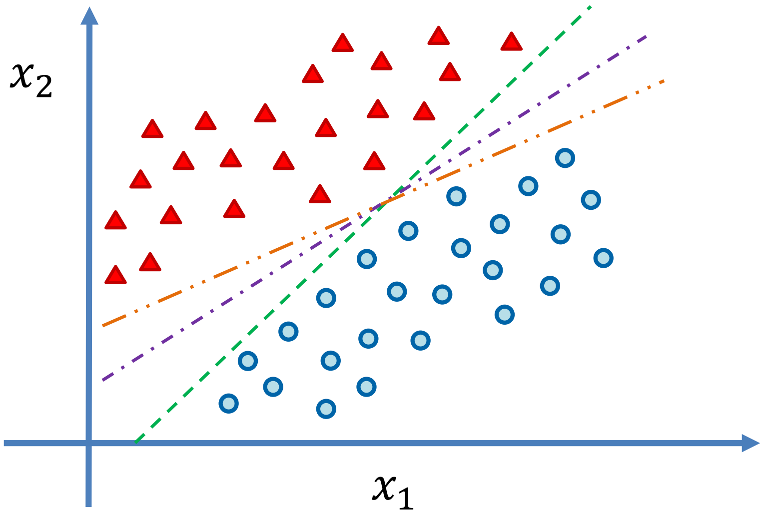
Modern Machine Learning — Support Vector Machines

Kenneth E. Barner
Department of Electrical and Computer Engineering
University of Delaware

Support Vector Machines produce nonlinear boundaries by constructing linear boundaries in a large, transformed version of the feature space

- First derive the case for linear support vector machines with separable data
  - Constructing an optimal separating hyperplane between two perfectly separated classes
- Generalize result to the linear support vector machines with nonseparable data
  - Construct an optimal separating hyperplane focusing on samples near the decision boundary
- Finally, employ the kernel trick to transform the data, yielding nonlinear boundaries in the observation space [covered in a subsequent section]

Assumption: two linearly separable classes



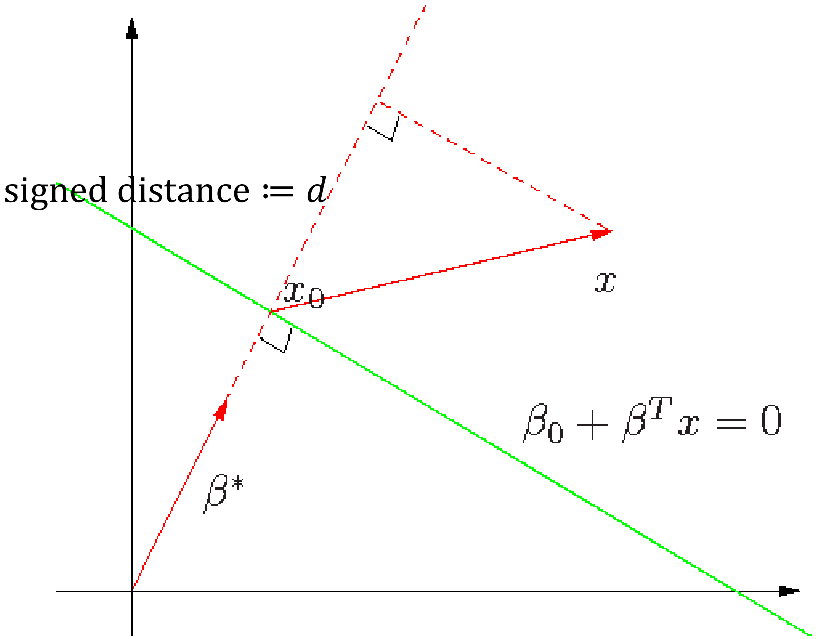
Hyperplane Equation

$\boldsymbol{x}^T\boldsymbol{\beta} + \beta_0 = 0$, where $\boldsymbol{x}^T = [x_1, \dots, x_p]$

and $\boldsymbol{\beta} = [\beta_1, \dots, \beta_p]^T$

Observation: There are an infinite number of separating hyperplanes

To establish the optimal hyperplane, determine the hyperplane vector normal and the signed distances of samples to the hyperplane



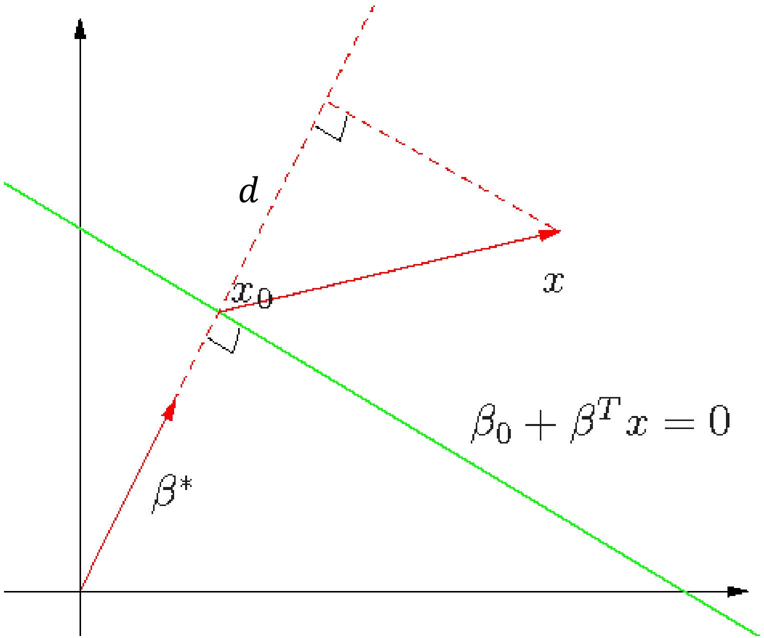
The hyperplane, or *afine set* L, is defined by:

$$f(\boldsymbol{x}) = \beta_0 + \mathbf{x}^T \boldsymbol{\beta} = 0$$

Observations & Definitions:

1. For any two points $\boldsymbol{x}_1, \boldsymbol{x}_2 \in L$, we have:
$$f(\boldsymbol{x}_1) - f(\boldsymbol{x}_2) = \boldsymbol{\beta}^T(\boldsymbol{x}_1 - \boldsymbol{x}_2) = 0$$
$\Rightarrow$ $\boldsymbol{\beta}^*$=$\boldsymbol{\beta}/\|\boldsymbol{\beta}\|$ is the unit vector normal to L
2. For any point $\boldsymbol{x}_0 \in L$, $\boldsymbol{\beta}^T \boldsymbol{x}_0 = -\beta_0$
3. This signed distance, d, of any point $\boldsymbol{x}$ to L is
$$\boldsymbol{\beta}^{*T}(\boldsymbol{x} - \boldsymbol{x}_0) = \frac{1}{\|\boldsymbol{\beta}\|}(\boldsymbol{\beta}^T \boldsymbol{x} + \beta_0)$$
$$= \frac{1}{\|f'(\boldsymbol{x})\|} f(\boldsymbol{x})$$

$\Longrightarrow$ $f(\boldsymbol{x})$ is proportional to the signed distance from $\boldsymbol{x}$ to the hyper plane defined by $f(\boldsymbol{x})$; signed distance $= f(\boldsymbol{x})$ when $\|\boldsymbol{\beta}\| = 1$



Observation: Signed distances can be used as metrics for classification, i.e.

$$G(\boldsymbol{x}) = \operatorname{sgn}(d)$$

Problem Setup:

Binary class case: $(y_i, \mathbf{x}_i) \in \{+1, -1\} \times \mathbb{R}^p$

Separating hyperplane: $\beta_0 + \mathbf{x}^T \boldsymbol{\beta} = 0$, where we enforce the normalization $||\boldsymbol{\beta}|| = 1$

Thus $G(\boldsymbol{x}_i) = \operatorname{sgn}(d) = \operatorname{sgn}(\mathbf{x}_i^T \boldsymbol{\beta} + \beta_0)$

positives signed distance $\Longrightarrow G(\boldsymbol{x}_i) = 1$

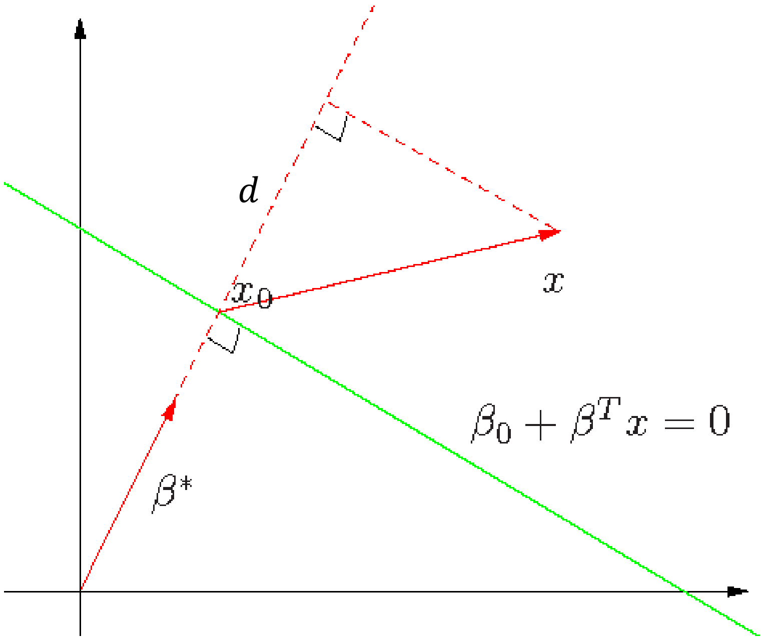
negative signed distance $\Longrightarrow G(\boldsymbol{x}_i) = -1$

Moreover

$$G(\boldsymbol{x}_i) y_i = y_i\left(\mathbf{x}_i^T \boldsymbol{\beta} + \beta_0\right) > 0 \rightarrow \text{correct classification}$$

$$G(\boldsymbol{x}_i) y_i = y_i\left(\mathbf{x}_i^T \boldsymbol{\beta} + \beta_0\right) < 0 \rightarrow \text{incorrect classification}$$

Finally, note that $\left|y_i\left(\mathbf{x}_i^T \boldsymbol{\beta} + \beta_0\right)\right|$ is the distance between $\boldsymbol{x}_i$ and the separating hyperplane



Perfectly Separable Case: In this case,

$$G(\boldsymbol{x})y_i = y_i\left(\mathbf{x}_i^T\boldsymbol{\beta} + \beta_0\right) > 0 \forall i$$

Question: How large can we make the margin, M, separating points from the hyper plane?

Cast the problem as:

$$\max_{\beta_0, \boldsymbol{\beta}, \|\boldsymbol{\beta}\|=1} M$$

subject to

$$y_i\left(\mathbf{x}_i^T\boldsymbol{\beta} + \beta_0\right) \geq M, i = 1, \ldots, n$$

Note: The condition $\|\boldsymbol{\beta}\| = 1$ can be included in the constraint as

$$\frac{1}{\|\boldsymbol{\beta}\|} y_i\left(\mathbf{x}_i^T \boldsymbol{\beta} + \beta_0\right) \geq M$$
$$\Rightarrow y_i\left(\mathbf{x}_i^T \boldsymbol{\beta} + \beta_0\right) \geq M\|\boldsymbol{\beta}\|$$

We can always rescale $(\beta_0, \boldsymbol{\beta})$ so that

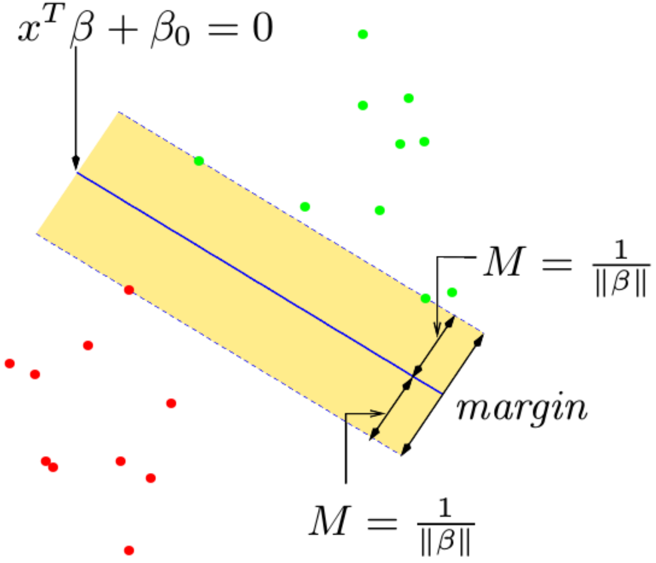
$$\|\boldsymbol{\beta}\| = \frac{1}{M}$$

Therefore, the optimization problem can be expressed as

$$\min_{\beta_0, \boldsymbol{\beta} \in \mathbb{R}^p} \|\boldsymbol{\beta}\|$$

subject to

$$y_i\left(\mathbf{x}_i^T \boldsymbol{\beta} + \beta_0\right) \geq 1, i = 1, \ldots, n$$



Non-Separable Case: Maximize M, but allow for some points to be on the wrong side of the margin.

Define the slack variables $\xi = (\xi_1, \xi_2, \ldots, \xi_n)$

Options to modify $y_i\left(\mathbf{x}_i^T\boldsymbol{\beta} + \beta_0\right) \geq M$

$$y_i\left(\mathbf{x}_i^T\boldsymbol{\beta} + \beta_0\right) \geq M - \xi_i \qquad (*)$$

$$\text{or}$$

$$y_i\left(\mathbf{x}_i^T\boldsymbol{\beta} + \beta_0\right) \geq M(1 - \xi_i) \qquad (**)$$

$\forall i, \xi_i \geq 0, \sum_{i=1}^n \xi_i \leq C$, where C is a constant

Observations:

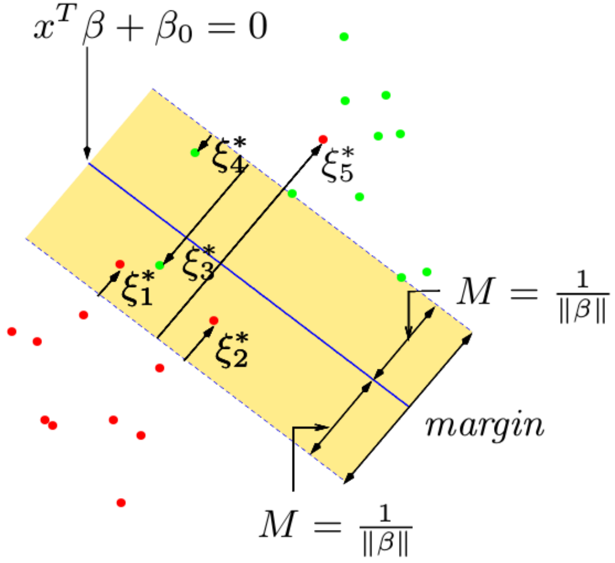
$(*)$ measures overlap in distance from the margin, but results in a nonconvex optimization

$(**)$ measures overlap in relative distance that scales with the margin M, but yields a convex optimization

Thus $(**)$ is referred to as the standard support vector classifier, which yields:

$$\min\|\boldsymbol{\beta}\| \text{ subject to } \begin{cases} y_i\left(\mathbf{x}_i^T\boldsymbol{\beta} + \beta_0\right) \geq 1 - \xi_i, \forall i \\ \xi_i \geq 0, \sum_{i=1}^n \xi_i \leq C \end{cases}$$

where as previously, we set $\mathrm{M} = 1/\|\boldsymbol{\beta}\|$



Summary:

- Linear Support Vector Machines determine the optimal separating hyperplane between two classes
- If the data is perfectly separable, the SVM problem statement is:

$$\min\|\boldsymbol{\beta}\| \ \text{ subject to } \ y_i\left(\mathbf{x}_i^T\boldsymbol{\beta}+\beta_0\right)\geq 1, \forall i$$

- If the data is not separable, the SVM problem statement is:

$$\min\|\boldsymbol{\beta}\| \ \text{ subject to } \begin{cases} y_i\left(\mathbf{x}_i^T\boldsymbol{\beta}+\beta_0\right)\geq 1-\xi_i, \forall i \\ \xi_i\geq 0, \sum_{i=1}^{n}\xi_i\leq C \end{cases}$$

  Where $\xi = (\xi_1, \xi_2, \dots, \xi_n)$ are slack variables and C is a constant that controls the total amount of slack
- Classification is determined as $G(x) = \operatorname{sgn}(\boldsymbol{x}^T\boldsymbol{\beta}+\beta_0)$ for binary (± 1) classes