Maximum Margin Classifiers

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

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

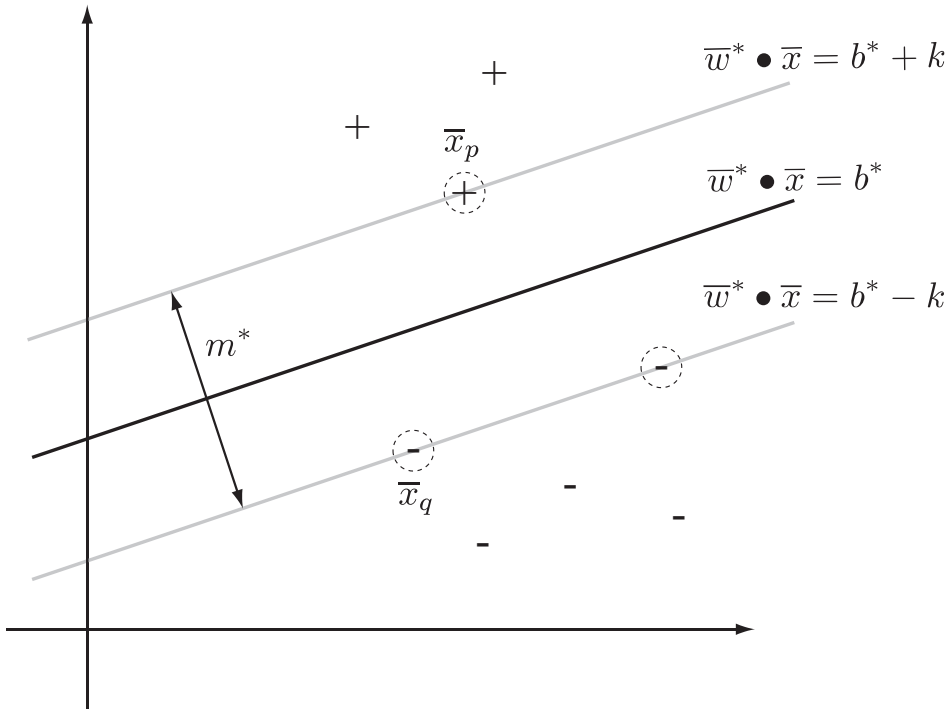
we can compute a maximum margin decision surface $\overline{w}^* \bullet \overline{x} = b^*$ with an optimization,

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

subject to the constraints,

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

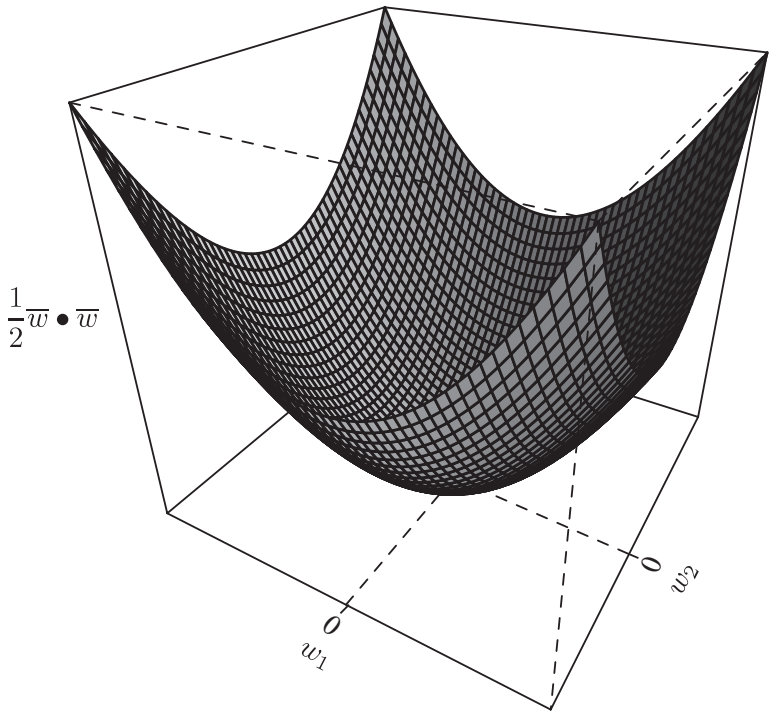
Maximum Margin Classifiers



Convex Optimization

Our objective function is convex,

$$\phi(\overline{w}, b) = \frac{1}{2}\overline{w} \bullet \overline{w} = \frac{1}{2}(w_1^2 + \ldots + w_n^2),$$



Here $\overline{w} \in \mathbb{R}^2$.

Quadratic Programming - QP

A *quadratic program* is a general convex optimization problem of the form

$$\overline{w}^* = \underset{\overline{w}}{\operatorname{argmin}} \left(\frac{1}{2} \overline{w}^T \mathbf{Q}\, \overline{w} - \overline{q} \bullet \overline{w} \right),$$

subject to the constraints

$$\mathbf{X}^T \overline{w} \geq \overline{c}.$$

Here, $\mathbf{Q}$ is an $n \times n$ matrix, $\mathbf{X}$ is an $l \times n$ matrix, the vectors $\overline{w}^*$, $\overline{w}$, $\overline{q}$ are n-dimensional vectors, and the vector $\overline{c}$ is an l-dimensional vector. [a]

In software packages this is usually given as function of the form,

$$\overline{w}^* = solve(\mathbf{Q}, \overline{q}, \mathbf{X}, \overline{c}).$$

[a] I have written the quadratic program in terms of $\overline{w}$; in the literature a different letter would typically be used for the optimization variable.

Quadratic Programming - QP

In order to bring the generalized quadratic program into a form that we can use for our maximum margin optimization we let

$$\mathbf{Q} = \mathbf{I},$$

and

$$\overline{q} = \overline{0},$$

then

$$\overline{w}^* = \underset{\overline{w}}{\operatorname{argmin}} \left(\frac{1}{2} \overline{w}^T \mathbf{I}\, \overline{w} - \overline{0} \bullet \overline{w} \right) = \underset{\overline{w}}{\operatorname{argmin}} \left(\frac{1}{2} \overline{w} \bullet \overline{w} \right),$$

Quadratic Programming - QP

Next, let us look at the original constraints,

$$(y_i \overline{x}_i) \bullet \overline{w} \geq 1 + y_i b,$$

for all $(\overline{x}_i, y_i) \in D$ with $i = 1, \ldots, l$ and $\overline{x}_i = (x_i^1 \ldots, x_i^n)$.

We have to rewrite these into the matrix form,

$$\mathbf{X}^T \overline{w} \geq \overline{c},$$

with

$$\mathbf{X} = \begin{pmatrix} y_1 x_1^1 & \cdots & y_i x_i^1 & \cdots & y_l x_l^1 \\ \vdots & & \vdots & & \vdots \\ y_1 x_1^n & \cdots & y_i x_i^n & \cdots & y_l x_l^n \end{pmatrix} \qquad \overline{c} = \begin{pmatrix} 1 + y_1 b \\ 1 + y_2 b \\ \vdots \\ 1 + y_l b \end{pmatrix}$$

Observation: b is now a free variable in the optimization problem, its value is not computed by the optimization algorithm but must be set by the user.

Quadratic Programming - QP

Proposition: (Quadratic Programming) Given a linearly separable training set

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

then we can compute a maximum margin decision surface $\overline{w}^* \bullet \overline{x} = b^*$ with a quadratic programming approach that solves the generalized optimization problem,

$$(\overline{w}^*, b^*) = \underset{\overline{w}, b}{\operatorname{argmin}} \left(\frac{1}{2} \overline{w}^T \mathbf{Q}\, \overline{w} - \overline{q} \bullet \overline{w} \right),$$

subject to the constraints

$$\mathbf{X}^T \overline{w} \geq \overline{c},$$

with $\mathbf{Q} = \mathbf{I}$, $\overline{q} = \overline{0}$, and where $\mathbf{X}$, and $\overline{c}$ are constructed according to the previous discussion.

QP - Algorithm

let $D = \{(\overline{x}_1, y_1), (\overline{x}_2, y_2), \ldots, (\overline{x}_l, y_l)\} \subset \mathbb{R}^n \times \{+1, -1\}$
$r \leftarrow \max\{|\overline{x}| \mid (\overline{x}, y) \in D\}$
$q \leftarrow 1000$
let $\overline{w}^*$ and b^* be undefined
construct $\mathbf{X}$
for each $b \in [-q, q]$ **do**
    construct $\overline{c}$
    $\overline{w} \leftarrow \mathit{solve}(\mathbf{I}, \overline{0}, \mathbf{X}, \overline{c})$
    if ($\overline{w}$ is defined **and** $\overline{w}^*$ is undefined) **or**
        ($\overline{w}$ is defined **and** $|\overline{w}| < |\overline{w}|^*$) **then**
        $\overline{w}^* \leftarrow \overline{w}$
        $b^* \leftarrow b$
    end if
end for
if $\overline{w}^*$ is undefined **then stop** constraints not satisfiable
else if $|\overline{w}|^* > q/r$ **then stop** bounding assumption of $|\overline{w}|$ violated
end if
return $(\overline{w}^*, b^*)$

QP - Example

Let

$$D = \{((1,6),-1),((3,7),-1),((1,4),+1),((2,1),+1)\}$$

be the training set and let

$$\overline{w}^* = solve(\mathbf{Q}, \overline{q}, \mathbf{X}, \overline{c}),$$

be a call to the solver, then

$$\mathbf{X} = \begin{bmatrix} -1 & -3 & 1 & 2 \\ -6 & -7 & 4 & 1 \end{bmatrix} \qquad \overline{c} = \begin{bmatrix} 1-b \\ 1-b \\ 1+b \\ 1+b \end{bmatrix}$$

$$\mathbf{Q} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \qquad \overline{q} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

QP - Free Parameter

Observation: Our optimization problem has a free parameter, the offset term b. Notice also that the constraints are dependent on this term. That is we need to pick b in such a way that the constraints are consistent. In other words, we need to pick b in such a way that the quadratic solver can actually find a minimum $\overline{w}$.

What are reasonable values for b to try?

QP - Free Parameter

First observation,

$$\begin{aligned} b = \overline{w} \bullet \overline{x} &\Rightarrow \\ b = |\overline{w}||\overline{x}| \cos\gamma &\Rightarrow \quad (-1 \leq \cos\gamma \leq 1, \text{for all } \gamma) \\ -|\overline{w}||\overline{x}| \leq b \leq |\overline{w}||\overline{x}| &\Rightarrow \quad (|\overline{x}| \leq r) \\ -|\overline{w}|r \leq b \leq |\overline{w}|r &\Rightarrow \end{aligned}$$

Second observation, $2r$ is the size of the largest margin and $\overline{w}$ is unbounded,

$$\frac{2}{|\overline{w}|} \leq 2r \Rightarrow \frac{1}{r} \leq |\overline{w}|$$

Third observation, we bound $\overline{w}$,

$$\frac{1}{r} \leq |\overline{w}| \leq \frac{q}{r}$$

Finally, we use $\overline{w} = q/r$ in the equation above,

$$-|\overline{w}|r \leq b \leq |\overline{w}|r \Rightarrow -q \leq b \leq q$$