UNIVERSITY
of
GLASGOW

Machine Learning

Lecture. 10.

Mark Girolami

girolami@dcs.gla.ac.uk

Department of Computing Science
University of Glasgow

SVM

UNIVERSITY of GLASGOW

- Discriminative classifiers directly provide a discriminant function of the form $f(\mathbf{x};\mathbf{w}) = \mathbf{w}^{\mathsf{T}}\phi(\mathbf{x})$

SVM

UNIVERSITY *of* GLASGOW

- Discriminative classifiers directly provide a discriminant function of the form $f(\mathbf{x};\mathbf{w}) = \mathbf{w}^{\mathsf{T}}\phi(\mathbf{x})$
- Simple binary linear discriminant on 2-d feature vector

$$g(\mathbf{x}; w_2, w_1, w_0) = w_2 x_2 + w_1 x_1 + w_0 = \mathbf{w}^{\mathsf{T}}\mathbf{x} + w_0$$

SVM

- Discriminative classifiers directly provide a discriminant function of the form $f(\mathbf{x}; \mathbf{w}) = \mathbf{w}^\mathsf{T}\phi(\mathbf{x})$
- Simple binary linear discriminant on 2-d feature vector

$$g(\mathbf{x}; w_2, w_1, w_0) = w_2 x_2 + w_1 x_1 + w_0 = \mathbf{w}^\mathsf{T}\mathbf{x} + w_0$$

- If target values ± 1 test $g(\mathbf{x}; w_2, w_1, w_0)$ positive or negative so $f(\mathbf{x}; w_2, w_1, w_0) = \mathrm{sign}(\mathbf{w}^\mathsf{T}\mathbf{x} + w_0)$

SVM

- Discriminative classifiers directly provide a discriminant function of the form $f(\mathbf{x};\mathbf{w}) = \mathbf{w}^{\mathsf{T}}\phi(\mathbf{x})$
- Simple binary linear discriminant on 2-d feature vector

$$g(\mathbf{x}; w_2, w_1, w_0) = w_2 x_2 + w_1 x_1 + w_0 = \mathbf{w}^{\mathsf{T}}\mathbf{x} + w_0$$

- If target values ± 1 test $g(\mathbf{x}; w_2, w_1, w_0)$ positive or negative so $f(\mathbf{x}; w_2, w_1, w_0) = \mathrm{sign}(\mathbf{w}^{\mathsf{T}}\mathbf{x} + w_0)$
- For N data points $(\mathbf{x}_1, t_1)\cdots(\mathbf{x}_N, t_N)$ assume classes completely linearly separable then training data correctly classified if

$$t_n(\mathbf{w}^{\mathsf{T}}\mathbf{x} + w_0) > 0 \;\; \forall \;\; n = 1 \cdots N$$

SVM

UNIVERSITY
of
GLASGOW

- Many possible solutions

SVM

UNIVERSITY of GLASGOW

- Many possible solutions

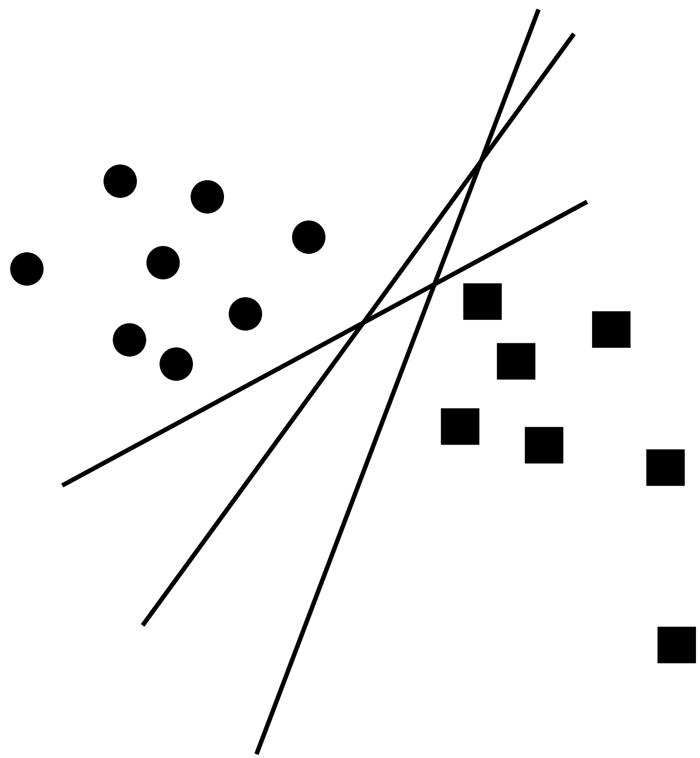


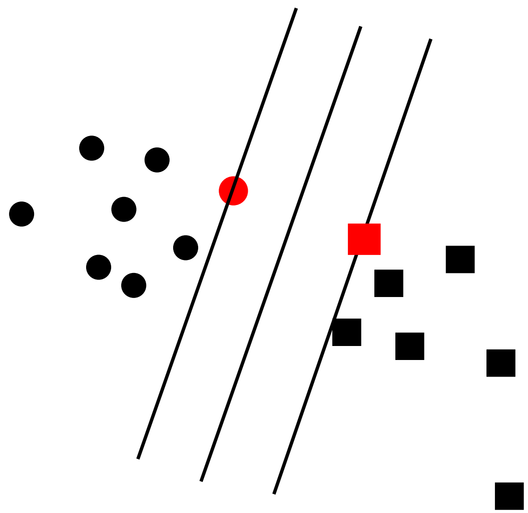
Figure 1: The samples of two classes denoted by sold circles and squares can be separated perfectly with no miss-classifications by a number of possible $\mathbf{w}$ some examples of which are drawn on this cartoon.

SVM

- Upper-bound on generalisation error inversely proportional to perpendicular distance from separating hyperplane, $\mathbf{w}$ and hyperplane through closest points from both classes

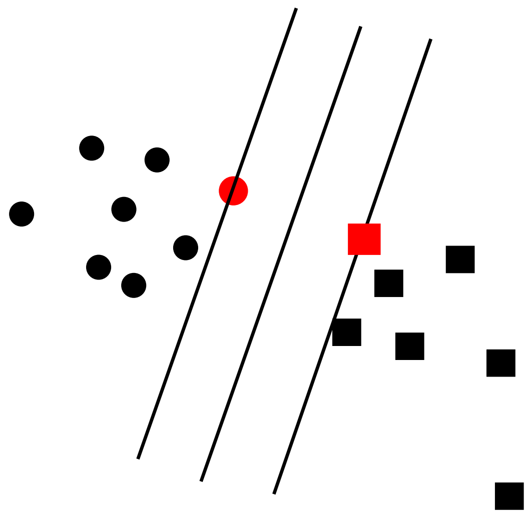
SVM

- Upper-bound on generalisation error inversely proportional to perpendicular distance from separating hyperplane, $\mathbf{w}$ and hyperplane through closest points from both classes



SVM

- Upper-bound on generalisation error inversely proportional to perpendicular distance from separating hyperplane, $\mathbf{w}$ and hyperplane through closest points from both classes



- Called the *margin* so to minimise bound on generalisation error we seek to maximise the *margin* of our classifier

SVM

UNIVERSITY *of* GLASGOW

- Distance of $\mathbf{x}$ to hyper-plane H defined by all points that satisfy $\mathbf{w}^{\mathsf{T}}\mathbf{x} + w_0 = 0$ is given by

$$\frac{\mathbf{w}^{\mathsf{T}}\mathbf{x} + w_0}{||\mathbf{w}||}$$

SVM

- Distance of $\mathbf{x}$ to hyper-plane H defined by all points that satisfy $\mathbf{w}^\mathsf{T}\mathbf{x} + w_0 = 0$ is given by

$$\frac{\mathbf{w}^\mathsf{T}\mathbf{x} + w_0}{||\mathbf{w}||}$$

- If $\mathbf{x}_1^*$ and $\mathbf{x}_2^*$ are closest points from each class to $\mathbf{w}$ margin of separation is

$$\frac{\mathbf{w}^\mathsf{T}\mathbf{x}_1^* + w_0}{||\mathbf{w}||} - \frac{\mathbf{w}^\mathsf{T}\mathbf{x}_2^* + w_0}{||\mathbf{w}||} = \frac{\mathbf{w}^\mathsf{T}}{||\mathbf{w}||}(\mathbf{x}_1^* - \mathbf{x}_2^*)$$

SVM

UNIVERSITY of GLASGOW

- SVM discriminant is $\mathrm{sign}(\mathbf{w}^\mathsf{T}\mathbf{x} + w_0)$ decision made invariant to arbitrary rescaling of $\mathbf{w}^\mathsf{T}\mathbf{x} + w_0$

SVM

UNIVERSITY *of* GLASGOW

- SVM discriminant is $\mathrm{sign}(\mathbf{w}^\mathsf{T}\mathbf{x} + w_0)$ decision made invariant to arbitrary rescaling of $\mathbf{w}^\mathsf{T}\mathbf{x} + w_0$
- Define *canonical* hyper-plane $\mathbf{w}$ such that $\mathbf{w}^\mathsf{T}\mathbf{x}_1^* + w_0 = 1$ and $\mathbf{w}^\mathsf{T}\mathbf{x}_2^* + w_0 = -1$ in which case the margin is now simply $\frac{2}{||\mathbf{w}||}$

SVM

UNIVERSITY *of* GLASGOW

- SVM discriminant is $\mathrm{sign}(\mathbf{w}^{\mathsf{T}}\mathbf{x} + w_0)$ decision made invariant to arbitrary rescaling of $\mathbf{w}^{\mathsf{T}}\mathbf{x} + w_0$
- Define *canonical* hyper-plane $\mathbf{w}$ such that $\mathbf{w}^{\mathsf{T}}\mathbf{x}_1^* + w_0 = 1$ and $\mathbf{w}^{\mathsf{T}}\mathbf{x}_2^* + w_0 = -1$ in which case the margin is now simply $\frac{2}{||\mathbf{w}||}$
- Maximise margin need to minimise $||\mathbf{w}||$ subject to all the points being correctly classified

SVM

UNIVERSITY *of* GLASGOW

- The SVM optimisation can be written as

$$\min \frac{1}{2}||\mathbf{w}||^2$$

subject to

$$t_n(\mathbf{w}^\mathsf{T}\mathbf{x} + w_0) \geq 1 \;\; \forall \;\; n = 1 \cdots N$$

and by finding the solution to the above we will be using the $\mathbf{w}$ in our classifier which will minimise the bound on the achievable generalisation error.

SVM Optimisation

UNIVERSITY of GLASGOW

- Given a constrained optimisation problem of the form

$$\min f(\mathbf{w})$$

subject to

$$\begin{aligned} g_i(\mathbf{w}) &\leq 0 \quad i = 1 \cdots K \\ h_i(\mathbf{w}) &= 0 \quad i = 1 \cdots M \end{aligned}$$

SVM Optimisation

UNIVERSITY of GLASGOW

- Given a constrained optimisation problem of the form

$$\min f(\mathbf{w})$$

subject to

$$\begin{aligned} g_i(\mathbf{w}) &\leq 0 \quad i = 1 \cdots K \\ h_i(\mathbf{w}) &= 0 \quad i = 1 \cdots M \end{aligned}$$

- Form the Lagrangian function as

$$\mathcal{L}(\mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = f(\mathbf{w}) + \sum_{i=1}^{K} \alpha_i g_i(\mathbf{w}) + \sum_{i=1}^{M} \beta_i h_i(\mathbf{w})$$

SVM Optimisation

UNIVERSITY *of* GLASGOW

- Find maximum of $\mathcal{L}(\mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ with respect to $\mathbf{w}$ denoted as $\theta(\boldsymbol{\alpha}, \boldsymbol{\beta})$

SVM Optimisation

UNIVERSITY of GLASGOW

- Find maximum of $\mathcal{L}(\mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ with respect to $\mathbf{w}$ denoted as $\theta(\boldsymbol{\alpha}, \boldsymbol{\beta})$
- Then solve the optimisation problem

$$\max \theta(\boldsymbol{\alpha}, \boldsymbol{\beta})$$

subject to

$$\alpha_i \geq 0 \quad \forall \;\; i = 1 \cdots K$$

SVM Optimisation

UNIVERSITY *of* GLASGOW

- The Lagrangian function for SVM, noting only one set of inequality constraints and no equality constraints then

$$\mathcal{L}(\mathbf{w}, w_0, \boldsymbol{\alpha}) = \frac{1}{2}||\mathbf{w}||^2 + \sum_{n=1}^{N} \alpha_n \left(1 - t_n(\mathbf{w}^{\mathsf{T}}\mathbf{x}_n + w_0)\right)$$

SVM Optimisation

UNIVERSITY *of* GLASGOW

- The Lagrangian function for SVM, noting only one set of inequality constraints and no equality constraints then

$$\mathcal{L}(\mathbf{w}, w_0, \boldsymbol{\alpha}) = \frac{1}{2}||\mathbf{w}||^2 + \sum_{n=1}^{N} \alpha_n \left(1 - t_n(\mathbf{w}^\mathsf{T}\mathbf{x}_n + w_0)\right)$$

- Have defined each $g_i(\mathbf{w}) = 1 - t_n(\mathbf{w}^\mathsf{T}\mathbf{x}_n + w_0) \leq= 0$ which comes from our original constraint.

SVM Optimisation

- Stationary point of $\mathcal{L}(\mathbf{w}, w_0, \boldsymbol{\alpha})$ so

$$\frac{\partial}{\partial \mathbf{w}} \mathcal{L}(\mathbf{w}, w_0, \boldsymbol{\alpha}) = \mathbf{w} - \sum_{n=1}^{N} \alpha_n t_n \mathbf{x}_n = 0 \Rightarrow \mathbf{w} = \sum_{n=1}^{N} \alpha_n t_n \mathbf{x}_n$$

SVM Optimisation

- Stationary point of $\mathcal{L}(\mathbf{w}, w_0, \boldsymbol{\alpha})$ so

$$\frac{\partial}{\partial \mathbf{w}}\mathcal{L}(\mathbf{w}, w_0, \boldsymbol{\alpha}) = \mathbf{w} - \sum_{n=1}^{N} \alpha_n t_n \mathbf{x}_n = 0 \Rightarrow \mathbf{w} = \sum_{n=1}^{N} \alpha_n t_n \mathbf{x}_n$$

- and

$$\frac{\partial}{\partial w_0}\mathcal{L}(\mathbf{w}, w_0, \boldsymbol{\alpha}) = -\sum_{n=1}^{N} \alpha_n t_n = 0 \Rightarrow \sum_{n=1}^{N} \alpha_n t_n = 0$$

SVM Optimisation

- Stationary point of $\mathcal{L}(\mathbf{w}, w_0, \boldsymbol{\alpha})$ so

$$\frac{\partial}{\partial \mathbf{w}}\mathcal{L}(\mathbf{w}, w_0, \boldsymbol{\alpha}) = \mathbf{w} - \sum_{n=1}^{N} \alpha_n t_n \mathbf{x}_n = 0 \Rightarrow \mathbf{w} = \sum_{n=1}^{N} \alpha_n t_n \mathbf{x}_n$$

- and

$$\frac{\partial}{\partial w_0}\mathcal{L}(\mathbf{w}, w_0, \boldsymbol{\alpha}) = -\sum_{n=1}^{N} \alpha_n t_n = 0 \Rightarrow \sum_{n=1}^{N} \alpha_n t_n = 0$$

- Using above define $\theta(\boldsymbol{\alpha})$ so plug-in results to $\mathcal{L}(\mathbf{w}, w_0, \boldsymbol{\alpha})$.

SVM Optimisation

- Using result $\mathbf{w} = \sum_{n=1}^{N} \alpha_n t_n \mathbf{x}_n$ we should see that

$$
\begin{aligned}
\frac{1}{2}||\mathbf{w}||^2 = \frac{1}{2}\mathbf{w}^\mathsf{T}\mathbf{w} &= \frac{1}{2}\left(\sum_{n=1}^{N} \alpha_n t_n \mathbf{x}_n^\mathsf{T}\right)\left(\sum_{m=1}^{N} \alpha_m t_m \mathbf{x}_m\right) \\
&= \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} \alpha_n \alpha_m t_n t_m \mathbf{x}_n^\mathsf{T}\mathbf{x}_m
\end{aligned}
$$

SVM Optimisation

UNIVERSITY *of* GLASGOW

- Now the second component of our Lagrangian needs to be considered

$$
\begin{aligned}
& \sum_{n=1}^{N} \alpha_n \left(1 - t_n(\mathbf{w}^{\mathsf{T}}\mathbf{x}_n + w_0)\right) \\
= \; & \sum_{n=1}^{N} \alpha_n - \sum_{n=1}^{N} \alpha_n \mathbf{w}^{\mathsf{T}}\mathbf{x}_n - w_0 \sum_{n=1}^{N} \alpha_n t_n \\
= \; & \sum_{n=1}^{N} \alpha_n - \sum_{n=1}^{N} \alpha_n \mathbf{w}^{\mathsf{T}}\mathbf{x}_n \\
= \; & \sum_{n=1}^{N} \alpha_n - \sum_{n=1}^{N} \sum_{m=1}^{N} \alpha_n \alpha_m t_n t_m \mathbf{x}_n^{\mathsf{T}}\mathbf{x}_m
\end{aligned}
$$

SVM Optimisation

- So combining the two parts we obtain

$$\theta(\boldsymbol{\alpha}) = \sum_{n=1}^{N} \alpha_n - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} \alpha_n \alpha_m t_n t_m \mathbf{x}_n^{\mathsf{T}} \mathbf{x}_m$$

SVM Optimisation

- So combining the two parts we obtain

$$\theta(\boldsymbol{\alpha}) = \sum_{n=1}^{N} \alpha_n - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} \alpha_n \alpha_m t_n t_m \mathbf{x}_n^{\mathsf{T}} \mathbf{x}_m$$

- This has to be maximised with respect to all α_n, the constraints that $\alpha_n \geq 0 \quad \forall \quad n = 1 \cdots N$ and the additional constraint which emerges from our stationary conditions that is $\sum_{n=1}^{N} \alpha_n t_n = 0$

SVM Optimisation

- SVM optimisation problem

$$\max \sum_{n=1}^{N} \alpha_n - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} \alpha_n \alpha_m t_n t_m \mathbf{x}_n^{\mathsf{T}} \mathbf{x}_m$$

subject to

$$\begin{aligned} \alpha_n &\geq 0 \quad \forall \; n = 1 \cdots N \\ \sum_{n=1}^{N} \alpha_n t_n &= 0 \end{aligned}$$

SVM Optimisation

UNIVERSITY of GLASGOW

- SVM optimisation problem

$$\max \sum_{n=1}^{N} \alpha_n - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} \alpha_n \alpha_m t_n t_m \mathbf{x}_n^\mathsf{T} \mathbf{x}_m$$

  subject to

$$\begin{aligned} \alpha_n &\geq 0 \;\; \forall \;\; n = 1 \cdots N \\ \sum_{n=1}^{N} \alpha_n t_n &= 0 \end{aligned}$$

- There are a number of ways to solve this problem and we will employ a simple quadratic optimisation solver which is written in Matlab.

Support Vectors

- A significant number of the α_n parameters are returned as having zero value from the optimisation

Support Vectors

UNIVERSITY *of* GLASGOW

- The α_n which have non-zero values are important and as they are associated with each vector in the training sample $\mathbf{x}_n$ these are referred to as the Support Vectors as the *support* the decision boundary between the two classes

Support Vectors

UNIVERSITY of GLASGOW

- Now discriminant function can be written as

$$
\begin{aligned}
f(\mathbf{x}_{new}; \mathbf{w}, w_0) &= \mathrm{sign}(\mathbf{w}^\mathsf{T}\mathbf{x}_{new} + w_0) \\
&= \mathrm{sign}\left(\sum_{n=1}^{N} t_n \alpha_n \mathbf{x}_n^\mathsf{T}\mathbf{x}_{new} + w_0\right) \\
&= \mathrm{sign}\left(\sum_{n \in SV} t_n \alpha_n \mathbf{x}_n^\mathsf{T}\mathbf{x}_{new} + w_0\right) \\
&= \mathrm{sign}\left(\sum_{n \in SV} t_n \alpha_n K(\mathbf{x}_n, \mathbf{x}_{new}) + w_0\right)
\end{aligned}
$$

Support Vectors

UNIVERSITY of GLASGOW

Figure (2) shows the SVM decision plane and the support vectors for this little toy data set.

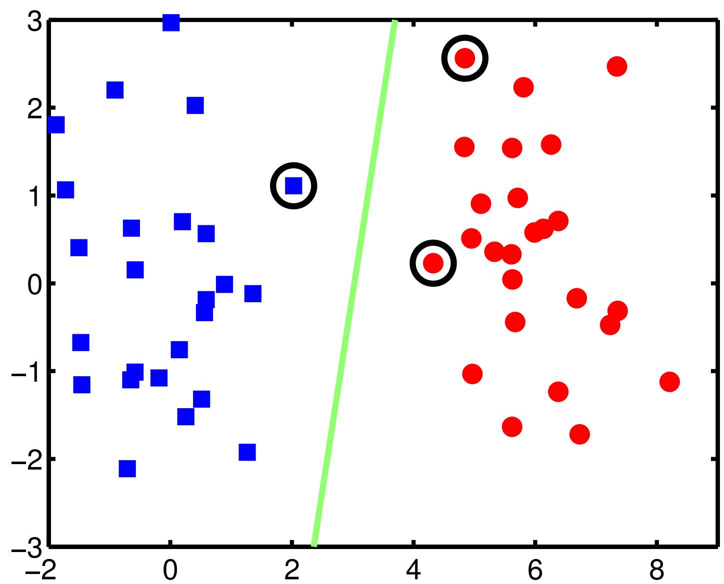


Figure 2: The SVM decision plane separating examples from two classes along with the support vectors which are highlighted. Note that there are only three non-zero α components and so only three points in the data set which are supporting the decision surface.

SVM Tuning Params

UNIVERSITY *of* GLASGOW

- For the case where the samples from the two classes may not be completely linearly separable then the SVM optimisation problem can be posed in such a way as to take these possible errors into account.

SVM Tuning Params

UNIVERSITY *of* GLASGOW

- For the case where the samples from the two classes may not be completely linearly separable then the SVM optimisation problem can be posed in such a way as to take these possible errors into account.
- It turns out that a very simple change to the SVM optimisation is required and it changes the positivity constraint from $\alpha_n \geq 0$ to $0 \leq \alpha_n \leq C$, for all n, where C is a box constraint parameter

SVM Tuning Params

UNIVERSITY *of* GLASGOW

- For the case where the samples from the two classes may not be completely linearly separable then the SVM optimisation problem can be posed in such a way as to take these possible errors into account.
- It turns out that a very simple change to the SVM optimisation is required and it changes the positivity constraint from $\alpha_n \geq 0$ to $0 \leq \alpha_n \leq C$, for all n, where C is a box constraint parameter
- Hyper-parameters consist of C and kernel parameters if any e.g. β from RBF kernel - LOO needed

SVM Tuning Params

UNIVERSITY
of
GLASGOW

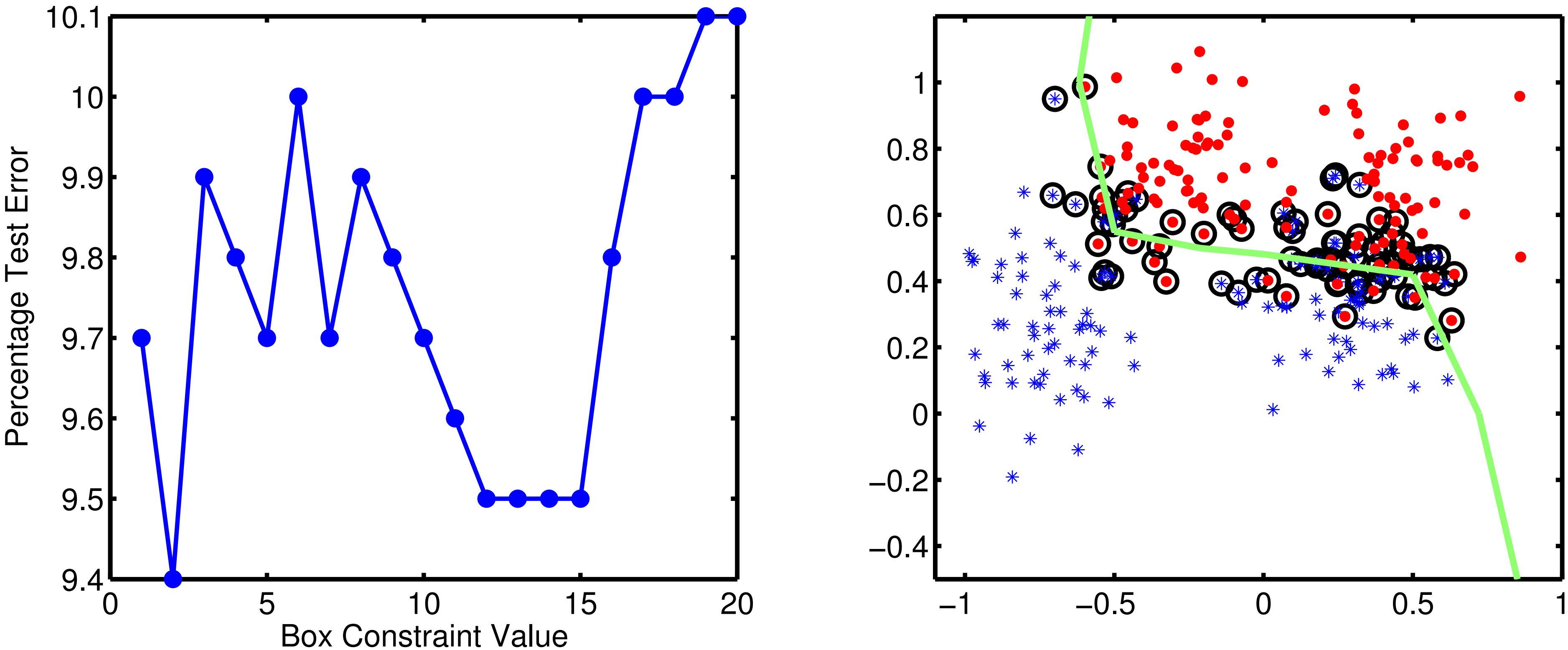


Figure 3: The left hand plot shows the test error achieved for varying values of C when using a polynomial order kernel function. The right hand plot shows the training data and the decision surface. The support vectors are highlighted and they can all be seen to be clumped around the decision surface.

SVM Tuning Params

UNIVERSITY of GLASGOW

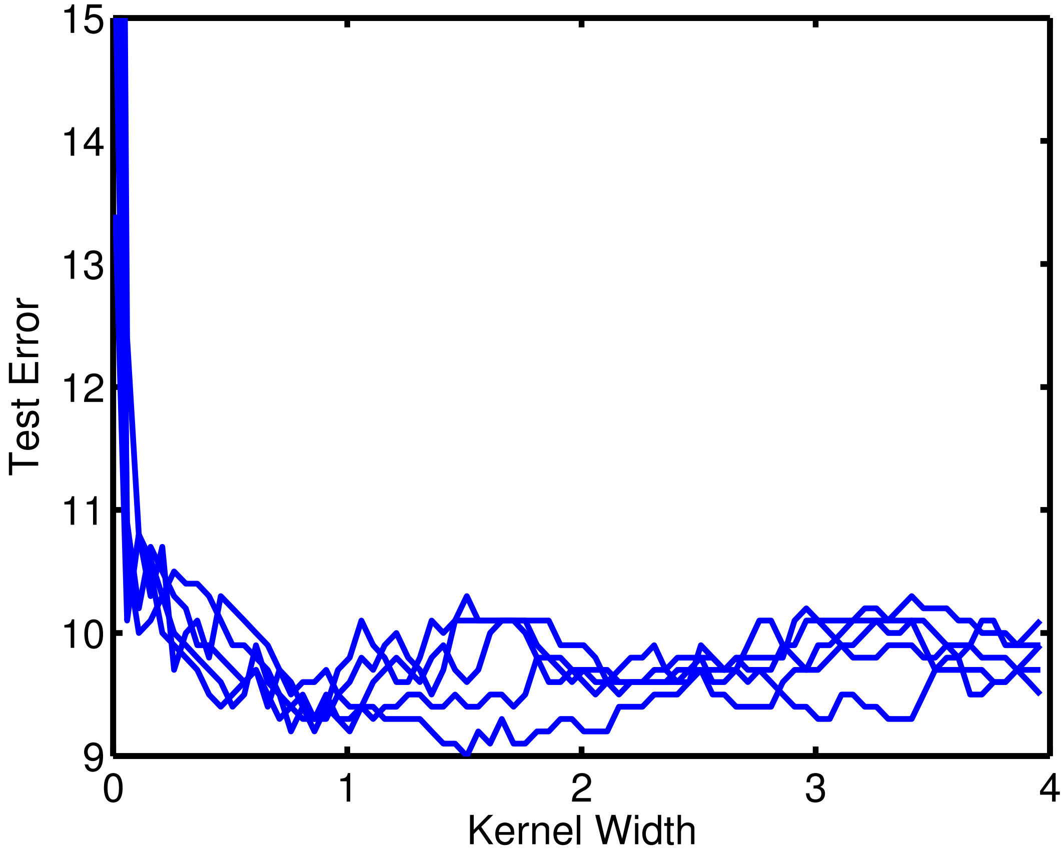


Figure 4: The percentage error achieved by an SVM using a Radial Basis Kernel function with a width parameter ranging from 0.01 to 4.0 in step sizes of 0.05. For each of these ranges a value of C was selected from 1 to 4 and we can see that the minimum test error of 9.0% was achieved with hyper-parameter values of $C = 1$ and $\beta = 1.4$.