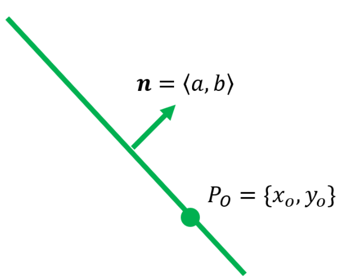
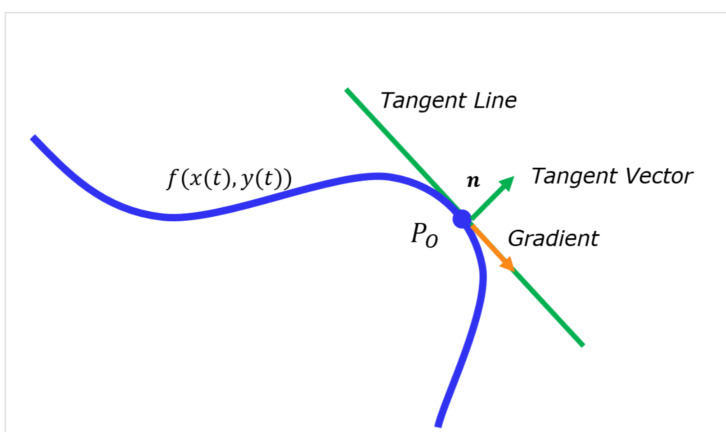
Recitation 02/21/20: Support Vector Machines

1. Geometric Interpretation of SVMs
2. Solving for $\mathbf{w}$
3. Slack Variables
4. Dual Formation
5. Kernels

Clarification on Lagrange Multiplier Derivation



$n \cdot P_o = 0$
$a(x - x_o) + b(y - y_o) = 0$

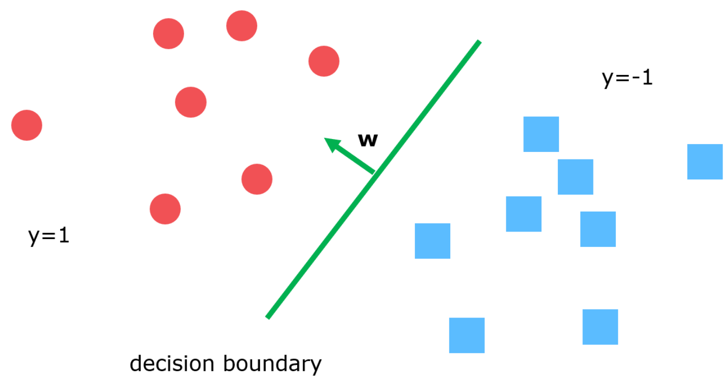


Geometric Interpretation of SVMs

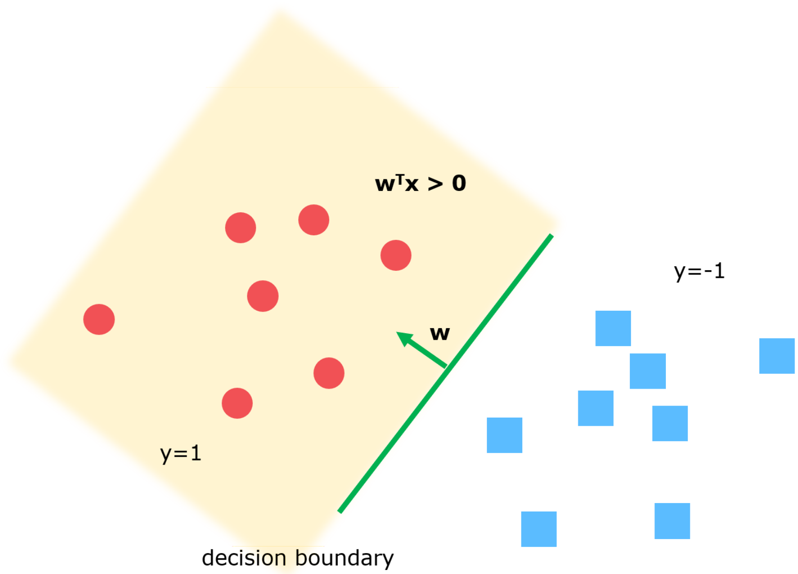
Predicting labels from a binary, linear classifier:

$$\hat{y}_i = sign(\mathbf{w}^T \mathbf{x_i})$$

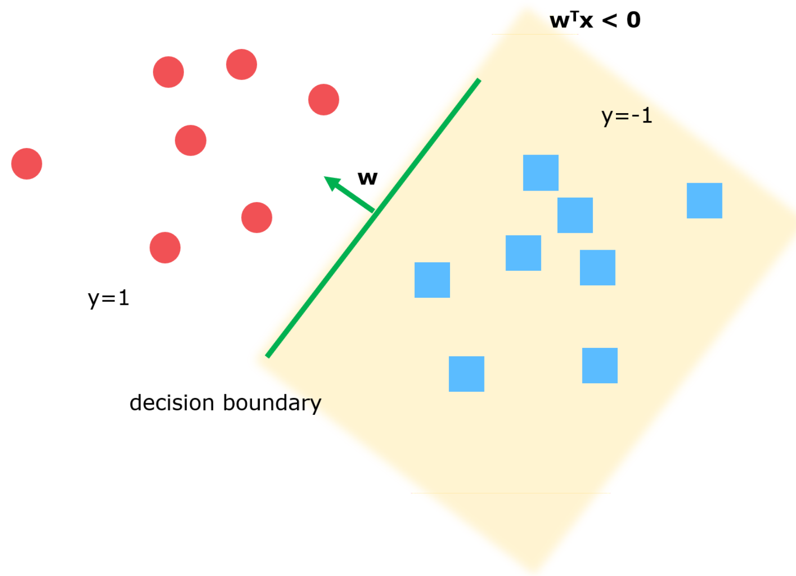
where $y \in \{-1, 1\}$.



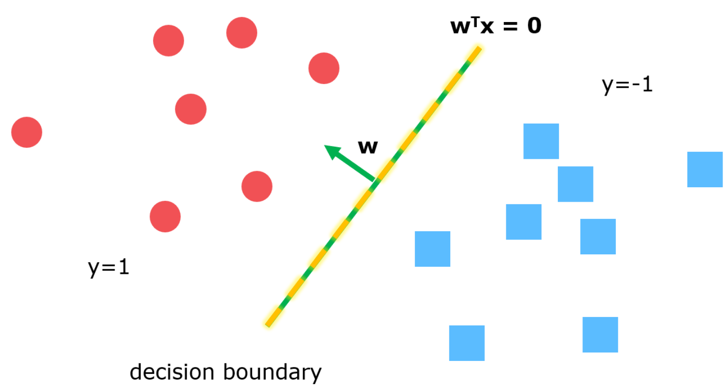
Positive Examples:



Negative Examples:



On the decision boundary:

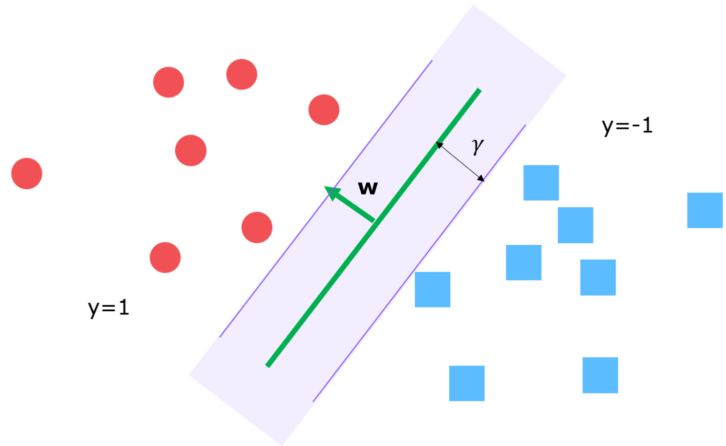


If we classify all the data points correctly:

$$y_i \mathbf{w}^T \mathbf{x_i} > 0, \forall i$$

Computing the Margin

The margin is the distance between the



$$\gamma_i = y_i \mathbf{w}^T \mathbf{x_i} \forall i$$

$$\gamma = min(\gamma_i) \forall i$$

If we classify all the data points correctly:

$$y_i \mathbf{w}^T \mathbf{x_i} > 0, \forall i$$

$$\gamma_i > 0, \forall i$$

The Primal SVM Formation

We want to find the decision boundary that maximizes the margin, while still correctly classifying all samples:

$$\max_{\mathbf{w}} \gamma \quad \text{s.t. } y_i \mathbf{w}^T \mathbf{x_i} \geq \gamma \quad \forall i$$

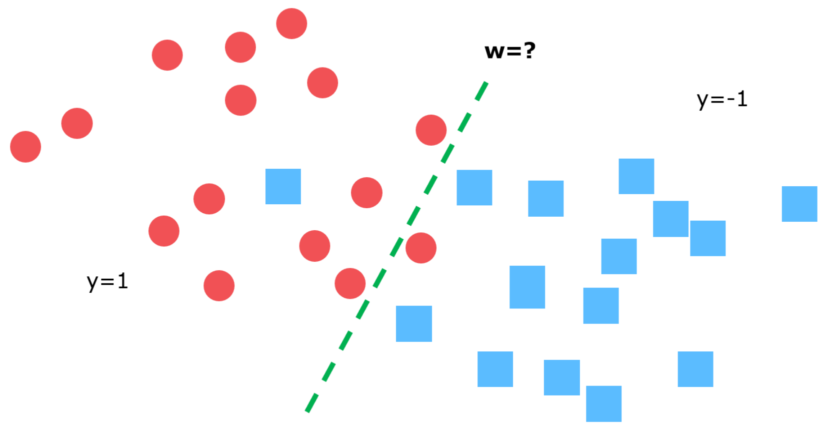
We can continue to increase γ by scaling $\mathbf{w}$. We will equivalently rewrite this as: find the smallest $\mathbf{w}$ that produces a margin of 1:

$$\min_{\mathbf{w}} \mathbf{w}^T \mathbf{w} \quad \text{s.t. } y_i \mathbf{w}^T \mathbf{x_i} \geq 1 \quad \forall i$$

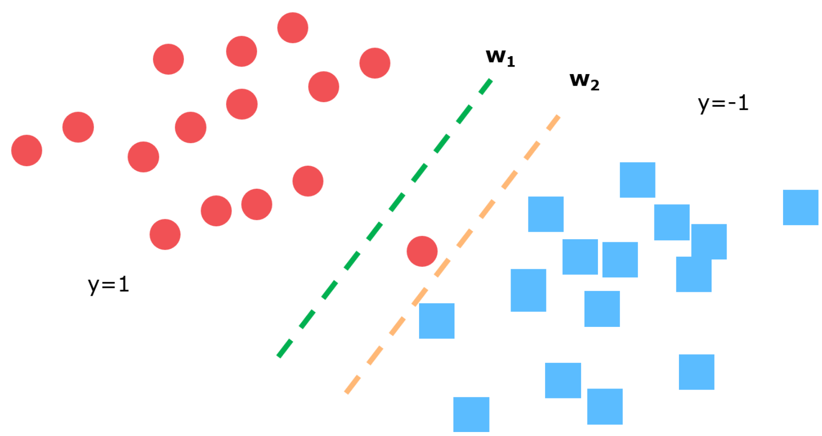
This is a quadratic minimzation problem with a linear constraint. Solvers exist to find the optimal solution.

Slack Variables

What happens if our data is not linearly separable? *We cannot find a solution that satisfies our constraint!*



What happens if our data is linearly separable but "the best" classifier would sacrifice a small number of misclassifications to increase the margin for all other data points?

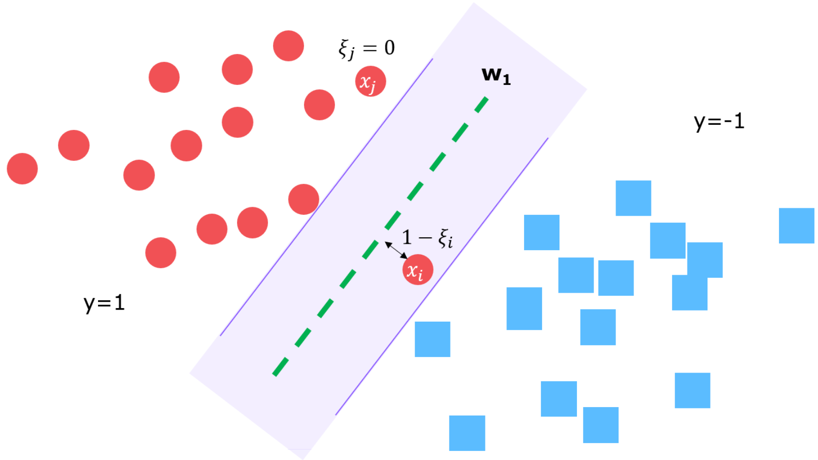


We address these two limitations of SVMs by adding **slack variables**. Slack variables allow some data points to be misclassified with a penalty. They find an optimal decision boundary with the minimal slack penalty:

$$\min_{\mathbf{w}} \mathbf{w}^T\mathbf{w} + \lambda \sum_{i=1}^{N} \xi_i$$

$$\text{s.t. } y_i \mathbf{w}^T \mathbf{x_i} + \xi_i \geq 1 \quad \forall i, \quad \xi_i \geq 0 \quad \forall i$$

What kind of regularization are we using on the ξ_i? This encourages ξ_i's to be ***sparse!***



Lagrange Multipliers

Recall, Lagrange Multipliers solve for the optimal value of a function $f(x, y)$ subject to the constraint $g(x, y) = c$. At the optimal value, $\mathbf{w_o}$:

$$\nabla f(\mathbf{w_o}) = \lambda \nabla g(\mathbf{w_o})$$

Using this we write the Langrangian:

$$\mathcal{L}(\mathbf{w}) = f(\mathbf{w}) + \lambda g(\mathbf{w})$$

Note that when we take the gradient and solve for zero we will solve:

$$\nabla \mathcal{L}(\mathbf{w}) = \nabla f(\mathbf{w}) + \lambda \nabla g(\mathbf{w}) = 0$$

If we solve the above equation for $\mathbf{w}$ we find:

$$\nabla f(\mathbf{w_o}) = \lambda \nabla g(\mathbf{w_o})$$

Let's formulate our SVM optimazation so we can solve it using Lagrange Multipliers:

$$f : \mathbf{w}^T \mathbf{w}$$

$$g : y_i \mathbf{w}^T \mathbf{x_i} - 1 \geq 0, \quad \forall i$$

In this formulation, we write the constant λ as $\alpha \in \mathbb{R}^N$. The Lagrangian can be written as

$$\mathcal{L}(\mathbf{w}, \alpha) = f(\mathbf{w}) + \sum_{i=1}^{N} \alpha_i g(y_i, \mathbf{w}, \mathbf{x_i})$$

$$\mathcal{L}(\mathbf{w}, \alpha) = \mathbf{w}^T \mathbf{w} - \sum_{i=1}^{N} \alpha_i [y_i \mathbf{w}^T \mathbf{x_i} - 1]$$

$$\nabla \mathcal{L}(\mathbf{w}, \alpha) = \mathbf{w} - \sum_{i=1}^{N} \alpha_i y_i \mathbf{x_i}$$

$$\mathbf{w} = \sum_{i=1}^{N} \alpha_i y_i \mathbf{x_i}$$

Let's rewrite $\mathcal{L}(\mathbf{w}, \alpha)$ as $\mathcal{L}(\alpha)$:

$$\mathcal{L}(\alpha) = (\sum_{i=1}^{N} \alpha_i y_i \mathbf{x_i})^T (\sum_{i=1}^{N} \alpha_i y_i \mathbf{x_i}) - \sum_{i=1}^{N} \alpha_i [y_i (\sum_{j=1}^{N} \alpha_j y_j \mathbf{x_j})^T \mathbf{x_i} - 1]$$

$$\mathcal{L}(\alpha) = \sum_{i=1}^{N} \alpha_i - \sum_{i=1}^{N} \alpha_i [y_i (\sum_{j=1}^{N} \alpha_j y_j \mathbf{x_j})^T \mathbf{x_i} - 1]$$

$$\text{s.t.} \alpha_i \geq 0, \quad \forall i$$

$$\text{s.t.} \sum_{i=1}^{N} \alpha_i y_i = 0, \quad \forall i$$

Prediction in the Dual:

$$\hat{y}_i = \mathbf{w}^T \mathbf{x_i}$$

$$\hat{y}_i = \sum_{j=1}^{N} \alpha_j y_j \mathbf{x_j}^T \mathbf{x_i}$$

Dual Formation

- Why solve the dual?
- How does solving the dual relate to overfitting?

Kernels

- Can you use kernels in the primal formation?
- Can you use the Kernel Trick in the primal formation?
- How do kernels relate to overfitting?

References

[1] Stanford CS229: SVMs