## Soft margins

Hard margins:

$$\text{for } i = 1, \ldots, m, \qquad y_i(w'x_i + b) \geq 1$$

Soft margins:

$$\text{for } i = 1, \ldots, m, \qquad y_i(w'x_i + b) \geq 1 - \zeta_i \qquad \zeta_i \geq 0$$

**The primal problem:**

Let $C$ be a constant that corresponds to the "amount of allowed softness". The function to be minimized and the linear inequality constraints are augmented to:

$$\text{Minimize} \quad \frac{1}{2}|w|^2 + C\sum_{i=1}^{m}\zeta_i$$

subject to the $2m$ linear inequality constraints:

$$\text{for } i = 1, \ldots, m, \qquad y_i(w'x_i + b) \geq 1 - \zeta_i, \qquad \zeta_i \geq 0$$

Intuitively, large values of $C$ would emphasize the requirement that the $\zeta_i$ are small, and thus *decrease* the softness.

### Derivation of the dual problem

The Lagrangian of the primal problem:

$$\begin{aligned} L(w, b, \zeta_1, \ldots, \zeta_m, \alpha_1, \ldots, \alpha_m, r_1, \ldots, r_m) &= \frac{1}{2}|w|^2 + C\sum_{i=1}^{m}\zeta_i \\ &\quad + \sum_{i=1}^{m}\alpha_i(1 - \zeta_i - y_i(w'x_i + b)) - \sum_{i=1}^{m} r_i\zeta_i \end{aligned} \tag{1}$$

To compute the dual problem we need to minimize $L$ with respect to $w, b, \zeta_i$ so that it is a function of only $\alpha_i, r_i$.

$$\begin{array}{ll} \text{The derivative of } L \text{ w.r.t. } w \text{ gives:} & w = \sum_{i=1}^{m}\alpha_i y_i x_i \\ \text{The derivative of } L \text{ w.r.t. } b \text{ gives:} & \sum_{i=1}^{m}\alpha_i y_i = 0 \\ \text{The derivative of } L \text{ w.r.t. } \zeta_i \text{ gives:} & C - \alpha_i - r_i = 0 \end{array} \tag{2}$$

Substituting these values in $L$ and simplifying we get::

$$L(\alpha_1, \ldots, \alpha_m, r_1, \ldots, r_m) = -\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j x_i' x_j + \sum_{i=1}^{m}\alpha_i$$

This is exactly the same dual function as in the hard-margins case. For the dual problem we also need the last two constraints in (2), and $\alpha_i \geq 0, r_i \geq 0$. The difference between the hard and the soft case is that from the third equation in (2) and the condition $r_i \geq 0$ we have: $\alpha_i \leq C$.

**The dual problem:**

$$\text{Maximize} \quad L(\alpha_1,\ldots,\alpha_m) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} \alpha_i\alpha_j y_i y_j x_i' x_j$$

$$\text{subject to:} \quad 0 \le \alpha_i \le C, \quad \sum_{i=1}^{m} \alpha_i y_i = 0$$

This is a quadratic programming problem and we assume that there is a black-box that solves it. The solution gives the values of the $\alpha_i$.

**The Karush-Kuhn-Tucker Complementary Conditions**

In this case the KKT condition gives:

$$\alpha_i \left( y_i(w'x_i + b) - 1 + \zeta_i \right) = 0$$
$$\zeta_i(\alpha_i - C) = 0$$

From the second condition it follows that either $\zeta_i = 0$, or $\alpha_i = C$. Therefore:

| | | | |
|---|---|---|---|
| $\alpha_i = 0$ | $\rightarrow$ | not support vector | |
| $0 < \alpha_i < C$ | $\rightarrow$ | $y_i(w'x_i + b) = 1$ | point on hard margin |
| $\alpha_i = C$ | $\rightarrow$ | $y_i(w'x_i + b) = 1 - \zeta_i$ | point on soft margin |

**Recovering $w, b$**

From (2) it follows that $w$ can be recovered from the support vectors in the same way as in the hard-margins case:

$$w = \sum_{j=1}^{k} \alpha_j y_j x_j \tag{3}$$

Once $w$ is determined the value of $b$ can be computed from any one of the hard margins support vectors (with $\alpha_i < C$), using the same formulas as in the hard-margins case:

$$0 < \alpha_s < C \quad \rightarrow \quad b = \frac{1}{y_s} - w'x_s \tag{4.1}$$

As in the hard-margins case it is also possible to compute the value of $b$ from all support vectors on the hard margins (satisfying $0 < \alpha_s < C$). Since the formulas for $w, b$ are the same as in the hard-margins case we can also use kernels.

**The value of $\zeta$**

In the hard-margins case the dual optimization problem can give infinite values, indicating that the primal problem has no solution (the data is not linearly separable.) This cannot happen in the soft-margins case. If the point $i$ is wrongly classified by the hyperplane then we can always choose $\zeta_i = 1 - y_i(w'x_i + b)$, since this gives $\zeta \ge 0$ (in fact it gives $\zeta \ge 1$). If the point $i$ is correctly classified but with distance from the margins that is too short, we can still choose $zeta_i = 1 - y_i(w'x_i + b)$, since we would still have $\zeta \ge 0$. The case in which $\zeta < 0$ corresponds to points that are correctly classified with $y_i(w'x_i + b) \ge 1$, and they are not inside the soft margins.

## Example

| $i$ | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| $x_i$ | 0 | 1 | 2 | 3 | 4 |
| $y_i$ | -1 | -1 | 1 | -1 | 1 |
| Lagrangian multiplier | $\alpha_0$ | $\alpha_1$ | $\alpha_2$ | $\alpha_3$ | $\alpha_4$ |

The dual problem:

$$\begin{aligned}
\text{maximize} \quad L(\alpha_0,\ldots,\alpha_4) &= \alpha_0+\alpha_1+\alpha_2+\alpha_3+\alpha_4 \\
&-\frac{1}{2}(\alpha_1^2+4\alpha_2^2+9\alpha_3^2+16\alpha_4^2 \\
&\quad -4\alpha_1\alpha_2+6\alpha_1\alpha_3-8\alpha_1\alpha_4-12\alpha_2\alpha_3+16\alpha_2\alpha_4-24\alpha_3\alpha_4) \\
&= \alpha_0+\alpha_1+\alpha_2+\alpha_3+\alpha_4-\frac{1}{2}(-\alpha_1+2\alpha_2-3\alpha_3+4\alpha_4)^2 \\
\text{subject to:} \ \ & 0\le\alpha_0\le C,\quad 0\le\alpha_1\le C,\quad 0\le\alpha_2\le C,\quad 0\le\alpha_3\le C,\quad 0\le\alpha_4\le C, \\
& -\alpha_0-\alpha_1+\alpha_2-\alpha_3+\alpha_4=0
\end{aligned}$$

With $C=10$ the solution (computed by the black box quadratic optimizer) is: $\alpha_0=0$, $\alpha_1=\alpha_4=3.55$, $\alpha_2=\alpha_3=10$. Therefore, the support vectors are $x_1,x_2,x_3,x_4$.
We can now compute $w$ from (3):

$$w=-3.55+20-30+4*3.55=0.66666$$

The value of $b$ can be computed, for example, from $x_1$, the first support vector, using (4.1):

$$b=-1-0.666=-1.666$$

It cannot be computed from $x_2,x_3$ since they satisfy $\alpha_i=C$. It can be computed from $x_4$: using (4.1):

$$b=1-0.6666\times 4=-1.6666$$

Observe that in this case $x_2,x_3$ are wrongly classified.

## Distances

In our case the "hyperplane" is the point satisfying $w'x+b=0$, which is $x=2.5$. The distance of the hard-margins support vectors from that hyperplane is 1.5. Observe that $1.5/|w|=1$, as expected. The $\zeta$ value for $x_2$ is $1-(-1/3)=4/3$. Its distance from the hyperplane is $(1-\zeta)/|w|=-1/2$.