Optimal Decision Surfaces

In our previous learning algorithms we have seen that there are many different ways to construct decision surfaces that separate two classes.

In the case of the "simple learning algorithm" we have seen that it attempts to construct a decision surface right in the middle between the two class means.

- Seems to work well and is intuitive, but outliers deep in the classes can distort the decision surface.

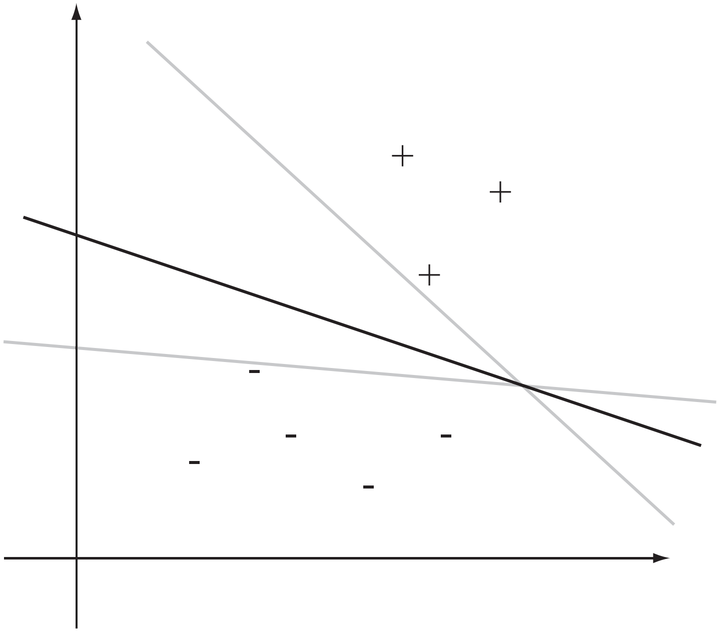
In the case of the perceptron we saw that the algorithm attempts for find decision surface that simply separates the two classes.

- from our discussion of the dual representation of the learning algorithm it follows that the only points that really influence the decision surface are the points close to it.
- but we also saw that because the algorithm is a greedy search no attempts at optimizing the position of the decision surface.

This raises the issue: what is ***optimality*** with respect to decision surface position?

Optimal Decision Surfaces

Definition: An ***optimal*** decision surface maximizes the probability of classifying unseen data points correctly.



Maximum Margin Classifiers

Characterizing the optimal decision surface.

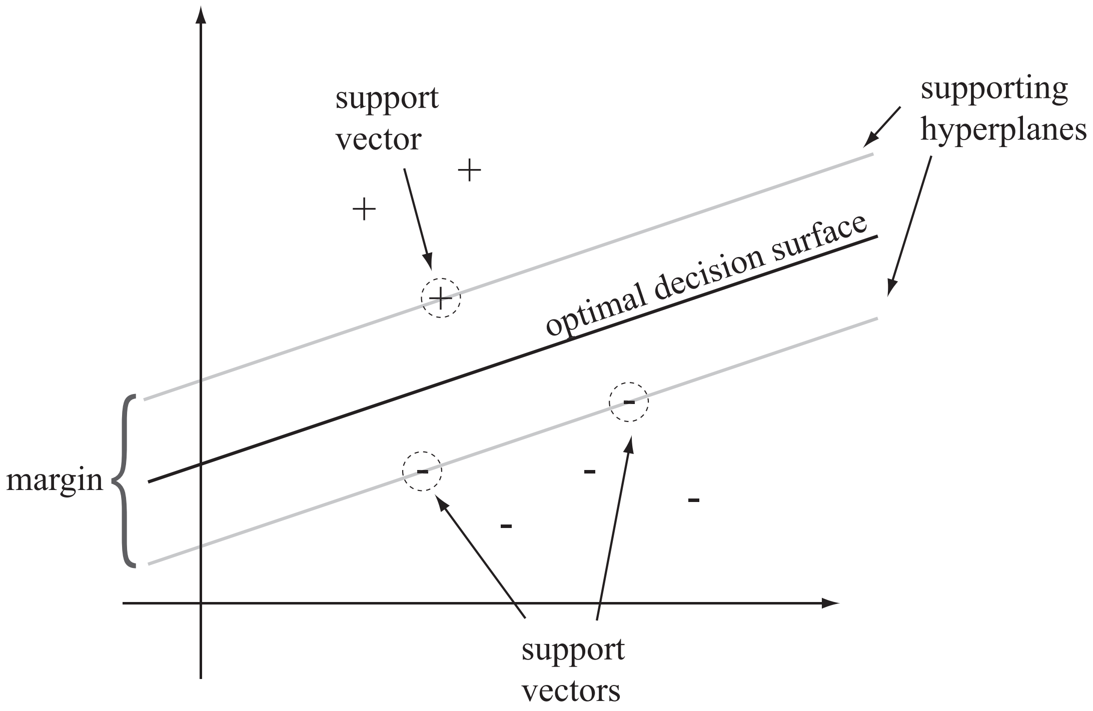
Definition: A hyperplane ***supports a class*** if it is parallel to the (linear) decision surface and all points of that class are either above or below the supporting plane.

Definition: The distance between two supporting hyperplanes is called a ***margin***.

Maximum Margin Classifiers

We can now define decision surface optimality in terms of the margin between supporting planes:

Definition: A (linear) decision surface for a binary classification problem is ***optimal*** if it is equidistant to two supporting hyperplanes and maximizes the margin between the two supporting hyperplanes.



Optimality implies Optimization

The fact that we are searching for a decision surface with an optimality criterion such as the ‘maximum margin’ implies an optimization problem.

Optimization problems are problems in which we want to select the best solution from a number of possible or feasible solutions. The feasible solutions are ranked by an objective function.

We define optimization problems formally,

$$\min_{\overline{x}} \phi(\overline{x}),$$

such that,

$$h_i(\overline{x}) \geq c_i,$$

with $i = 1, \ldots, l$ and for all $\overline{x} \in \mathbb{R}^n$.

Here, the function $\phi : \mathbb{R}^n \rightarrow \mathbb{R}$ is the ***objective function*** and each function $h_i : \mathbb{R}^n \rightarrow \mathbb{R}$ is called a ***constraint*** with ***bound*** c_i.

Optimality implies Optimization

Any point $\overline{x}^*$ such that $h_i(\overline{x}^*) \geq c_i$ for all i is called a ***feasible solution***. In addition we call $\overline{x}^*$ an ***optimal solution*** if

$$\phi(\overline{x}^*) = \min_{\overline{x}} \phi(\overline{x}),$$

or

$$\phi(\overline{x}^*) \leq \phi(\overline{q}),$$

for all $\overline{q} \in \mathbb{R}^n$ and $h_i(\overline{q}) \geq c_i$ for all i.

$\Rightarrow$ That is the optimization operator $\min$ finds the smallest feasible solution.

Optimality implies Optimization

We have defined optimization only in terms of minimization.

However, we can turn any maximization problem into a minimization problem with one of the following identities,

$$\max \phi(\overline{x}) = \min -\phi(\overline{x}),$$

$$\max \phi(\overline{x}) = \min \frac{1}{\phi(\overline{x})}.$$

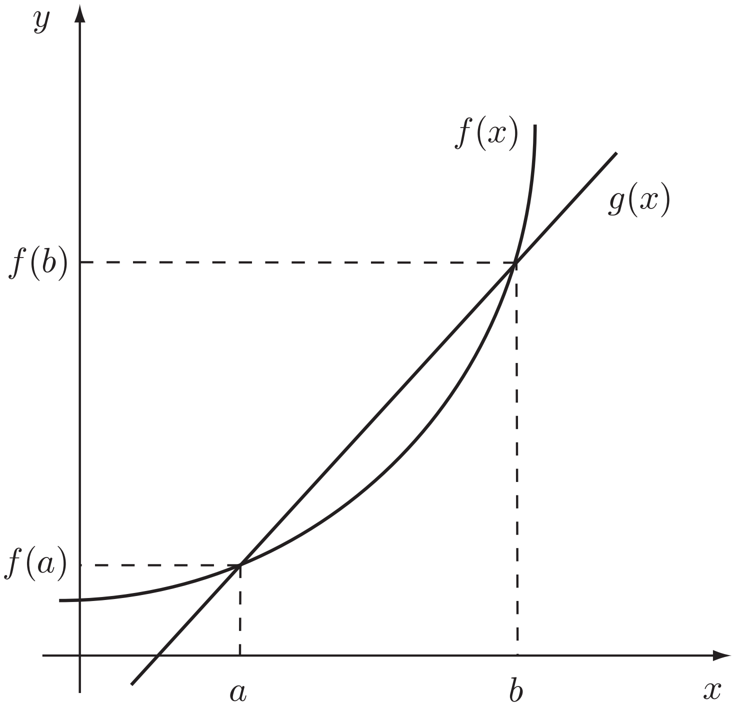
With the second identity we have to be careful that $1/\phi(\overline{x})$ is well defined for all $\overline{x}$.

Convex Optimization

A convex optimization problem has a convex objective function and linear constraints.

Convex optimization problems are particularly well behaved:

- the objective function has a global minimum,
- the function surface is smooth in that we can connect any two points on the function surface with a linear function without crossing the function surface itself.

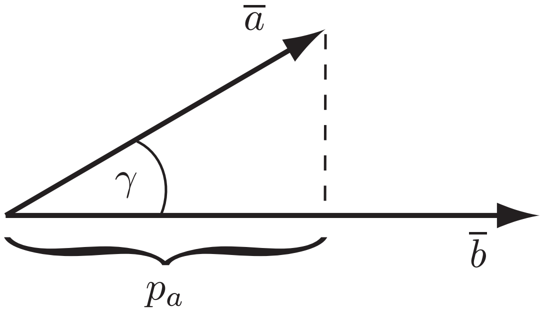


Maximizing the margin is a convex optimization problem!

Projections

Definition: Let $\overline{a}$ and $\overline{b}$ be vectors in $\mathbb{R}^n$ that form an angle γ between them, then we say that p_a is the ***projection*** of $\overline{a}$ in the direction of $\overline{b}$, such that,

$$p_a = |\overline{a}| \cos\gamma = \frac{\overline{a} \bullet \overline{b}}{|b|}.$$



Optimizing the Margin

Let us assume that we have a linearly separable training set,

$$D = \{(\overline{x}_1, y_1), (\overline{x}_2, y_2), \ldots, (\overline{x}_l, y_l)\} \subseteq \mathbb{R}^n \times \{+1, -1\}.$$

Let us also assume that we have the optimal decision surface for this training set,

$$\overline{w}^* \bullet \overline{x} = b^*.$$

Since this decision surface is optimal, the following identities must hold,

$$m^* = \phi(\overline{w}^*, b^*) = \max \phi(\overline{w}, b),$$

where m^* is the maximum margin.

Observation: Our goal is to derive an expression for the objective function so that we can use it to optimize $\overline{w}$ and b and in this way obtain an optimal margin.

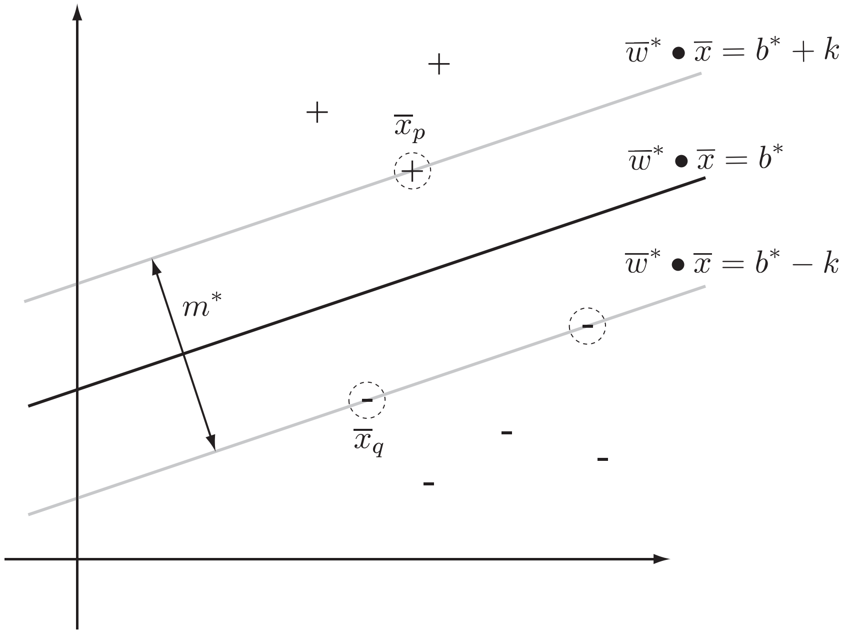
Optimizing the Margin

Since $\overline{w}^* \bullet \overline{x} = b^*$ is the optimal decision surface we have two supporting hyperplanes, say

$$\overline{w}^* \bullet \overline{x} = b^* + k,$$
$$\overline{w}^* \bullet \overline{x} = b^* - k.$$

with points $(\overline{x}_p, +1)$ and $(\overline{x}_q, -1)$ in D such that

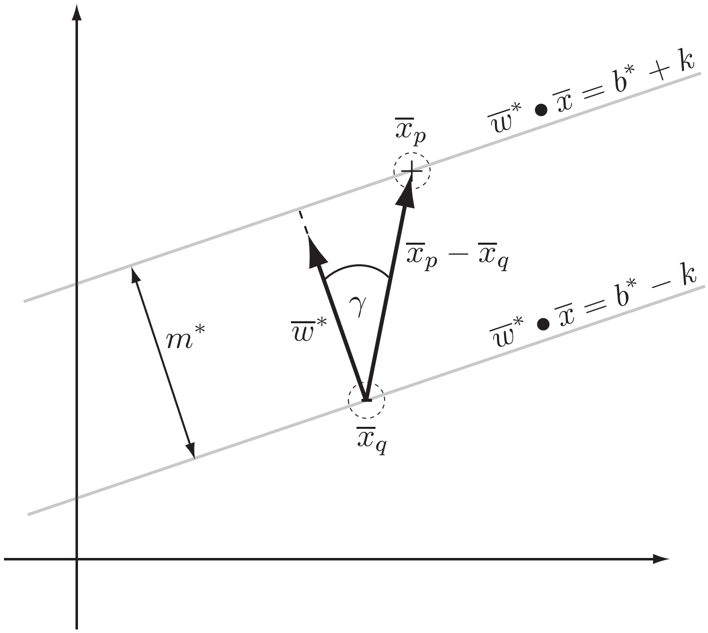
$$\overline{w}^* \bullet \overline{x}_p = b^* + k,$$
$$\overline{w}^* \bullet \overline{x}_q = b^* - k.$$


$\overline{w}^* \bullet \overline{x} = b^* + k$
$\overline{w}^* \bullet \overline{x} = b^*$
$\overline{w}^* \bullet \overline{x} = b^* - k$
$\overline{x}_p$
$\overline{x}_q$
m^*
+
+
+
-
-
-
-
-

Optimizing the Margin

The margin m^* is the projection of $\overline{x}_p - \overline{x}_q$ onto $\overline{w}^*$.

$$\begin{aligned} m^* &= |\overline{x}_p - \overline{x}_q| \cos\gamma \\ &= \frac{\overline{w}^* \bullet (\overline{x}_p - \overline{x}_q)}{|\overline{w}^*|} \\ &= \frac{\overline{w}^* \bullet \overline{x}_p - \overline{w}^* \bullet \overline{x}_q}{|\overline{w}^*|} \\ &= \frac{(b^* + k) - (b^* - k)}{|\overline{w}^*|} \\ &= \frac{2k}{|\overline{w}^*|}. \end{aligned}$$


$\overline{w}^* \bullet \overline{x} = b^* + k$
$\overline{x}_p$
$\overline{x}_p - \overline{x}_q$
γ
m^*
$\overline{w}^*$
$\overline{w}^* \bullet \overline{x} = b^* - k$
$\overline{x}_q$

This gives us the optimization expression,

$$m^* = \max \frac{2k}{|\overline{w}|}.$$

Optimizing the Margin

Expressing our maximization problem as a minimization problem,

$$\begin{aligned} m^* &= \max \frac{2k}{|\overline{w}|} \\ &= \min \frac{|\overline{w}|}{2k} \\ &= \min \frac{|\overline{w}|^2}{2k} \\ &= \min \frac{1}{2k} \overline{w} \bullet \overline{w} \\ &= \min \frac{1}{2} \overline{w} \bullet \overline{w}, \end{aligned}$$

with $k = 1$ (optimization problems are invariant under constants).

This gives us our objective function as,

$$\boxed{\phi(\overline{w}, b) = \frac{1}{2} \overline{w} \bullet \overline{w}.}$$

Note that the objective function is a convex function!

Optimizing the Margin

We can now derive the constraints for the optimization problem. For our optimal supporting hyperplanes the following identities hold,

$$\begin{aligned} \overline{w}^* \bullet \overline{x}_i &\geq b^* + k && \text{for all } (\overline{x}_i, y_i) \in D \text{ s.t. } y_i = +1, \\ \overline{w}^* \bullet \overline{x}_i &\leq b^* - k && \text{for all } (\overline{x}_i, y_i) \in D \text{ s.t. } y_i = -1. \end{aligned}$$

Taking our choice of $k = 1$ into account, then any feasible solution $\overline{w} \bullet \overline{x} = b$ has to fulfill the constraints,

$$\begin{aligned} \overline{w} \bullet \overline{x}_i &\geq 1 + b && \text{for all } (\overline{x}_i, y_i) \in D \text{ s.t. } y_i = +1, \\ \overline{w} \bullet (-\overline{x}_i) &\geq 1 - b && \text{for all } (\overline{x}_i, y_i) \in D \text{ s.t. } y_i = -1, \end{aligned}$$

or

$$\overline{w} \bullet (y_i \overline{x}_i) \geq 1 + y_i b \qquad \text{for all } (\overline{x}_i, y_i) \in D.$$

Maximum Margin Classifiers

Proposition: (Maximum Margin Classifier) Given a linearly separable training set

$$D = \{(\overline{x}_1, y_1), (\overline{x}_2, y_2), \ldots, (\overline{x}_l, y_l)\} \subseteq \mathbb{R}^n \times \{+1, -1\},$$

we can compute a maximum margin decision surface $\overline{w}^* \bullet \overline{x} = b^*$ with an optimization,

$$\min \phi(\overline{w}, b) = \min_{\overline{w}, b} \frac{1}{2} \overline{w} \bullet \overline{w}$$

subject to the constraints,

$$\overline{w} \bullet (y_i \overline{x}_i) \geq 1 + y_i b \qquad \text{for all } (\overline{x}_i, y_i) \in D.$$