IAML: Support Vector Machines I

Nigel Goddard
School of Informatics

Semester 1

Outline

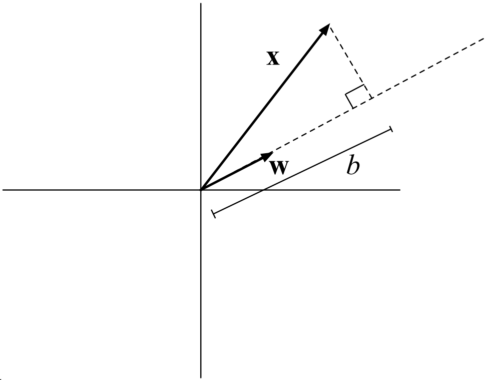
- Separating hyperplane with maximum margin
- Non-separable training data
- Expanding the input into a high-dimensional space
- Support vector regression
- Reading: W & F sec 6.3 (maximum margin hyperplane, nonlinear class boundaries), SVM handout. SV regression not examinable.

Overview

- Support vector machines are one of the most effective and widely used classification algorithms.
- SVMs are the combination of two ideas
  - Maximum margin classification
  - The “kernel trick”
- SVMs are a linear classifier, like logistic regression and perceptron

Stuff You Need to Remember

$\mathbf{w}^\top \mathbf{x}$ is length of the projection of $\mathbf{x}$ onto $\mathbf{w}$ (if $\mathbf{w}$ is a unit vector)



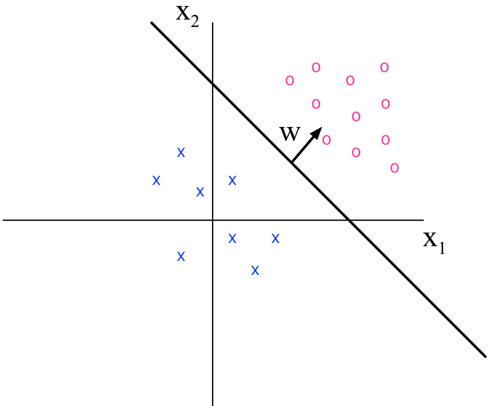
i.e., $b = \mathbf{w}^T \mathbf{x}$.

(If you do not remember this, see supplementary maths notes on course Web site.)

Separating Hyperplane

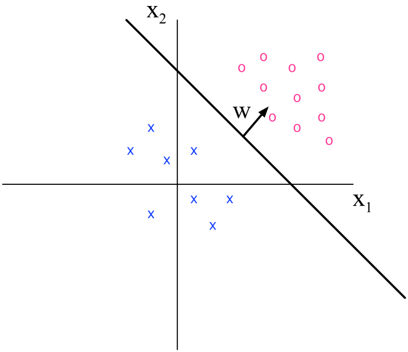
For any linear classifier

- Training instances $(\mathbf{x}_i, y_i)$, $i = 1, \ldots, n$. $y_i \in \{-1, +1\}$
- Hyperplane $\mathbf{w}^\top \mathbf{x} + w_0 = 0$
- Notice for this lecture we use -1 rather than 0 for negative class. This will be convenient for the maths.

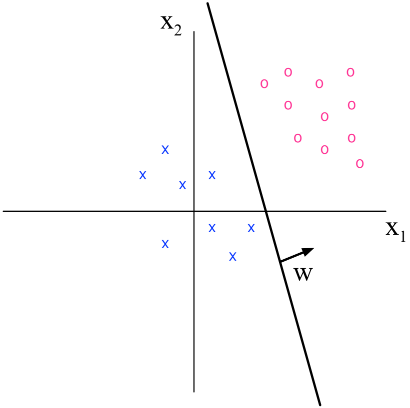


A Crap Decision Boundary

Seems okay

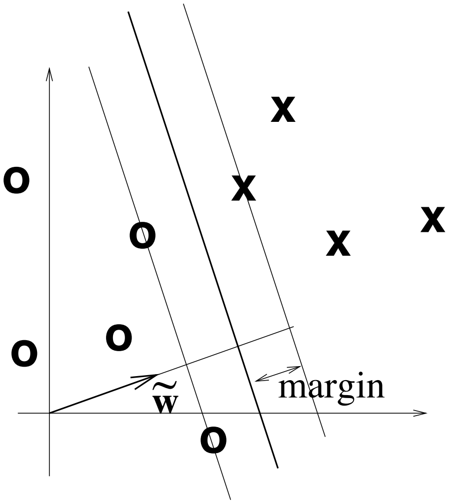


This is crap



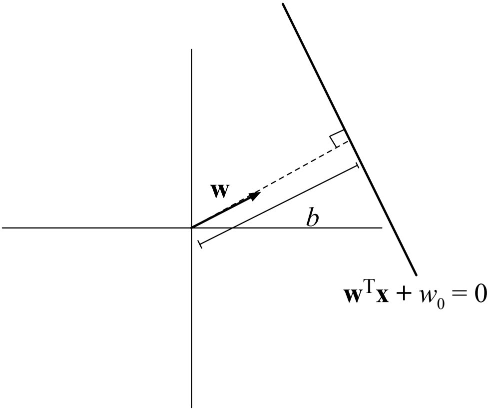
Idea: Maximize the Margin

The **margin** is the distance between the decision boundary (the hyperplane) and the closest training point.

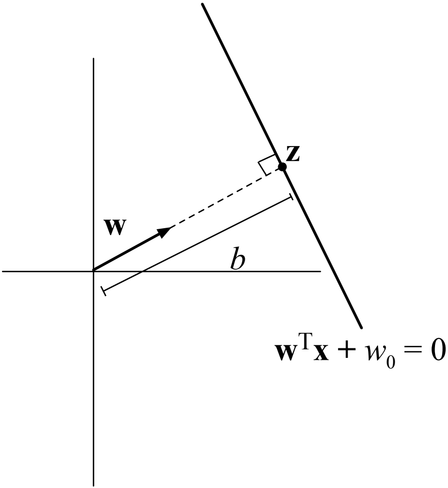


Computing the Margin

- The tricky part will be to get an equation for the margin
- We'll start by getting the distance from the origin to the hyperplane
- i.e., We want to compute the scalar *b* below



Computing the Distance to Origin



- Define $\mathbf{z}$ as the point on the hyperplane closest to the origin.
- $\mathbf{z}$ must be proportional to $\mathbf{w}$, because $\mathbf{w}$ normal to hyperplane
- By definition of b, we have the norm of z given by:

$$||\mathbf{z}|| = b$$

So

$$b\frac{\mathbf{w}}{||\mathbf{w}||} = \mathbf{z}$$

Computing the Distance to Origin

- We know that (a) $\mathbf{z}$ on the hyperplane and (b) $b\frac{\mathbf{w}}{||\mathbf{w}||} = \mathbf{z}$.
- First (a) means $\mathbf{w}^T\mathbf{z} + w_0 = 0$
- Substituting we get

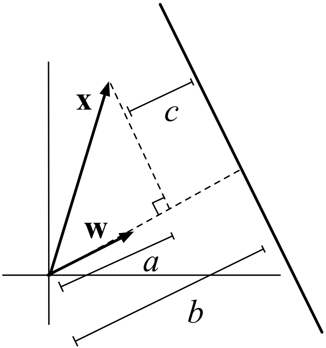
$$\mathbf{w}^T \frac{b\mathbf{w}}{||\mathbf{w}||} + w_0 = 0$$

$$\frac{b\mathbf{w}^T\mathbf{w}}{||\mathbf{w}||} + w_0 = 0$$

$$b = -\frac{w_0}{||\mathbf{w}||}$$

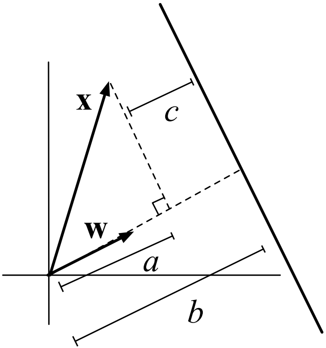
- Remember $||\mathbf{w}|| = \sqrt{\mathbf{w}^T\mathbf{w}}$.
- Now we have the distance from the origin to the hyperplane!

Computing the Distance to Hyperplane



- Now we want c, the distance from **x** to the hyperplane.
- It's clear that $c = |b - a|$, where a the length of the projection of **x** onto **w**. Quiz: What is a?

Computing the Distance to Hyperplane



- Now we want c, the distance from $\mathbf{x}$ to the hyperplane.
- It's clear that $c = |b - a|$, where a the length of the projection of $\mathbf{x}$ onto $\mathbf{w}$. Quiz: What is a?

$$a = \frac{\mathbf{w}^T\mathbf{x}}{||\mathbf{w}||}$$

Equation for the Margin

- The perpendicular distance from a point $\mathbf{x}$ to the hyperplane $\mathbf{w}^T\mathbf{x} + w_0 = 0$ is

$$\frac{1}{||\mathbf{w}||}|\mathbf{w}^T\mathbf{x} + w_0|$$

- The **margin** is the distance from the closest training point to the hyperplane

$$\min_i \frac{1}{||\mathbf{w}||}|\mathbf{w}^T\mathbf{x}_i + w_0|$$

The Scaling

- Note that $(\mathbf{w}, w_0)$ and $(c\mathbf{w}, cw_0)$ defines the same hyperplane. The scale is arbitrary.
- This is because we predict class $y = 1$ if $\mathbf{w}^T\mathbf{x} + w_0 \geq 0$. That's the same thing as saying $c\mathbf{w}^T\mathbf{x} + cw_0 \geq 0$
- To remove this freedom, we will put a constraint on $(\mathbf{w}, w_0)$

$$\min_i |\mathbf{w}^\top \mathbf{x}_i + w_0| = 1$$

- With this constraint, the margin is always $1/||\mathbf{w}||$.

First version of Max Margin Optimization Problem

- Here is a first version of an optimization problem to maximize the margin (we will simplify)

$$\begin{aligned} \max_{\mathbf{w}} \quad & 1/||\mathbf{w}|| \\ \text{subject to} \quad & \mathbf{w}^\top \mathbf{x}_i + w_0 \geq 0 && \text{for all } i \text{ with } y_i = 1 \\ & \mathbf{w}^\top \mathbf{x}_i + w_0 \leq 0 && \text{for all } i \text{ with } y_i = -1 \\ & \min_i |\mathbf{w}^\top \mathbf{x}_i + w_0| = 1 \end{aligned}$$

- The first two constraints are too lose. It's the same thing to say

$$\begin{aligned} \max_{\mathbf{w}} \quad & 1/||\mathbf{w}|| \\ \text{subject to} \quad & \mathbf{w}^\top \mathbf{x}_i + w_0 \geq 1 && \text{for all } i \text{ with } y_i = 1 \\ & \mathbf{w}^\top \mathbf{x}_i + w_0 \leq -1 && \text{for all } i \text{ with } y_i = -1 \\ & \min_i |\mathbf{w}^\top \mathbf{x}_i + w_0| = 1 \end{aligned}$$

- Now the third constraint is redundant

First version of Max Margin Optimization Problem

- That means we can simplify to

$$\begin{aligned} \max_{\mathbf{w}} \;& 1/||\mathbf{w}|| \\ \text{subject to} \;& \mathbf{w}^\top \mathbf{x}_i + w_0 \geq 1 && \text{for all } i \text{ with } y_i = 1 \\ & \mathbf{w}^\top \mathbf{x}_i + w_0 \leq -1 && \text{for all } i \text{ with } y_i = -1 \end{aligned}$$

- Here's a compact way to write those two constraints

$$\begin{aligned} \max_{\mathbf{w}} \;& 1/||\mathbf{w}|| \\ \text{subject to} \;& y_i(\mathbf{w}^\top \mathbf{x}_i + w_0) \geq 1 && \text{for all } i \end{aligned}$$

- Finally, note that maximizing $1/||\mathbf{w}||$ is the same thing as *minimizing* $||\mathbf{w}||^2$

The SVM optimization problem

- So the SVM weights are determined by solving the optimization problem:

$$\begin{aligned} \min_{\mathbf{w}} \quad & ||\mathbf{w}||^2 \\ \text{s.t.} \quad & y_i(\mathbf{w}^\top \mathbf{x}_i + w_0) \geq +1 \qquad \text{for all } i \end{aligned}$$

- Solving this will require maths that we don't have in this course. But I'll show the form of the solution next time.

Fin (Part I)