Classification – recap.

The goal for classification is to draw a hyperplane of dimension $d-1$ that linearly separates data points of different classes.

Consider it the separating line. The key properties of the line $y = f(x)$:

$f(x) = \boldsymbol{w}^T\boldsymbol{x} + b$ where $\boldsymbol{x} = [x_1, x_2, \dots, x_N]^T$

This can be extended to non-linear spaces:

$y(\boldsymbol{x}) = \boldsymbol{w}^T\phi(\boldsymbol{x}) + b$

Sparse Kernel Machines

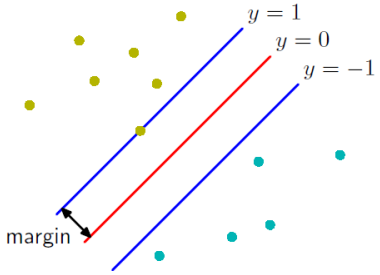
Maximum Margin Classifiers

Consider binary classification using linear models with $\{x_n\}_{n=1}^N$ inputs with corresponding target values $\{t_n\}_{n=1}^N$ where $t_n \in \{-1,1\}$. Then:

$$y(\boldsymbol{x}) = \boldsymbol{w}^T\phi(\boldsymbol{x}) + b$$

ϕ is a feature-space transformation and $\begin{cases} t_n = +1,\ if\ y(x_n) > 0 \\ t_n = -1,\ if\ y(x_n) < 0 \end{cases}$

Margin: the smallest distance between the decision boundary and any of the samples, which we want to maximize:



We want to maximize the distance of a point x_n to the decision boundary:

$$\frac{t_n y(x_n)}{\|w\|} = \frac{t_n(w^T\phi(x_n) + b)}{\|w\|}$$

We want to find the parameters that maximize the minimal distance to the boundary:

$$\arg\max_{\mathrm{w,b}}\left\{\frac{1}{\|w\|}\min_n[t_n(w^T\phi(x_n) + b]\right\}$$

Since we are taking the norm $\|w\|$, we can w.l.o.g. define that the closest distance $t_n(w^t\phi(x_n) + b) = 1$

Which means: $t_n(w^T\phi(x_n) + b) \geq 1$ - the distance of all data points is at least 1.

We call the data points closest to the boundary active points.

Then the max problem turns into a min problem: $\arg\min_{w,b}\frac{1}{2}\|w\|^2$

Adding Lagrange multipliers $a_n \geq 0$:

$$L(w, b, a) = \frac{1}{2}\|w\|^2 - \sum_{n=1}^{N} a_n\{t_n(w^T\phi(x) + b) - 1\}$$

From deriving $\frac{\partial L}{\partial w} = 0$ and $\frac{\partial L}{\partial b} = 0$ we get:

$$w = \sum_{n=1}^{N} a_n t_n \phi(x_n), \quad \sum_{n=1}^{N} a_n t_n = 0$$

$$L(w,b,a)=\frac{1}{2}\|w\|^2-\sum_{n=1}^{N}a_n\{t_n(w^T\phi(x_n)+b)-1\}$$

$$\frac{\partial L}{\partial w}=0\Rightarrow w=\sum_{n=1}^{N}a_n t_n\phi(x_n)$$

$$\frac{\partial L}{\partial b}=0\Rightarrow\sum_{n=1}^{N}a_n t_n=0$$

Let's plug these expressions back:

$$\hat{L}(a)=\frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}a_n a_m t_n t_m\phi(x_n)^T\phi(x_m)-\sum_{n=1}^{N}\sum_{m=1}^{N}a_n\{t_n(a_m t_m\phi(x_m)^T\phi(x_n)+b)-1\}=$$

$$-\frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}a_n a_m t_n t_m\phi(x_n)^T\phi(x_m)+\sum_{n=1}^{N}a_n=$$

$$-\frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}a_n a_m t_n t_m k(x_n,x_m)+\sum_{n=1}^{N}a_n$$

Where $k(x_n,x_m)\coloneqq\phi(x_n)^T\phi(x_m)$ is the <u>kernel function</u> – the dot-product the feature transform gives.

The condition $\sum_{n=1}^{N}a_n t_n=0$ is still satisfied where $a_n\geq 0$.

<u>Classification</u>:

To classify a new data point x we just need to run it through the kernel function with every data point, sum over it and see if it's positive or negative:

$$y(x)=w^T\phi(x)+b=\sum_{n=1}^{N}a_n t_n k(x,x_n)+b$$

<u>The KKT conditions</u>: For an equation of the form (λ is the Lagrange multipliers)

$L(x,\lambda)=f(x)-\lambda g(x)$

$g(x)\geq 0$

$\lambda\geq 0$

$\lambda g(x)\geq 0$

Here:

$a_n\geq 0$

$t_n y(x_n)-1\geq 0$

$a_n\{t_n y(x_n)-1\}=0$

Which means for every data point either $a_n=0$ or $t_n y(x_n)=1$ – the support vectors.

If $a_n=0$ for a particular data point, it is thrown away – so only part of the data is used for the classification (where a_n is not 0).

Note: the kernel has to be a PSD.

For instance, in a poly-kernel $k(x,x')=(x,x')^d$; for a Gaussian kernel: $k(x,x')=\exp\left[-\frac{|x-x'|^2}{\gamma}\right]$

The support vectors are just the closest points to the boundary – those of distance 1.

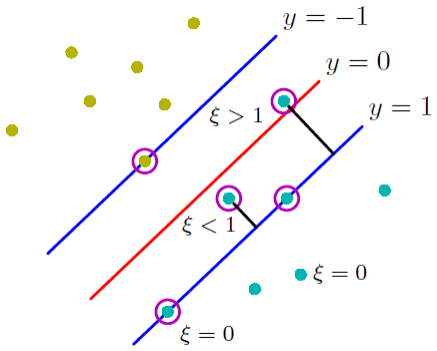
This works only if the points are linearly separable. Now we want to extend this concept to cases where the classes overlap, as long as the overlap is not significant.

Overlapping class distributions

We can draw the decision surface such that some tolerance is allowed, using Slack Variables:

$\xi_n \geq 0 (n = 1, \dots, N)$ – one for each data point

Where: $\xi_n = \begin{cases} 0, & x_n \; correctly \; classified \\ |t_n - y(x_n)|, & otherwise \end{cases}$



Correct points: $\xi_n = 0$

Points inside the margin on the correct side: $0 < \xi_n \leq 1$

Points on the wrong side: $\xi_n > 1$

Now the classification constraint is:

$$t_n y(x_n) \geq 1 - \xi_n, n = 1, \dots, N$$

And minimize:

$$C \sum_{n=1}^{N} \xi_n + \frac{1}{2} \|w\|^2$$

Where $C \to \infty$ reduces to linearly-separable SVM.

So now:

$$L(w, b, a) = \frac{1}{2} \|w\|^2 + C \sum_{n=1}^{N} \xi_n - \sum_{n=1}^{N} a_n \{t_n y(x_n) - 1 + \xi_n\} - \sum_{n=1}^{N} \mu_n \xi_n$$

Where a_n, μ_n are Lagrange multipliers.

The derivation is similar to before, and is on slide 28 – **may be in the Homework**

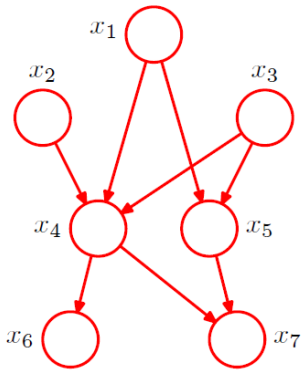
It's like the previous SVM only including the slack variables.

Multiple-Class SVMs

- One-versus-rest: K separate SVMs – the more conventional of the two.
- One-versus-one: $\frac{K(K-1)}{2}$ 2-class SVMs

Graphical Models

The joint dist. Of the graphical model on slide 37:



$p(x_1)p(x_2)p(x_3)p(x_4|x_1,x_2,x_3)p(x_5|x_1,x_3)p(x_6|x_4)p(x_7|x_4,x_5)$

You can't have a cycle, otherwise there's a cyclic dependency.