Chapter 8
Classification
Support Vector Machines

Statistical Methods and Learning in Computer Vision
WS 2012/13

Claudia Nieuwenhuis
Lehrstuhl für Computer Vision and Pattern Recognition
Fakultät für Informatik
Technische Universität München

Overview

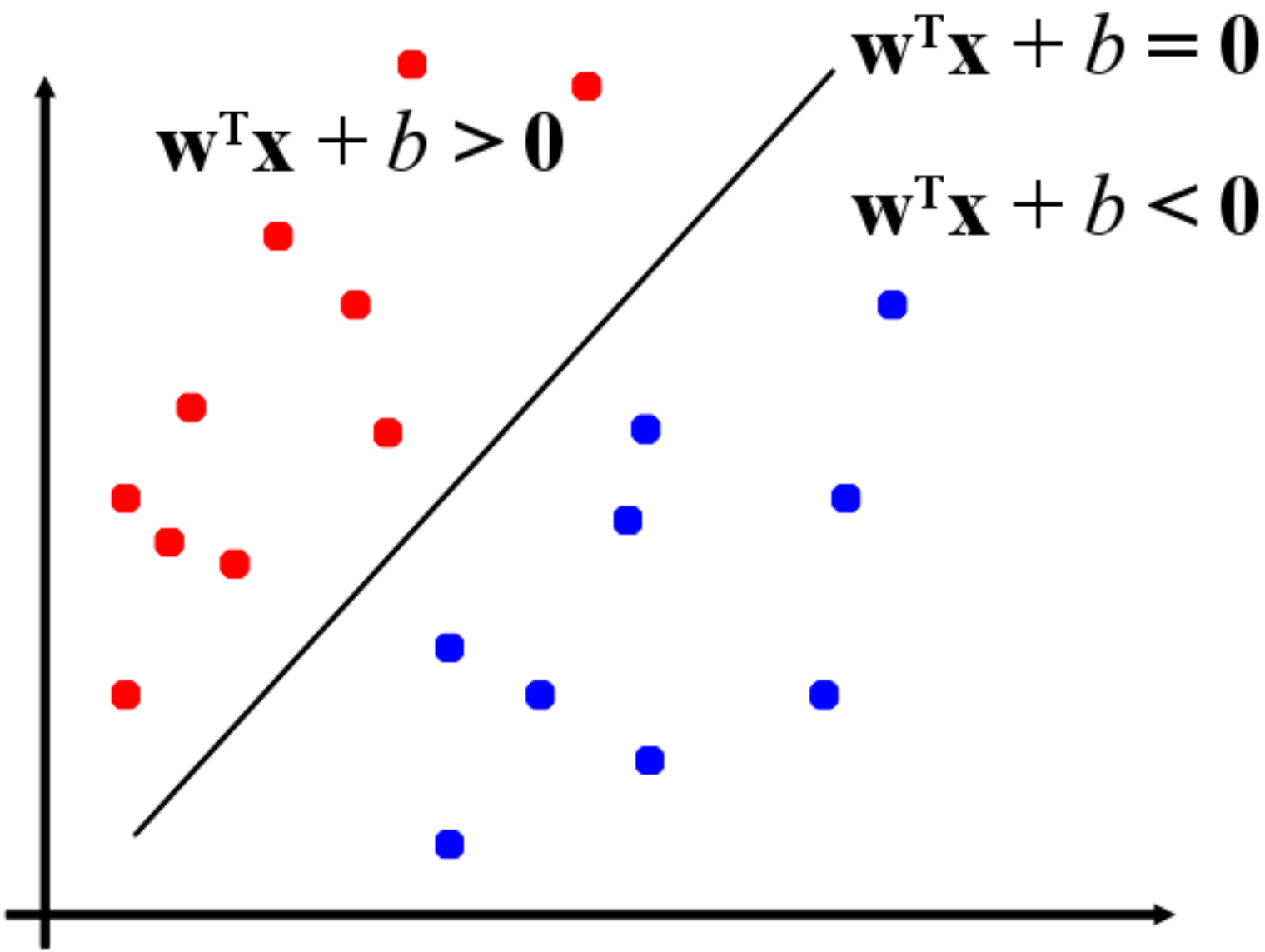
Support Vector Machines

We want to classify unknown data points given a training dataset $(x_1, ..., x_N)$ with indicated classes $(t_1, ..., t_N) \in \{-1, 1\}$. Support Vector Machines define hyperplanes, which separate the data points. Such a hyperplane is given by

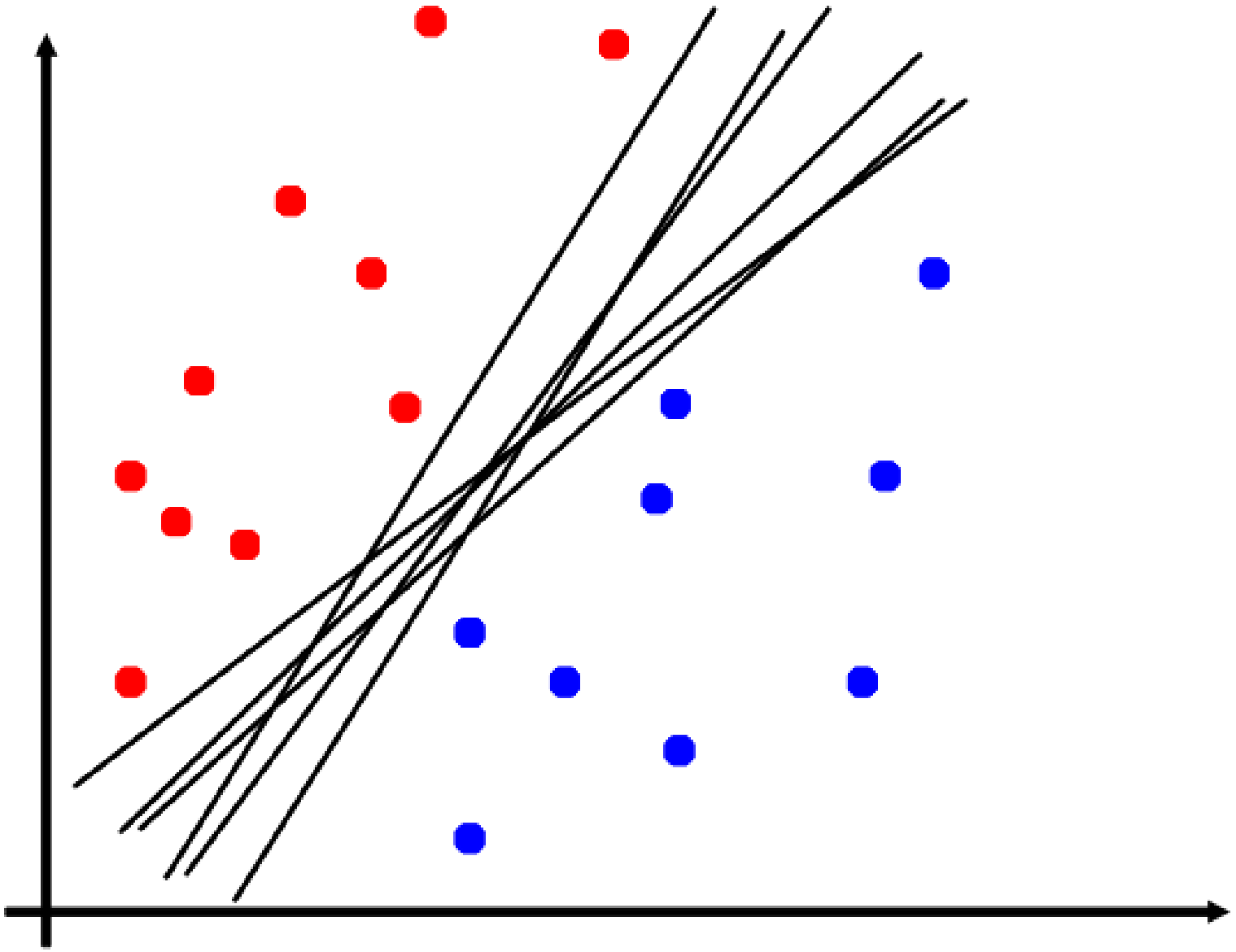
$$y(x) = w^T x + b$$

We want to find the 'optimal' hyperplane for class separation.

Support Vector Machines



Support Vector Machines



Support Vector Machines

Optimal hyperplane means that the margin (the perpendicular distance between the decision boundary and the closest point of the data points) is maximized.

Distance d between point x_n and plane defined by $y(x_n)$:

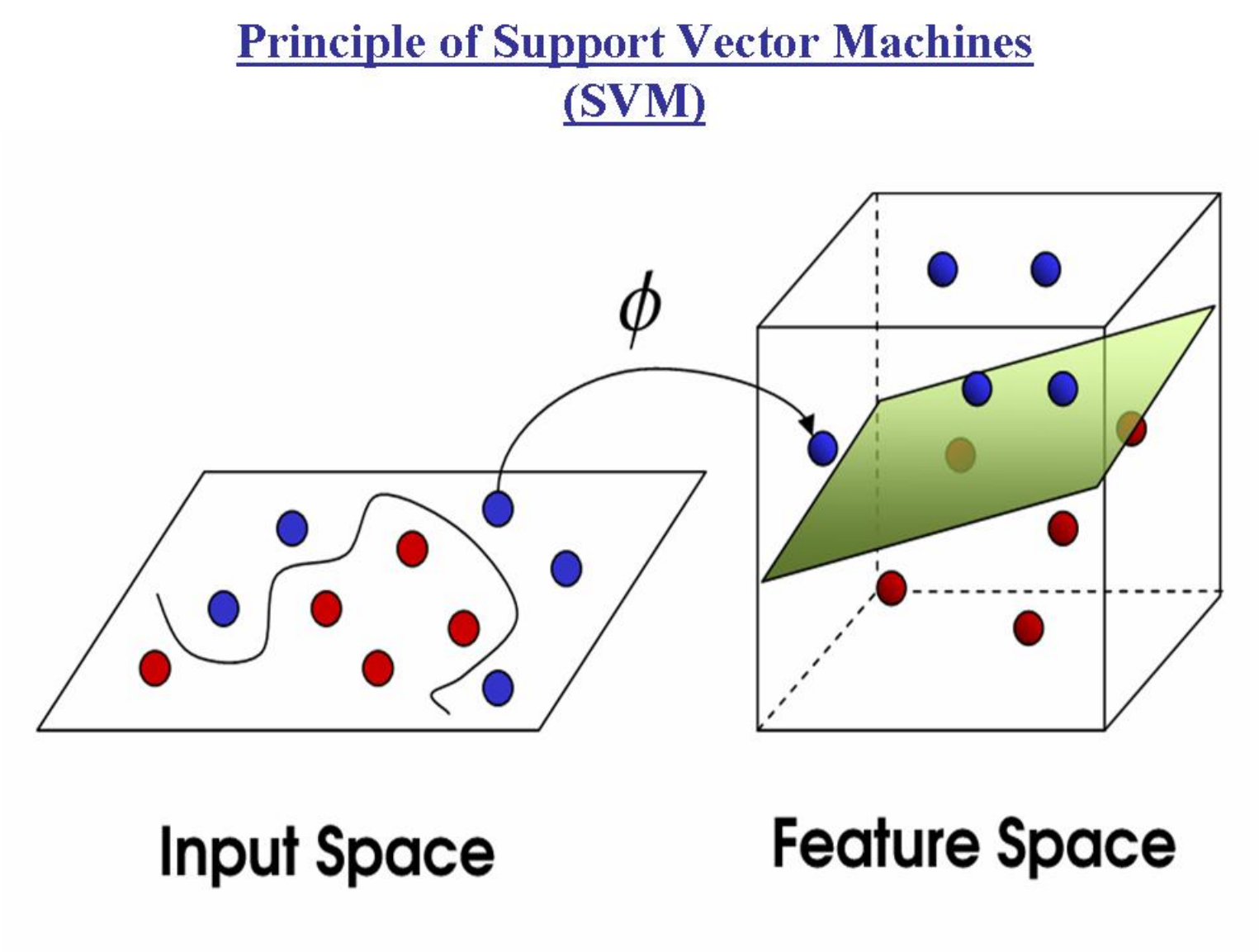
$$y(x) = w^T x + b \Rightarrow d_n = \frac{t_n y(x_n)}{\|w\|}$$

To maximize the margin, we solve the optimization problem

$$\underset{w,b}{\text{argmax}} \left\{ \frac{1}{\|w\|} \min_n [t_n(w^T \phi(x_n) + b)] \right\}$$

where ϕ means a kernel function which allows for non-linear decision boundaries.

Kernel Functions



Optimization Problem

Note that any rescaling κw and κb does not change the distance of any point from the hyperplane. Hence, the hyperplane is not uniquely defined. This degree of freedom can be used to set

$$\min_n t_n(w^T \phi(x_n) + b) = 1 \tag{1}$$

for the point closest to the hyperplane. It follows for all points that

$$t_n(w^T \phi(x_n) + b) \geq 1 \tag{2}$$

By definition, there will always be at least one active constraint, because there will always be a closest point. The optimization problem then simply requires that we maximize $\|w\|^{-1}$, which is equivalent to minimizing $\|w\|^2$. This leads to the optimization problem

$$\underset{w,b}{\operatorname{argmin}} \frac{1}{2}\|w\|^2$$

subject to the constraints (2) above.

Optimization Problem

To solve the constrained optimization problem, we use Lagrange multipliers to incorporate the inequality constraints and obtain the Karush-Kuhn-Tucker conditions

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

$$t_n(w^T\phi(x_n) + b) - 1 \geq 0$$

$$a_n \geq 0$$

$$a_n\{t_n(w^T\phi(x_n) + b) - 1\} = 0 \quad (3)$$

We minimize over w and b.

Optimization Problem

Taking the derivative with respect to b and w and eliminating b and w from $L(w, b, a)$ leads to the following maximization problem in a

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

$$a_n \geq 0$$

$$\sum_{n=1}^{N} a_n t_n = 0$$

where $k(x_n, x_m) := \phi(x_n)^T \phi(x_m)$. This is a quadratic programming problem, which can be solved for a.

After computing a we can compute w from

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

From the complementary slackness condition

$$a_n\{t_n(w^T\phi(x_n) + b) - 1\} = 0$$

we see that either $a_n = 0$ or the inequality constraint is exactly fulfilled. Only the set of points S, for which the constraint is exactly fulfilled and $a_n > 0$, play a role in the classification of points by means of

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

If $y(x)$ is negative x belongs to one class, otherwise to the other.

Support Vectors

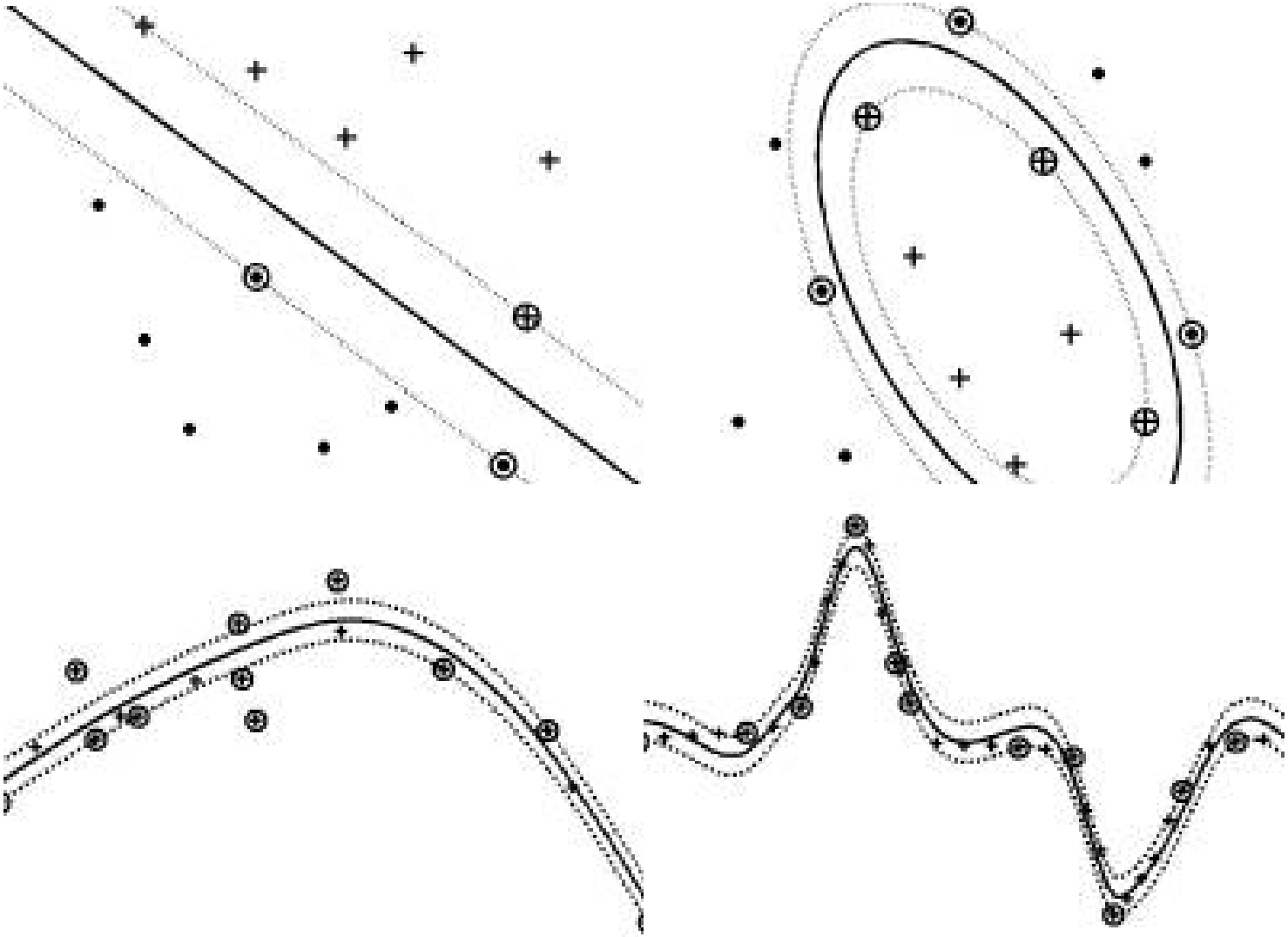
These points in the set S are the so-called support vectors, which lie directly on the margin boundary. For these points we have (see (1))

$$t_n y(x_n) = t_n(w^T \phi(x_n) + b) = 1.$$

b can be computed from one of these points. To make the computation more stable we multiply by t_n, use $t_n^2 = 1$ and use averaging

$$b = \frac{1}{N_S} \sum_{n \in S} \left(t_n - \sum_{m \in S} a_m t_m k(x_n, x_m) \right)$$

Kernel Functions

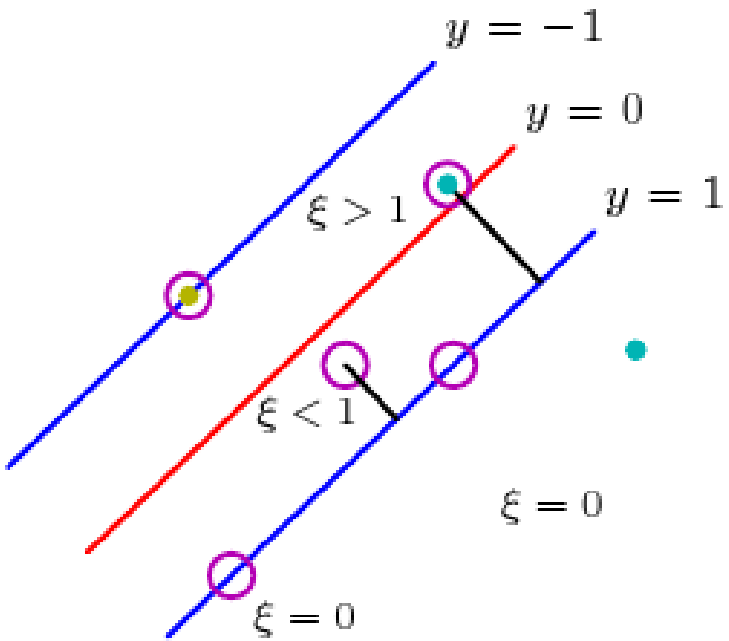


Overlapping Class Distributions

So far, we have assumed that the classes are linearly separable in the feature space. This may not be the case. We introduce variables ξ_n, which allow for classification errors.

$$\xi_n = \max\{|t_n - y(x_n)|, 0\}$$

$$\xi_n \begin{cases} = 0 & x_n \text{ lies on or outside the correct margin boundary} \\ < 1 & x_n \text{ lies within the margin but is classified correctly} \\ = 1 & x_n \text{ lies on the class separating plane} \\ > 1 & x_n \text{ is classified incorrectly} \end{cases}$$



Overlapping Classes

Instead of the original constraints (2)

$$t_n(w^T\phi(x_n) + b) \geq 1$$

we obtain the relaxed constraints

$$t_n(w^T\phi(x_n) + b) \geq 1 - \xi_n,\ \xi_n \geq 0 \tag{4}$$

We want to maximize the margin and softly penalize points lying on the wrong side of the margin boundary. This leads to the new target function

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

which is to be minimized subject to the above constraints (4). C can be understood as a regularization parameter, which determines the influence of misclassified points.

Optimization Problem

The corresponding Lagrangian is given by

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

$$\begin{aligned}
&t_n y(x_n) - 1 + \xi_n \geq 0, \quad \xi_n \geq 0 \text{ ("primal feasibility")} \\
&a_n \geq 0, \quad \mu_n \geq 0 \text{ ("dual feasibility")} \\
&a_n(t_n y(x_n) - 1 + \xi_n) = 0, \quad \mu_n\xi_n = 0 \text{ ("comp. slackness")}
\end{aligned} \tag{5}$$

Optimization Problem

Eliminating w, b and ξ leads to the same optimization problem as before with slightly different constraints

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

$$0 \leq a_n \leq C$$

$$\sum_{n=1}^{N} a_n t_n = 0$$

Support Vectors

For classification we have again

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

As before, only the support vectors with $a_n > 0$ play an important role for the classification of new points x.

After solving the optimization problem for the parameters a, we obtain the same hyperplane parameter equations as before - but computed from different values for a_n.

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

$$b = \frac{1}{N_S} \sum_{n \in S} \left(t_n - \sum_{m \in S} a_m t_m k(x_n, x_m) \right)$$