)/2$

subject to the following inequality constraints:

For all x in class + 1

$w^Tx+b \ge 1$

For all x in class - 1

$w^Tx+b \le -1$

Min $(w^Tw)/2$

subject to the following inequality constraints:

For all x in class + 1

$y(w^Tx+b) \ge 1$

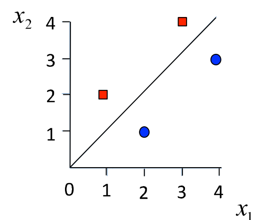
For all x in class - 1

$y(w^Tx+b) \ge 1$

The same constraint!!
So much easier to handle!

Example

Training tuples:
([1, 2], +1)
([2, 1], −1)
([3, 4], +1)
([4, 3], −1)



$$\min_{w_1,w_2} \frac{1}{2}\left(w_1^2 + w_2^2\right)$$

subject to

$$(+1)(w_1 + 2w_2) \ge 1$$
$$(-1)(2w_1 + w_2) \ge 1$$
$$(+1)(3w_1 + 4w_2) \ge 1$$
$$(-1)(4w_1 + 3w_2) \ge 1$$

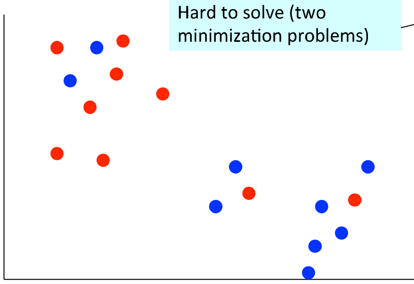
For this example, solution is easy to see:
b = 0 and w = [−1, +1], i.e. the line is:
$-x_1+x_2=0$
All conditions are satisfied.

$$M = \frac{2}{\|\mathbf{w}\|} = \frac{2}{\sqrt{1+1}} = \frac{2}{\sqrt{2}} = \sqrt{2}$$

Non linearly separable case

- So far we assumed that a linear plane can perfectly separate the points
- But this is not usally the case

- noise, outliers



Hard to solve (two minimization problems)

How can we convert this to a QP problem?

- Minimize training errors?

min w^Tw

min #errors

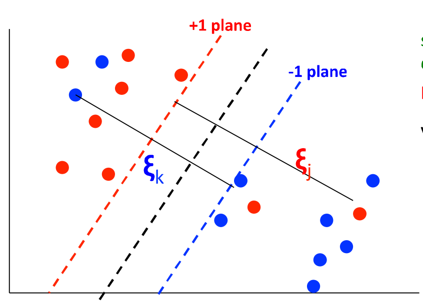
- Penalize training errors:

min w^Tw+C*(#errors)

Hard to encode in a QP problem

Non linearly separable case

- Instead of minimizing the number of misclassified points we can minimize the *distance* between these points and their correct plane



The new optimization problem is:

$$\min_w \frac{w^T w}{2} + \sum_{i=1}^{n} C\xi_i$$

subject to the following inequality constraints:

For all x_i in class +1 or class -1

$y^*(w^Tx+b) \ge 1 - \xi_i$

Wait. Are we missing something?

Final optimization for non linearly separable case

The new optimization problem is:

$$\min_w \frac{w^T w}{2} + \sum_{i=1}^{n} C\xi_i$$

subject to the following inequality constraints:

For all x_i in class +1

or class -1

$y*(w^Tx+b) \geq 1- \xi_i$

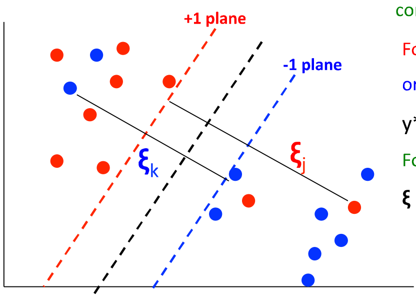
n constraints

For all i

$\xi_i \geq 0$

Another n constraints

Slack variables (ε)



Where we are

Two optimization problems: For the separable and non separable cases

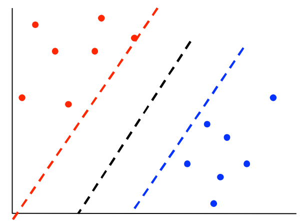
$$\min_w \frac{w^T w}{2}$$

For all x in class + 1

$w^Tx+b \geq 1$

For all x in class - 1

$w^Tx+b \leq -1$



$$\min_w \frac{w^T w}{2} + \sum_{i=1}^{n} C\xi_i$$

For all x_i in class + 1

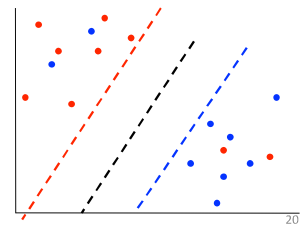
$w^Tx+b \geq 1- \xi_i$

For all x_i in class - 1

$w^Tx+b \leq -1+ \xi_i$

For all I

$\xi_I \geq 0$



Where we are

Two optimization problems: For the separable and non separable cases

$$\min_w \frac{w^T w}{2}$$

For all x in class + 1

$w^Tx+b \geq 1$

For all x in class - 1

$w^Tx+b \leq -1$

$$\min_w \frac{w^T w}{2} + \sum_{i=1}^{n} C\xi_i$$

For all x_i in class + 1

$w^Tx+b \geq 1- \xi_i$

For all x_i in class - 1

$w^Tx+b \leq -1+ \xi_i$

For all I

$\xi_I \geq 0$

- Instead of solving these QPs directly we will solve a dual formulation of the SVM optimization problem
- The main reason for switching to this type of representation is that it would allow us to use a neat trick that will make our lives easier (and the run time faster)