Week 7
SVM

SVM – Basic Concepts

training set:

N input vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$

N target value: $t_1, t_2, \ldots, t_N$ and $t_n \in \{-1, 1\}, n = 1, 2, \ldots N$

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

Now we assume data is linearly separable, then we got:

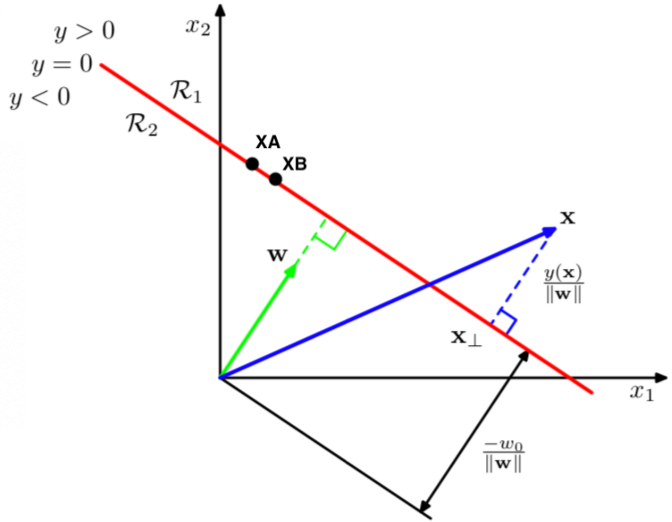
- $y(\mathbf{x}_n) > 0$ for $t_n = +1$
- $y(\mathbf{x}_n) < 0$ for $t_n = -1$

Notice: $t_n y(\mathbf{x}_n) > 0$ holds for all training data points.

SVM – Basic Concepts

margin:

Margin is the smallest distance between decision boundary and any of the samples and our object is to maximise this margin.



$y(\mathbf{x}_A) = y(\mathbf{x}_B) = 0$ (because they are both in the decision boundary)

so we got: $\mathbf{w}^T(\mathbf{x}_A - \mathbf{x}_B) = 0$

The vector $\mathbf{w}$ is orthgonal to every vetor within decision surface.

SVM – The large margin principal

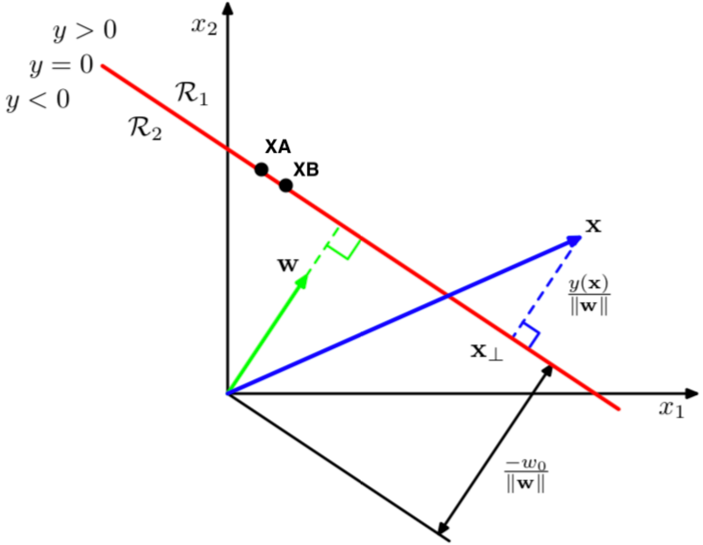
Assuming $\phi(\mathbf{x}) = \mathbf{x}$, so $y(x) = \mathbf{w}^T\mathbf{x} + w_0$

For points in the boundary:

$$y(\mathbf{x}) = 0$$

$$\mathbf{w}^T\mathbf{x} + w_0 = 0$$

$$\frac{\mathbf{w}^T\mathbf{x}}{||\mathbf{w}||} = -\frac{w_0}{||\mathbf{w}||}$$



For arbitrary $\mathbf{x}$

$\mathbf{x} = \mathbf{x}_\perp + r\frac{\mathbf{w}}{||\mathbf{w}||}$ ($\mathbf{x}_\perp$ is the orthgonal projection onto the decision surface)

$\frac{\mathbf{w}}{||\mathbf{w}||}$ provides the unit vector, so how to calculate r (the distance from the decision boundary whose normal vector is $\mathbf{w}$?

SVM – The large margin principal

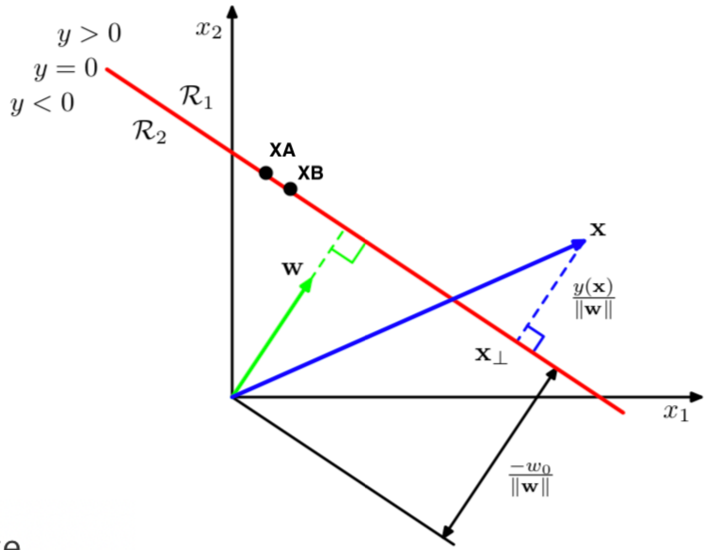
substract $\mathbf{x} = \mathbf{x}_\perp + r\frac{\mathbf{w}}{||\mathbf{w}||}$ into $y(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0$

$$f(x) = \mathbf{w}^T\mathbf{x} + w_0 = (\mathbf{w}^T\mathbf{x}_\perp + w_0) + r\frac{\mathbf{w}^T\mathbf{w}}{||\mathbf{w}||}$$

$$\because y(\mathbf{x}_\perp) = \mathbf{w}^T x_\perp + w_0 = 0$$

$$\therefore y(\mathbf{x}) = r\frac{\mathbf{w}\mathbf{w}^T}{||\mathbf{w}||}$$

$$\therefore r = \frac{y(\mathbf{x})}{||\mathbf{w}||}$$



Notice: $\mathbf{w}$ determines the orientation of the decision surface

w_0 dertermines the location of the decision surface

SVM – The large margin principal

Now, we would like to make this distance $r = \frac{y(\mathbf{x})}{||\mathbf{w}||}$ as large as possible.

In addition, we want to ensure each point is on the correct side of the boundary. For all points correctly classified:

$$t_n y(\mathbf{x}_n) > 0$$

The distance of a point x_n to the decision surface is defined by:

$$\frac{t_n y(\mathbf{x}_n)}{||\mathbf{w}||} = \frac{tn(\mathbf{w}^T \phi(\mathbf{x}_n) + w_0)}{||\mathbf{w}||}$$

So, maximise margin is found by:

$$\underset{\mathbf{w}, w_0}{\operatorname{argmax}}\{\frac{1}{||\mathbf{w}||} \min_n [\mathbf{t}_n (\mathbf{w}^T \phi(\mathbf{x}_n) + w_0)]\}$$

We notice that,

if $\mathbf{w} \rightarrow k\mathbf{w}, w_0 \rightarrow kw_0$, $\frac{\mathbf{t}_n y(\mathbf{x}_n)}{||\mathbf{w}||}$ is unchanged. Since the k factor cancels out when we divide by $||\mathbf{w}||$

SVM – The large margin principal

So now we can choose a factor k such that $\mathbf{t}_n(\mathbf{w}^T\phi(\mathbf{x}_n) + w_0) = 1$ for support vectors, then all data points will satisfy:

$$\mathbf{t}_n(\mathbf{w}^T\phi(\mathbf{x}_n) + w_0) \geq 1, n = 1, \ldots, N$$

The problem now is:

maximise $\frac{1}{||\mathbf{w}||}$, and this is equivalent to minimise $||\mathbf{w}||^2$

so,

$$\underset{\mathbf{w},w_0}{\operatorname{argmax}}\{\frac{1}{||\mathbf{w}||}\min_n[\mathbf{t}_n(\mathbf{w}^T\phi(\mathbf{x}_n) + w_0)]\} \Rightarrow \underset{\mathbf{w},w_0}{\operatorname{argmin}} \frac{1}{2}||\mathbf{w}||^2$$

s.t. $\mathbf{t}_n(\mathbf{w}^T\phi(\mathbf{x}_n) + w_0) \geq 1$

SVM – The large margin principal

****

In order to solve this, we use Lagrange multipliers: $a_n \geq 0$

$$\mathcal{L}(\mathbf{w}, w_0, a) = \frac{1}{2}||\mathbf{w}||^2 - \sum_{n=1}^{N} a_n\{t_n(\mathbf{w}^T\phi(\mathbf{x}_n) + w_0) - 1\}$$

Why "$-$" sign? $\Rightarrow$ Because we want to minimise with respect to $\mathbf{w}$ and b, and maximise with respect to a .

Then we compute the partial derivative with respect to $\mathbf{w}$ and b respectively and set them to 0:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}} = 0 \qquad \frac{\partial \mathcal{L}}{\partial w_0} = 0$$

$$\mathbf{w} = \sum_{n=1}^{N} a_n t_n \phi(\mathbf{x}_n) \qquad \sum_{n=1}^{N} a_n t_n = 0$$

Then we eliminate $\mathbf{w}$ and b from $\mathcal{L}(\mathbf{w}, w_0, a)$:

$$\tilde{\mathcal{L}}(a) = \sum_{n=1}^{N} a_n - \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} a_n a_m t_n t_m k(\mathbf{x}_n, \mathbf{x}_m)$$

s.t. $a_n \geq 0, \sum_{n=1}^{N} a_n t_n = 0$

****

SVM – Soft margin

In the previous section, we assume the data is linearly separable, what if the data is not linearly separable?

$\Rightarrow$ we can use kernel trick

But what if the data is still not linearly separable after using kernel trick?

$\Rightarrow$ We introduce slack variables $\xi_n \geq 0$.

$$\begin{cases} \xi_n = 0 \text{ if the point is on or inside the correct margin boundary} \\ \xi_n = |t_n - y(\mathbf{x}_n)| \text{ othereise} \end{cases}$$

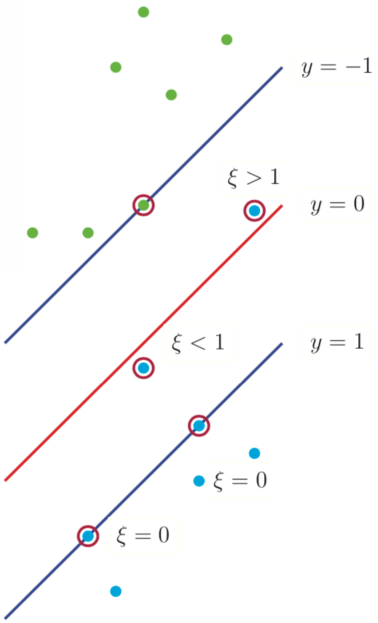
SVM – Soft margin

e.g.

If $0 < \xi_n \leq 1$, the point lies inside the margin, but on the correct side of the decision boundary.

If $\xi_n > 1$, the point lies on the wrong side of the decision boundary.

We now replace the hard constraints that $t_n y(\mathbf{x}_n) \geq 0$ with the **soft margin constraints** that $t_n y(\mathbf{x}_n) \geq 1 - \xi_n$



The new objective becomes :

$\min_{\mathbf{w}, w_0, \xi} \frac{1}{2}||\mathbf{w}||^2 + C \sum_{n=1}^{N} \xi_n$

s.t. $\xi_n \geq 0$ and $t_n(\mathbf{x}_n^T \mathbf{w} + w_0) \geq 1 - \xi_n$

C is a regularisation parameter to control the number of erros we can tolerate.

< we usually define $C = \frac{1}{\nu N}$, ν is used to contro the fraction of misclassified points, and $0 < \nu \leq 1$. This is called ν-SVM. >

SVM

****

Probabilistic output

Notice, the ouput of SVM is a hard-labeling (sign$(y(\mathbf{x})$)), unlike logistic regression, which produces probabilistic output. So how can we convert the hard-labeling into probabilistic output?

$\Rightarrow p(t=1|\mathbf{x},\theta)=\sigma(ay(\mathbf{x})+b)$

where a, b can be estimated by maximum likelihood on a separate validation set.

****

Multi-class classification

$\Rightarrow$ One-versus-all (OVA)

$\Rightarrow$ One-versus-one (OVO)

SVM – example

Given a training set consisting of four points $([1,2],+1)$, $([3,4],+1)$, $([2,1],-1)$ and $([4,3],-1)$, find the corrdinates of $\mathbf{w}=[u,v]$ and b that minimise $||\mathbf{w}||$ subject to:

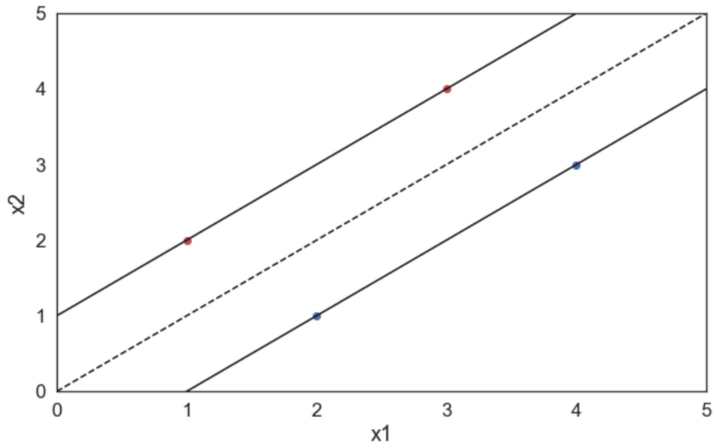
$$y_i(\mathbf{w}\cdot\mathbf{x}_i+b)\geq 1 \text{ for all } i=1,\ldots,4$$

SVM

Given a training set consisting of four points $([1,2],+1)$, $([3,4],+1)$, $([2,1],-1)$ and $([4,3],-1)$, find the corrdinates of $\mathbf{w}=[u,v]$ and b that minimise $||\mathbf{w}||$ subject to:

$$y_i(\mathbf{w}\cdot\mathbf{x}_i+b)\geq 1 \text{ for all } i=1,\ldots,4$$

x1	x2	y
1	2	+1
3	4	+1
2	1	-1
4	3	-1



SVM

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

we find $y(x) = 0$ intersects $(0, 0)$, so $x_1 = x_2$

$y(x) = w_1x_1 + w_2x_2 + b$

$cx_1 = cx_2$

$cx_1 - cx_2 = 0$

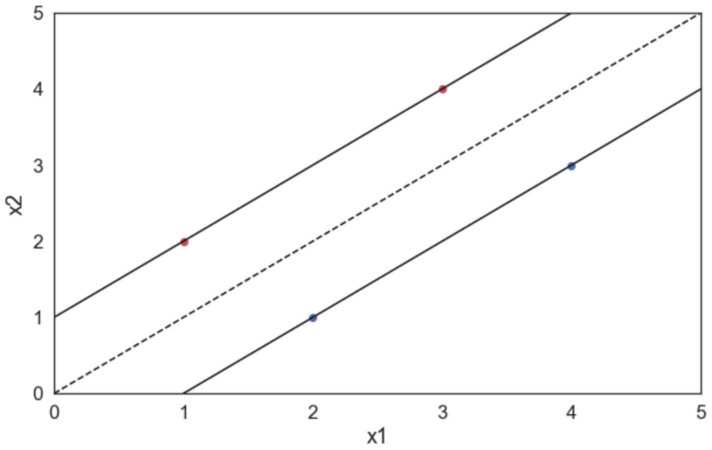
$[c - c]^T \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + 0 = 0$

$\therefore \mathbf{w} = \begin{bmatrix} c \\ -c \end{bmatrix}, b = 0$

substract $\mathbf{w}$ and b into the equation with four points:

$$\begin{cases} c - 2c \geq 1 \\ 3c - 4c \geq 1 \\ 2x - c \leq -1 \\ 4c - 3c \leq -1 \end{cases} \quad \Rightarrow c \leq -1$$

x1	x2	y
1	2	+1
3	4	+1
2	1	-1
4	3	-1



$\min_{\mathbf{w}} ||\mathbf{w}|| = \min_{c} |c|$

and this is subject to $c \leq -1$

$\therefore c = -1$, $\mathbf{w} = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$ and $b = 0$