

We briefly describe linear and non-linear SVMs. Let  $\vec{x}_i$  denote the feature vector, and let  $y_i$  denote its class label (e.g.,  $y_i = +1$  if  $\vec{x}_i$  corresponds to a clean audio, and  $y_i = -1$  if  $\vec{x}_i$  corresponds to a stega audio). In a linear SVM, we seek a linear decision function determined by a unit vector  $\vec{w}$  and an offset  $b$  as:

$$f(\vec{x}) = \text{sgn}(\vec{w}^T \vec{x} - b), \quad (1)$$

such that after projecting a data point,  $\vec{x}$ , onto  $\vec{w}$ , a positive labeled data will have an output  $+1$  while a negative labeled data will have an output  $-1$ . The decision function,  $f(\cdot)$ , is estimated by maximizing  $\gamma$  subject to the following constraints;

$$\begin{aligned} \vec{w}^T \vec{x}_i - b &\geq \gamma, & \text{if } y_i = +1 \\ \vec{w}^T \vec{x}_i - b &\leq -\gamma, & \text{if } y_i = -1 \\ \|\vec{w}\| &= 1, \end{aligned} \quad (2)$$

where  $\gamma$  is the classification margin (dotted line in Figure 1(a)). The margin is the distance that the classification surface can translate without touching any data point. These constraints force all of the data to be outside the margin region, and constrains  $\vec{w}$  to be a unit vector. This optimization problem can be further transformed into a constrained convex quadratic programming problem, whose solution is found by efficient iterative algorithms.

In the case where the data do not afford any linear separation, the optimization problem is adjusted to tolerate some classification errors. Specifically, slack variable  $\xi_i$  is introduced for each data point  $\vec{x}_i$  to indicate its violation from a linear separation. The constraints are changed accordingly to:

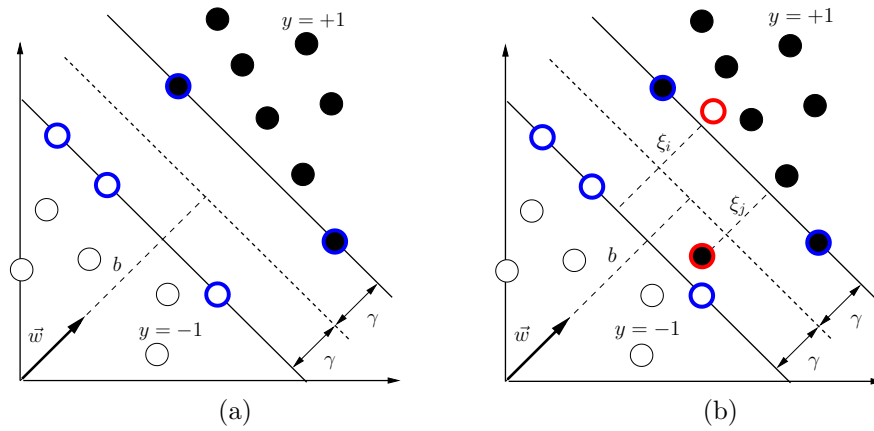
$$\begin{aligned} \vec{w}^T \vec{x}_i - b &\geq \gamma - \xi_i, & \text{if } y_i = +1 \\ \vec{w}^T \vec{x}_i - b &\leq -\gamma + \xi_i, & \text{if } y_i = -1 \\ \|\vec{w}\| &= 1, & \xi_i \geq 0. \end{aligned} \quad (3)$$

The overall classification error is thus measured by the sum of such slack variables. The objective function is changed to reflect the compromise between minimizing the classification error and maximizing the classification margin as:

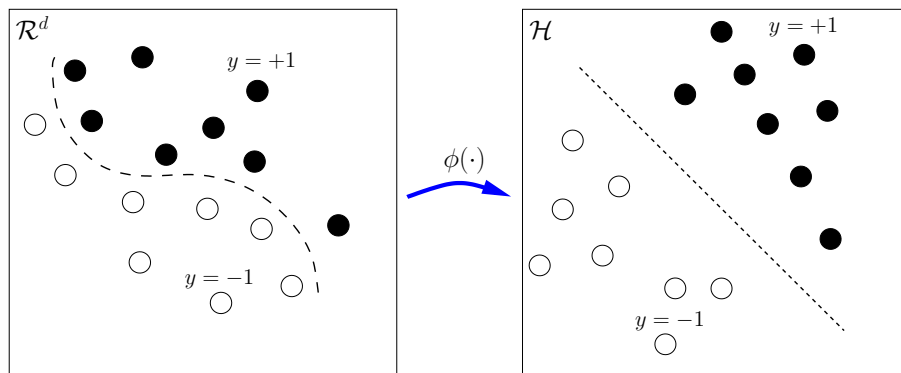
$$\max \quad \gamma - C \sum_{i=1}^N \xi_i, \quad (4)$$

where  $C > 0$  is the penalty on classification errors.

Linear SVM can also be performed in a nonlinearly mapped space to achieve nonlinear separation of data, Figure 2. First, all data points are mapped into another space  $\mathcal{H}$  where they afford a linear separation by a nonlinear map  $\phi$ . Linear SVM algorithm is run in  $\mathcal{H}$  to find a linear decision function as Equation (??). Such a linear decision function corresponds in the original space as a nonlinear classification. There is one further step in nonlinear SVM where a kernel function computing inner products of two mapped data points in  $\mathcal{H}$  is introduced. Such a kernel function is plugged into the optimization algorithm and solves the problem more efficiently.



**Figure 1.** Linear SVM classification in a 2D toy example for (a) linearly separable data and (b) linearly nonseparable data.



**Figure 2.** Non-linear SVM as linear SVM performed in a non-linearly mapped space.