0.1 Support Vector Machine

- *Algorithm:* Support Vector Machine (algo. 1)
- *Input:* set of (input, output) training pair samples, where input has n features such that $x_1, x_2, ..., x_n$ and output as y.
- *Complexity:* $\mathcal{O}(n^2)$ for linear, $\mathcal{O}(n^2) - \mathcal{O}(n^3)$ for nonlinear.
- *Data structure compatibility:* N/A
- *Common applications:* Machine Learning, for data classification,

Problem. Support Vector Machine

Given training vectors $x_i \in \mathbb{R}^p$, i=1,. . . , n, in two classes, and a vector $y \in \{1, -1\}^n$, our goal is to find $w \in \mathbb{R}^p$ and $b \in \mathbb{R}$ such that the prediction given by sign$(w^T \phi(x) + b)$ is correct for most samples. .

Description

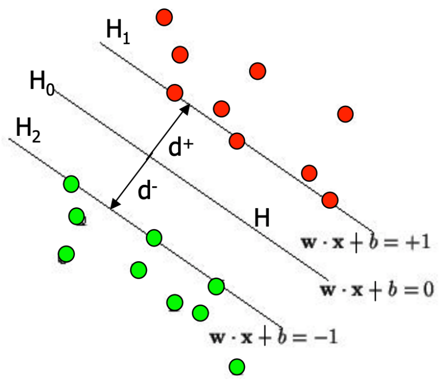
Basic idea: The objective of support vector machine algorithm is to find a hyperplane in an N-dimensional space (N is the number of features). Then this optimal hyperplane could be used for linearly separable patterns (distinctly classifies the data points).

More specific, the hyperplanes is the decision boundaries used to classify the data points. For example, if the number of input features is 2, then a *line* is the hyperplane in a 2-dimensional space, which is binary classification. So with the **input** as a set of (input, output) training pair samples, where input has n features such that $x_1, x_2, ..., x_n$ and output as y, the **output** is a set of weights $\mathbf{w}$ $(w_1, w_2, ..., w_n)$, one for each feature. Then the linear combination of those weights predicts the value of y.

Important Concepts:

- **Support Vector:** Data points that are close to the hyperplane, influence the position and orientation of the hyperplane, which could be used to maximize the margin of the classifier [**ts**]. This is a key difference between SVM and neural nets. Because there are finite sets of weights $\mathbf{w}$ that we could choose from, all the data points will influence the optimality for neural nets, while SVM only consider the support vectors (so other points will not affect the boundary).
- **Hyperplane** H defined as [**slides**]

$$w \cdot x_i + b \geq +1 \text{ when } \mathrm{y_i} = +1$$
$$w \cdot x_i + b \leq -1 \text{ when } \mathrm{y_i} = -1$$
$$H_1 : w \cdot x_i + b = +1$$
$$H_2 : w \cdot x_i + b = -1$$
$$H_0 : w \cdot x_i + b = 0$$



$d+/d-$: the shortest distance to the closet positive/negative point. And no points between H_1 and H_2. Points on plane H_1 and H_2 are support vectors.

- **Optimal Hyperplane:** H with the maximum margins. Notice the distance between H_0, H_1 is $2/||w||$, so need to minimize $||w||$

Minimize $||w||$ and no points between H_1, H_2 together defined a constrained optimization problem, which could be solved by the Lagrangian multiplier method. Also we could solve it by just computing the inner products of x_i, y_i [**slides**].

We minimize $||w||$ while add penalty when sample misclassified (penalty strength controlled by C). Allow some distance ζ_i between points and correct boundary. $\min_{w,b,\zeta} \frac{1}{2} w^T w + C \sum_{i=1}^{n} \zeta_i$ subject to $y_i(w^T \phi(x_i) + b) \geq 1 - \zeta_i, \zeta_i \geq 0, i = 1, ..., n$ [**scikit**]. The function finally obtained to optimize is $L_d = \sum a_i - 1/2 \sum a_i a_j y_i y_j (x_i \cdot x_j)$.

Non-linear SVMs: As we compute $(x_i \cdot x_j)$, if transform the vector to get a linear situation, need to compute $\phi(x_i), \phi(x_j)$, which will be time consuming. However, with **kernel function** K such $K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j)$, no need to know or compute ϕ. The kernel function defines inner products in the transformed space [**slides**]. The function finally obtained to optimize is $L_d = \sum a_i - 1/2 \sum a_i a_j y_i y_j K(x_i, x_j)$. Generally speaking, for a nonlinear SVM is between $\mathcal{O}(n^2)$ (small C) and $\mathcal{O}(n^3)$ (large C) with n amounts of training instances [**paper1**].

Complexity:

The core of an SVM is the separation between support vectors and the rest of the data. The training time complexity depends on the number of samples, type of the kernel function and the penalization C (regularization parameter). For linear case, generally could expect training time of $\mathcal{O}(n^2)$. It is highly depends on how we solve the quadratic problem and choose the support vectors. The training time for a linear SVM to reach a certain level of generalization error actually decreases as training set size increases

---

Algorithm 1: Training an SVM

Input : $\mathbf{X}$ and y loaded with training labeled data, $\alpha \leftarrow 0$ or partially trained SVM
Output: trained SVM

$C \leftarrow$ value specified by the case
while *changes in* α *and other resource constraint not met* **do**
    for $\forall$ $(x_i, y_i), (x_j, y_j)$ **do**
        Optimize α_i and α_j
    end for
end while

To demonstrate, consider a simple two dimensional case

This simple binary classification problem has two informative features (data points generated by *Python* $make_classification$). So the two classes are separable by a linear classifier. yellow and black points represent different class of points. It illustrates how C parameter performs regularization. Large C represent less regularization and will fit the training set with as few errors as possible, even if it means using a small immersion decision boundary. Very small values of C on the other hand use more regularization that encourages the classifier to find a large marge on decision boundary, even if that decision boundary leads to more points being misclassified.

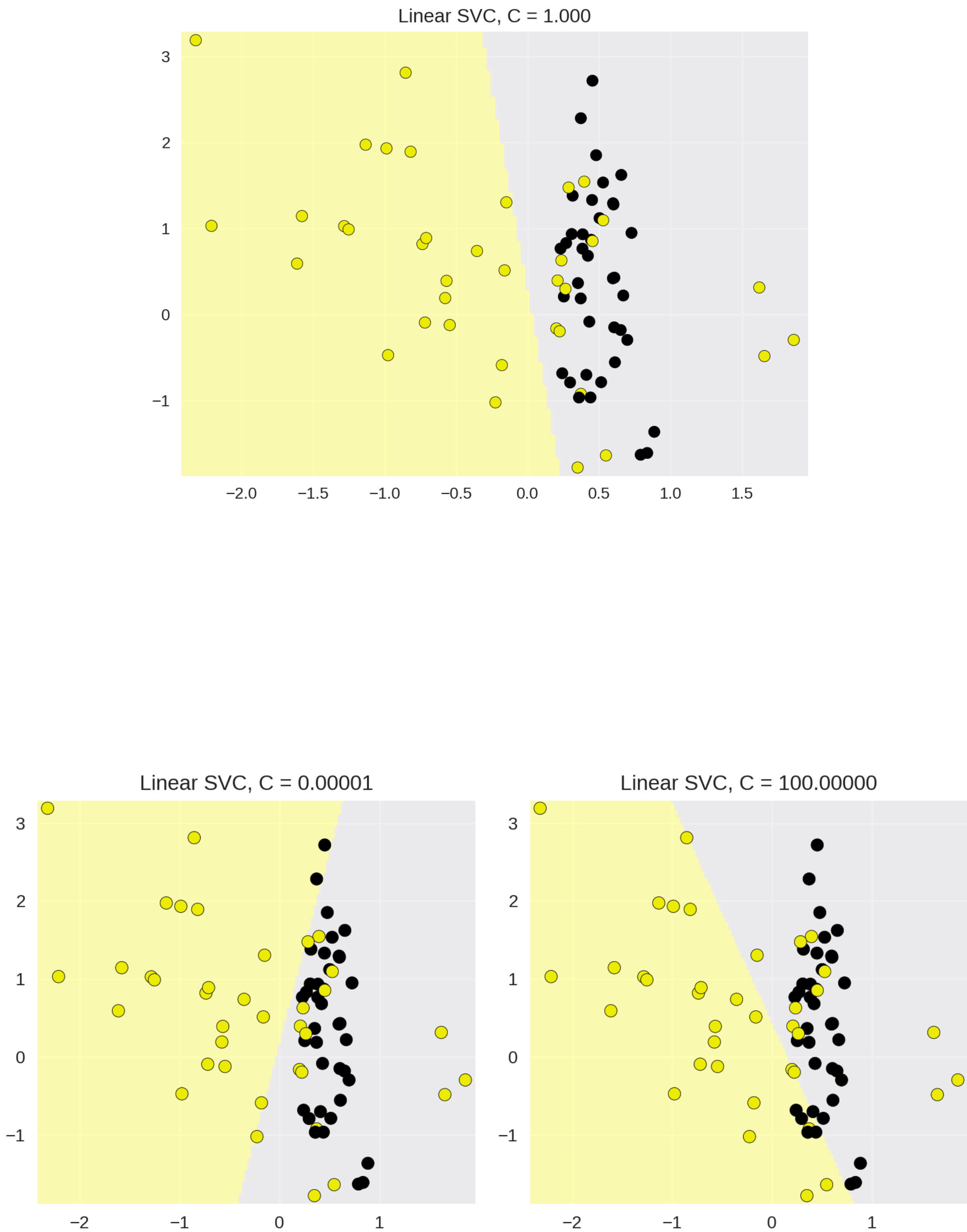


Figure 1: C parameter affect SVM