Support Vector Machines

CSCI-B555

Martha White
Department of Computer Science and Informatics
Indiana University, Bloomington

Fall, 2016

Reminders

- Assignment 3 released
  - Provided python code solves some of assignment 1
- Thought questions due this Wednesday
- Changed to two quizzes instead of three
  - Each quiz administered in class, for 30 minutes

Your Feedback

- Additional notes: recommended textbooks
  - Pattern Recognition by Bishop, Ch 1-5
- Difficulty of course/assignments
  - Assignments are meant to challenge you
  - Quiz/Final will have simpler questions that can be answered in 2 hours
  - You will be allowed a 4 page cheat sheet
- More real-world examples and intuition

Real-world example

- Much of machine learning is about prediction
- Imagine you are google, and you want to **predict** if a user will buy a product
  - How could you make this prediction?
  - What data can you leverage?
  - What learning methods could you use?
  - Any considerations based on the amount of data?
  - Any considerations based on the fact that the prediction has to be made quickly?

Feedback question

- Imagine you have a dataset of 5 points, with d-dimensional features.
  (a) If the corresponding targets are {-3.0, 2.2, -5.3, -1.0, 4.3}, then what estimation technique might you use?

Feedback question

- (b) If the corresponding targets are {1.0, 6.0, 3.0, 2.0, 2.0} and you know y is always a positive integer, then what estimation technique might you use?

Feedback question

- (c) If the corresponding targets are {1, 2, 3, 2, 1} and you know y is always in {1, 2, 3}, then what estimation technique might you use?

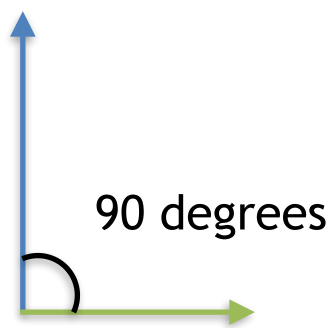
Hyperplanes for SVMs

- For linear classification, would like to separate two classes with a hyperplane
  - Plane characterized by: $\mathbf{w}^\top \mathbf{x} + w_0 = 0$
- We want a hyperplane that separates these classes "the most"
- How do we characterize such a maximal separation?
  - let's talk about vectors in a d-dimensional space
  - let's talk about the distance to a plane

Orthogonality

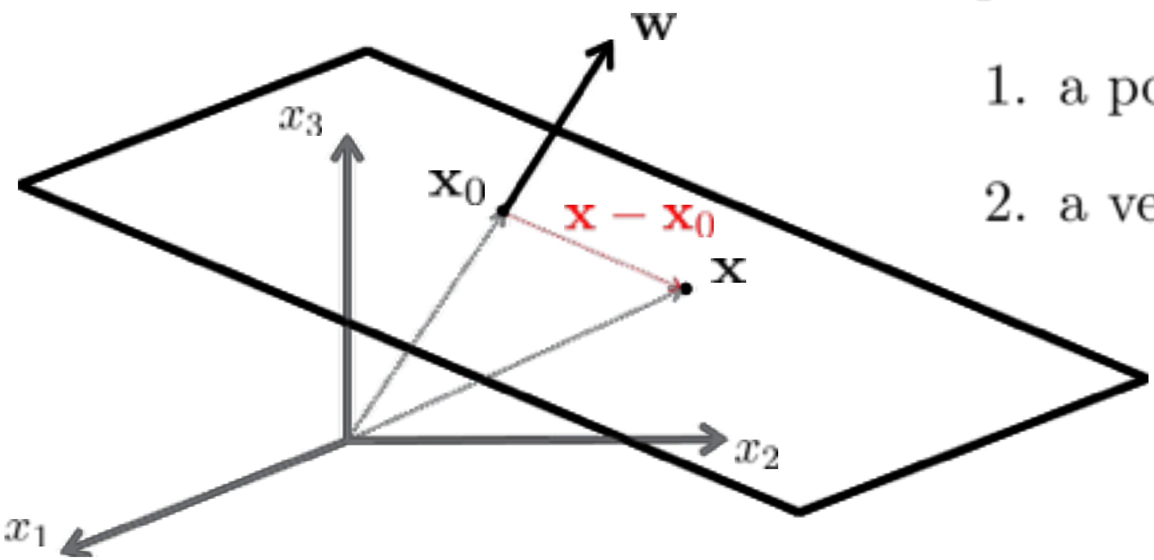
- Two points are orthogonal if dot product is 0
- Cosine similarity: theta angle between w and x

$$\mathbf{w}^\top \mathbf{x} = \|\mathbf{w}\| \|\mathbf{x}\| \cos(\theta)$$



cos(0 degrees) = 0

Equation of the Plane



A plane is defined using:

1. a point $\mathbf{x}_0$ lying in the plane
2. a vector $\mathbf{w}$ normal to the plane

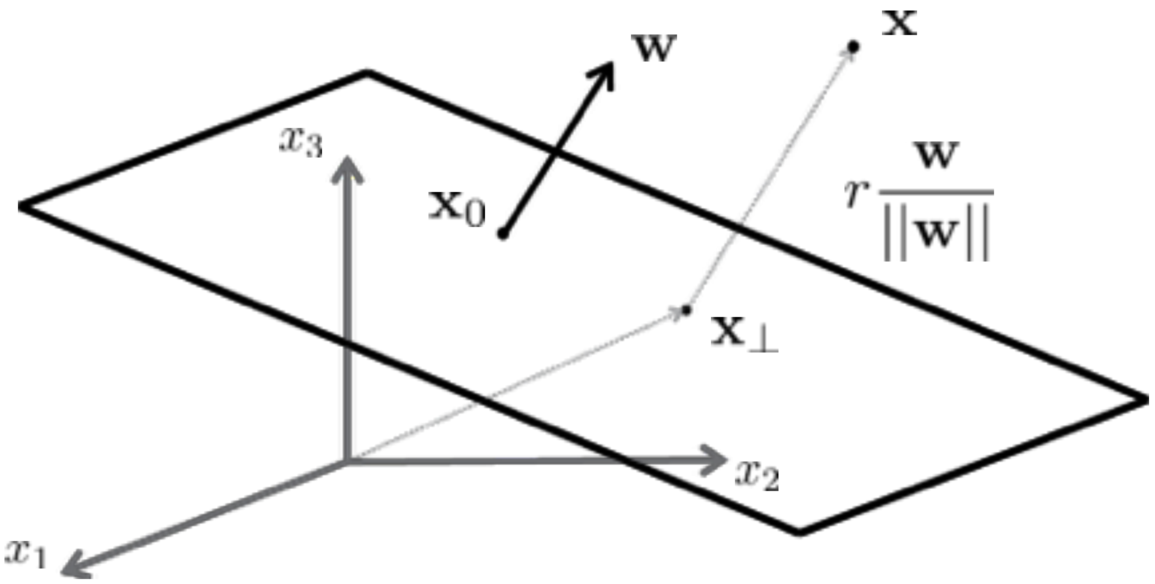
Let $\mathbf{x}$ be on the plane defined by $\mathbf{w}$ and $\mathbf{x}_0$:

$$\mathbf{w}^T(\mathbf{x} - \mathbf{x}_0) = 0$$

$$\mathbf{w}^T\mathbf{x} - \mathbf{w}^T\mathbf{x}_0 = 0$$

$$\boxed{\mathbf{w}^T\mathbf{x} + w_0 = 0}$$

Distance from Point to the Plane



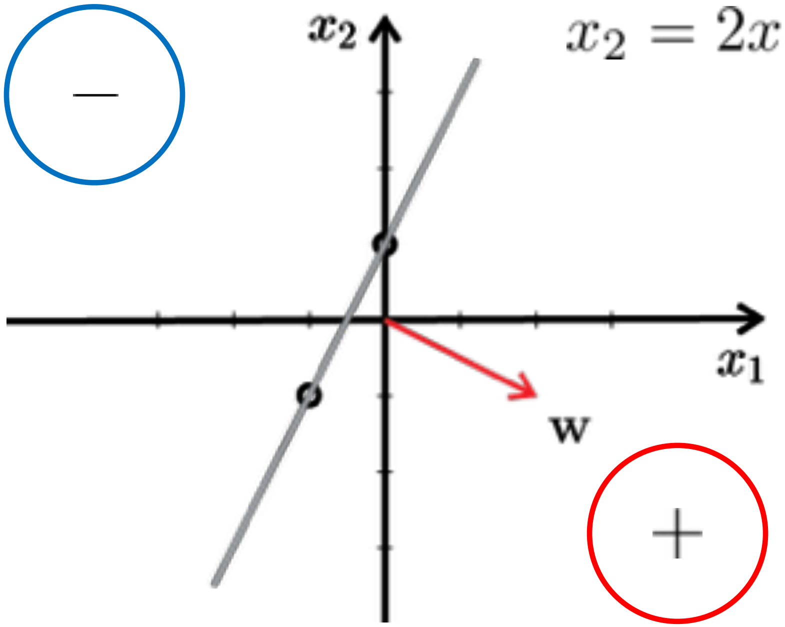
$\mathbf{x}$ = outside the plane

$$\mathbf{x} = \mathbf{x}_\perp + r\frac{\mathbf{w}}{||\mathbf{w}||}$$

$$\mathbf{w}^T\mathbf{x} + w_0 = \underbrace{\mathbf{w}^T\mathbf{x}_\perp + w_0}_{0} + r||\mathbf{w}||$$

$$\boxed{r = \frac{\mathbf{w}^T\mathbf{x} + w_0}{||\mathbf{w}||}}$$

Example



$$x_2 = 2x_1 + 1 \quad \text{or} \quad 2x_1 - x_2 + 1 = 0$$

$$\mathbf{x}, \mathbf{w} \in \mathbb{R}^2$$

$$\mathbf{w}^T\mathbf{x} + w_0 = 0$$

where $\mathbf{w} = (2, -1)$ and $w_0 = 1$.

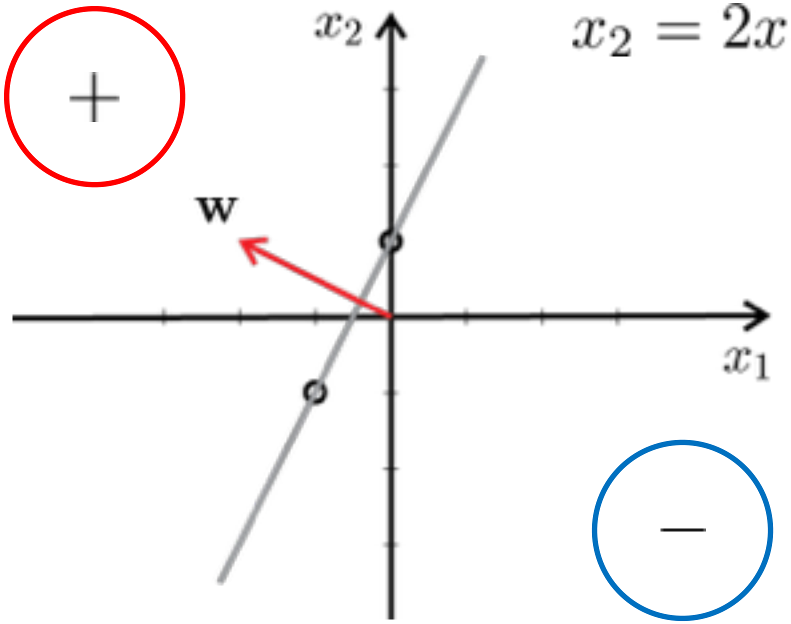
$$r = \frac{\mathbf{w}^T\mathbf{x} + w_0}{||\mathbf{w}||}$$

$$\mathbf{x} = (0,0) \implies r = \frac{1}{\sqrt{5}}$$

$$\mathbf{x} = (-1,1) \implies r = -\frac{2}{\sqrt{5}}$$

The vector $\mathbf{w}$ defines what side of the plane is positive.

EXAMPLE



$x_2 = 2x_1 + 1$

What if $\mathbf{w} = (-2, 1)$?

$\mathbf{x}, \mathbf{w} \in \mathbb{R}^2$

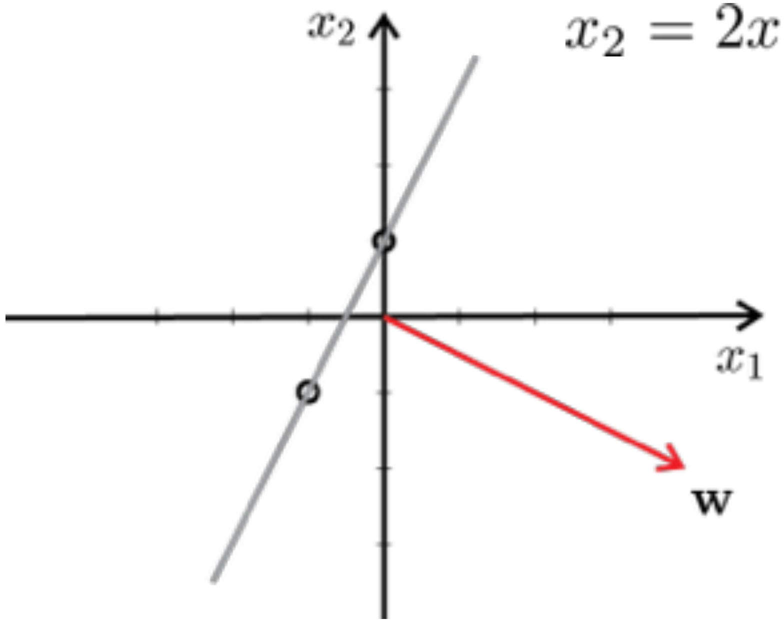
$\mathbf{w}^T\mathbf{x} + w_0 = 0$
where $\mathbf{w} = (-2, 1)$ and $w_0 = -1$.

$$r = \frac{\mathbf{w}^T\mathbf{x} + w_0}{||\mathbf{w}||}$$

$$\mathbf{x} = (0, 0) \implies r = -\frac{1}{\sqrt{5}}$$

$$\mathbf{x} = (-1, 1) \implies r = \frac{2}{\sqrt{5}}$$

Example



$$x_2 = 2x_1 + 1$$

What if $\mathbf{w} = (4, -2)$ and $w_0 = 2$?

$$4x_1 - 2x_2 + 2 = 0$$

$\mathbf{w}^T\mathbf{x} + w_0$ is "bigger"!!!

$$r = \frac{\mathbf{w}^T\mathbf{x} + w_0}{||\mathbf{w}||}$$

$$\mathbf{x} = (0,0) \quad \Longrightarrow r = \frac{1}{\sqrt{5}}$$

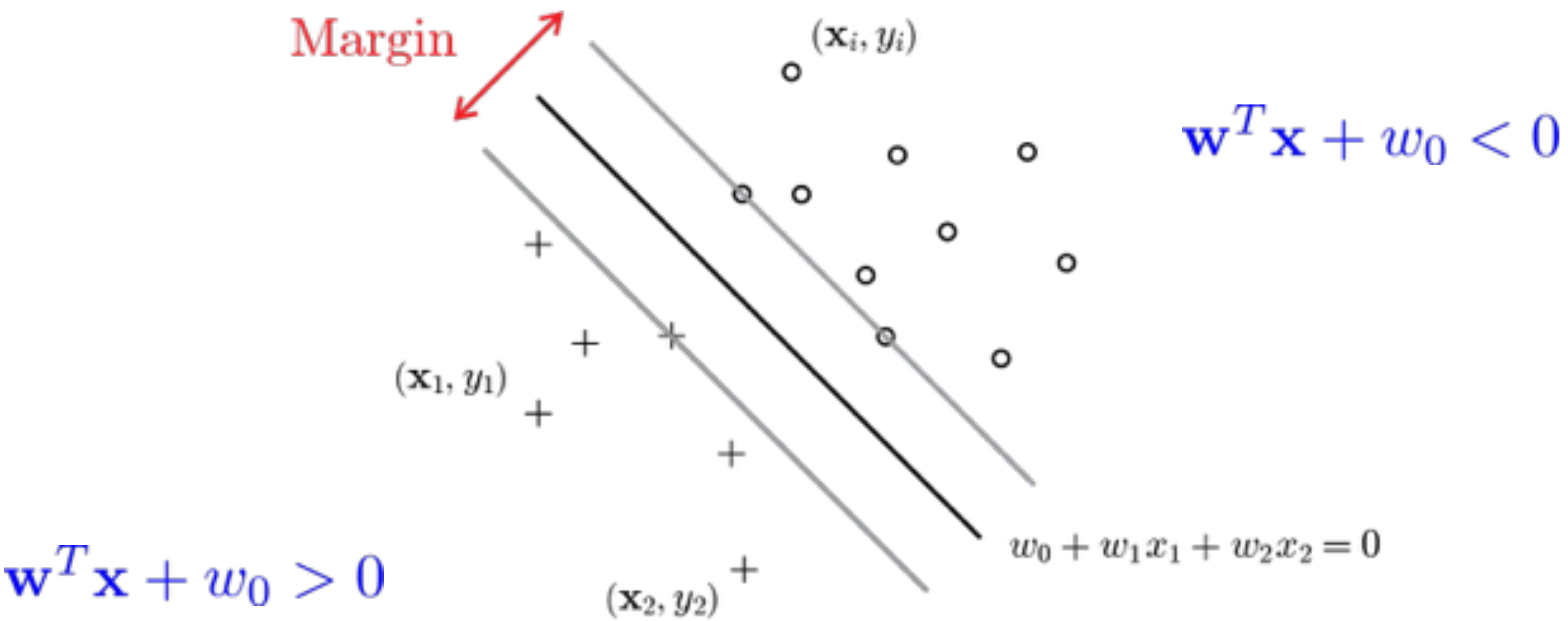
$$\mathbf{x} = (-1,1) \quad \Longrightarrow r = -\frac{2}{\sqrt{5}}$$

Distances are unchanged when $\mathbf{w}$ and w_0 are multiplied by a constant!

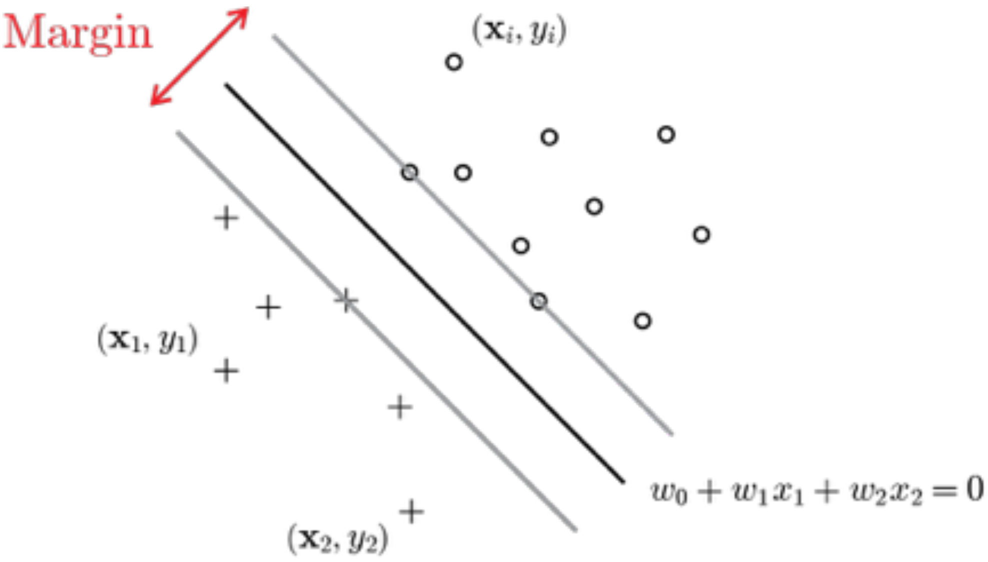
Problem Formulation

Given: $\mathcal{D} = \{(\mathbf{x}_i, y_i\}_{i=1}^{n}$, where $\mathbf{x}_i \in \mathbb{R}^k$ and $y_i \in \{-1, +1\}$.
Data is linearly separable.

Objective: Find hyperplane such that the minimum distance from any data point to the hyperplane is maximized.



Maximizing Margin



$$\mathbf{w}^T\mathbf{x}_i + w_0 > 0 \implies y_i = +1$$
$$\mathbf{w}^T\mathbf{x}_i + w_0 < 0 \implies y_i = -1$$

$$y_i(\mathbf{w}^T\mathbf{x}_i + w_0) > 0$$

$$i \in \{1, 2, \ldots, n\}$$

Idea: find $\mathbf{w}$ to maximize unsigned distance $d_i = \dfrac{y_i(\mathbf{w}^T\mathbf{x} + w_0)}{||\mathbf{w}||}$

$$\boxed{(\mathbf{w}^*, w_0^*) = \arg\max_{\mathbf{w}, w_0} \left\{ \frac{1}{||\mathbf{w}||} \min_i \left(y_i(\mathbf{w}^T\mathbf{x}_i + w_0) \right) \right\}}$$

Reformulating the Problem

$$(\mathbf{w}^*, w_0^*) = \arg\max_{\mathbf{w}, w_0} \left\{ \frac{1}{||\mathbf{w}||} \min_i \left(y_i(\mathbf{w}^T \mathbf{x}_i + w_0) \right) \right\}$$

Scale $\mathbf{w}$ and w_0 such that $\min_i y_i(\mathbf{w}^\top \mathbf{x}_i + w_0) = 1$

$$\mathbf{w} \leftarrow k \cdot \mathbf{w}$$
$$w_0 \leftarrow k \cdot w_0$$

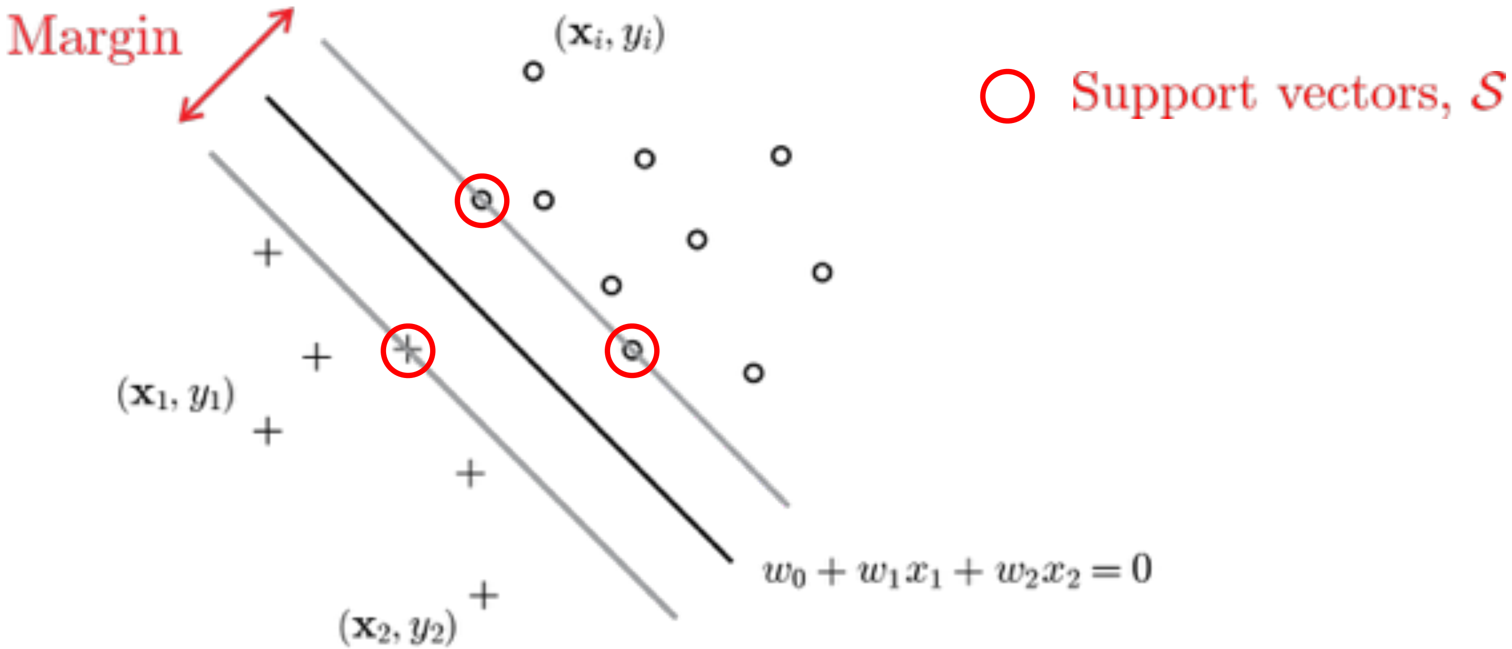
Equivalence class of w, since distance is the same for all of these points, objective the same

$$(\mathbf{w}^*, w_0^*) = \arg\min_{\mathbf{w}} \{||\mathbf{w}||\}$$

Subject to:

$$y_i(\mathbf{w}^T \mathbf{x}_i + w_0) \geq 1 \quad \forall i \in \{1, 2, \ldots, n\}$$

Final Problem Formulation



$$(\mathbf{w}^*, w_0^*) = \arg\min_{\mathbf{w}} \left\{ \frac{1}{2} \mathbf{w}^T \mathbf{w} \right\}$$ ← Convex function!

Subject to:

$$y_i(\mathbf{w}^T \mathbf{x}_i + w_0) \geq 1 \quad \forall i \in \{1, 2, \ldots, n\}$$ ← Linear constraints!

How can We Solve it?

$$(\mathbf{w}^*, w_0^*) = \arg\min_{\mathbf{w}} \left\{ \frac{1}{2} \mathbf{w}^T \mathbf{w} \right\}$$

Subject to:

$$y_i(\mathbf{w}^T \mathbf{x}_i + w_0) \geq 1 \quad \forall i \in \{1, 2, \ldots, n\}$$

Need to know more about constrained optimization

Constrained Optimization

Objective: solve the following optimization problem

$$\mathbf{x}^* = \arg\max_{\mathbf{x}} \{f(\mathbf{x})\}$$

Subject to:

$$g_i(\mathbf{x}) = 0 \qquad \forall i \in \{1, 2, \ldots, m\}$$

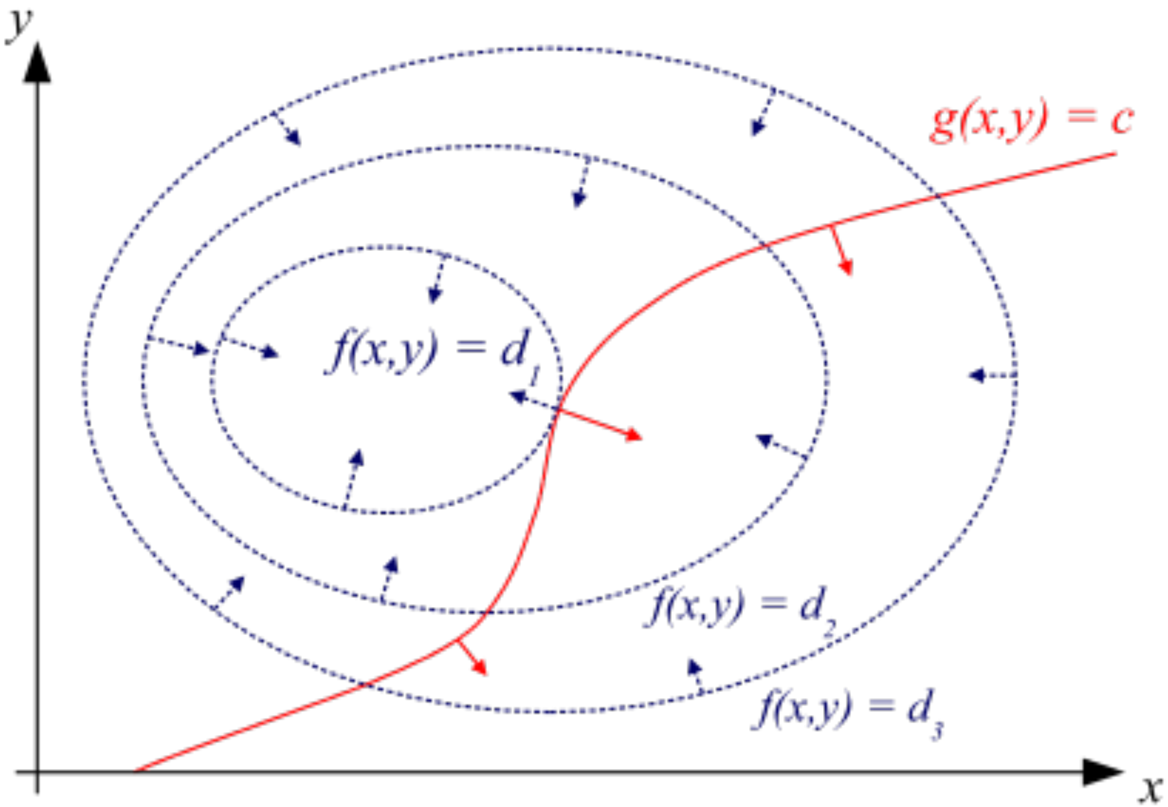
$$h_j(\mathbf{x}) \geq 0 \qquad \forall j \in \{1, 2, \ldots, n\}$$

Or, in a shorter notation, to:

$$\mathbf{g}(\mathbf{x}) = \mathbf{0}$$

$$\mathbf{h}(\mathbf{x}) \geq \mathbf{0}$$

INTUITION ON LAGRANGE MULTIPLIERS



Lagrange Multipliers

Taylor's expansion for $g(\mathbf{x})$, where $\mathbf{x}+\boldsymbol{\epsilon}$ is on the surface of $g(\mathbf{x})$

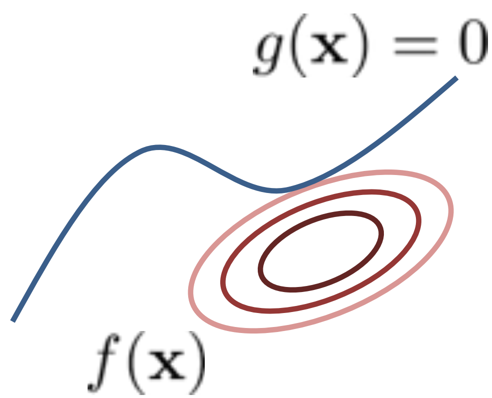
$$g(\mathbf{x}+\boldsymbol{\epsilon}) \approx g(\mathbf{x}) + \boldsymbol{\epsilon}^T \nabla g(\mathbf{x})$$

We know that $g(\mathbf{x}) = g(\mathbf{x}+\boldsymbol{\epsilon})$

$$\boldsymbol{\epsilon}^T \nabla g(\mathbf{x}) \approx 0$$

when $\boldsymbol{\epsilon} \rightarrow \mathbf{0}$

$$\boldsymbol{\epsilon}^T \nabla g(\mathbf{x}) = 0 \quad \Longrightarrow \quad \nabla g(\mathbf{x}) \text{ is orthogonal to the surface}$$



$\nabla g(\mathbf{x})$ and $\nabla f(\mathbf{x})$ are parallel!

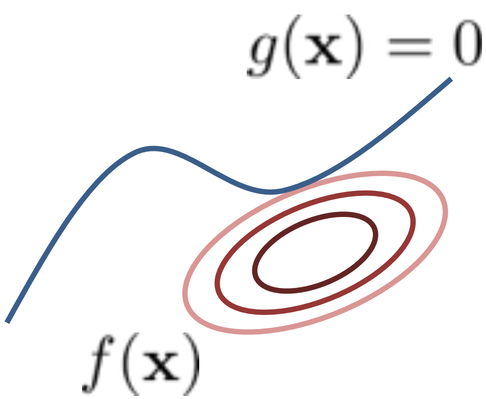
$$\nabla f(\mathbf{x}) + \alpha \nabla g(\mathbf{x}) = 0 \qquad \alpha \neq 0$$

Not a step-size
This is a Lagrange multiplier

$$L(\mathbf{x}, \alpha) = f(\mathbf{x}) + \alpha g(\mathbf{x})$$

More intuition on Lagrange multipliers

The two gradients are parallel,
but not necessarily of the same magnitude
The Lagrange multiplier adapts to
this difference in magnitude



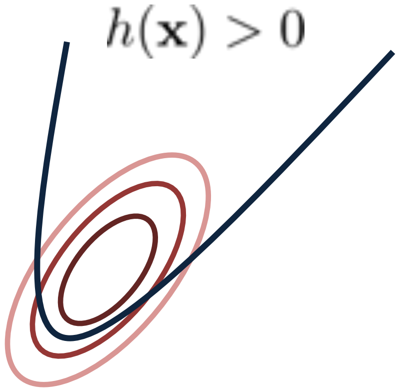
$\nabla g(\mathbf{x})$ and $\nabla f(\mathbf{x})$ are parallel!

$$\nabla f(\mathbf{x}) + \alpha \nabla g(\mathbf{x}) = 0$$

$$L(\mathbf{x}, \alpha) = f(\mathbf{x}) + \alpha g(\mathbf{x})$$

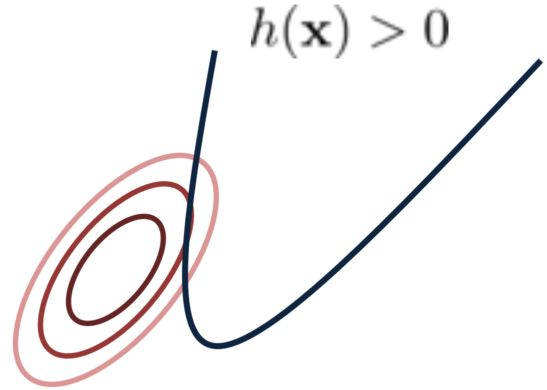
Lagrange Multipliers

Inactive constraint



$$\nabla f(\mathbf{x}) = 0$$

Active constraint



$$\nabla f(\mathbf{x}) = -\mu \nabla h(\mathbf{x}) \qquad \mu > 0$$

It holds that:

$$h(\mathbf{x}) \geq 0$$
$$\mu \geq 0$$
$$\mu \cdot h(\mathbf{x}) = 0$$

Karush-Kuhn-Tucker (KKT) conditions

Note: alpha rather than mu is used for inequality constraint in SVMs; an unfortunate historical choice, but we stick with it next

$$L(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\mu}) = f(\mathbf{x}) + \boldsymbol{\alpha}^T \mathbf{g}(\mathbf{x}) + \boldsymbol{\mu}^T \mathbf{h}(\mathbf{x})$$

How can We Solve it?

$$(\mathbf{w}^*, w_0^*) = \arg\min_{\mathbf{w}} \left\{ \frac{1}{2}\mathbf{w}^T\mathbf{w} \right\}$$

Subject to:

$$y_i(\mathbf{w}^T\mathbf{x}_i + w_0) \geq 1 \quad \forall i \in \{1, 2, \ldots, n\}$$

Solution: use Lagrangian multipliers!

$$L(\mathbf{w}, w_0, \boldsymbol{\alpha}) = \frac{1}{2}\mathbf{w}^T\mathbf{w} - \sum_{i=1}^{n} \alpha_i \left(y_i \left(\mathbf{w}^T\mathbf{x}_i + w_0\right) - 1\right)$$

$$\max_{\alpha} \min_{\mathbf{w}, w_0} L(\mathbf{w}, w_0, \boldsymbol{\alpha}) \qquad \alpha_i \geq 0$$

Solving it

$$\frac{\partial}{\partial w_j} L(\mathbf{w}, w_0, \boldsymbol{\alpha}) = 0 \quad \Longrightarrow \quad w_j = \sum_{i=1}^{n} \alpha_i y_i x_{ij}$$

$$\Longrightarrow \quad \mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i$$

$$\frac{\partial}{\partial w_0} L(\mathbf{w}, w_0, \boldsymbol{\alpha}) = 0 \quad \Longrightarrow \quad \sum_{i=1}^{n} \alpha_i y_i = 0$$

Dual Problem

$$\boxed{\sum_{i=1}^{n} \alpha_i y_i = 0 \qquad \mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i}$$

$$\begin{aligned}
L(\mathbf{w}, w_0, \boldsymbol{\alpha}) &= \frac{1}{2}\mathbf{w}^T\mathbf{w} - \sum_{i=1}^{n} \alpha_i y_i \mathbf{w}^T \mathbf{x}_i - \sum_{i=1}^{n} \alpha_i y_i w_0 + \sum_{i=1}^{n} \alpha_i \\
&= \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j - \sum_{i=1}^{n} \alpha_i y_i (\sum_{j=1}^{n} \alpha_j y_j \mathbf{x}_j)^T \mathbf{x}_i + \sum_{i=1}^{n} \alpha_i \\
&= \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j \\
&= \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j)
\end{aligned}$$

kernel property

$$k(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^\top \mathbf{x}_j$$

Subject to:

$$\alpha_i \geq 0 \qquad \forall i \in \{1, 2, \ldots, n\}$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0$$

Solving the Dual Problem

Use quadratic programming to solve for $\boldsymbol{\alpha}$

Then set

$$\Longrightarrow \qquad \mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i$$

$$k(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^\top \mathbf{x}_j$$

$$\Longrightarrow \qquad \begin{aligned} f(\mathbf{x}) &= \mathbf{w}^T \mathbf{x} + w_0 \\ &= \frac{1}{2} \sum_{i=1}^{n} \alpha_i y_i k(\mathbf{x}, \mathbf{x}_i) + w_0 \end{aligned}$$

Analysis of the Solution

Karush-Kuhn-Tucker (KKT) conditions:

$$\alpha_i \geq 0$$

$$y_i(\mathbf{w}^T\mathbf{x}_i + w_0) - 1 \geq 0 \qquad\qquad \forall i \in \{1, 2, \ldots, n\}$$

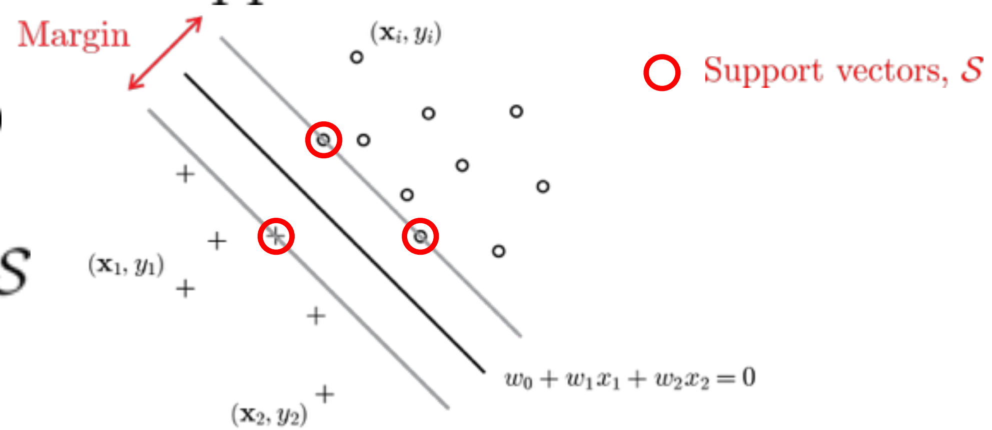
$$\alpha_i \left(y_i \left(\mathbf{w}^T\mathbf{x}_i + w_0\right) - 1\right) = 0$$

This means that for $\forall i$, either $\alpha_i = 0$ or $y_i \left(\mathbf{w}^T\mathbf{x}_i + w_0\right) = 1$

$\alpha_i = 0$ for all vectors that are not support vectors

$$f(\mathbf{x}) = \sum_{\mathbf{x}_i \in \mathcal{S}} \alpha_i y_i k(\mathbf{x}, \mathbf{x}_i) + w_0$$

$w_0 = 1 - \mathbf{w}^T\mathbf{x}_s$, where $\mathbf{x}_s \in \mathcal{S}$



A Support Vector Machine