SUPPORT VECTOR MACHINES

Optimal Soft Margin Hyperplane

This linear classifier is given by the solution to

$$(OSM) \quad \min_{w,b,\xi} \quad \frac{1}{2}\|w\|^2 + \frac{C}{n}\sum_{i=1}^{n}\xi_i$$

$$\text{s.t.} \quad y_i\left(\langle w, x_i\rangle + b\right) \geq 1-\xi_i, \quad i=1,\dots,n$$

$$\xi_i \geq 0, \quad i=1,\dots,n$$

The SVM is a kernelized version of this method, that is, it uses inner product kernels to extend to a nonlinear classification method. To kernelize the method, we must first express the classifier so that feature vectors are only involved via inner products. To do this we will show that the dual of (OSM) has this property, and that the solution of (OSM) can be recovered from the dual solution.

The Optimal Soft Margin Dual

The Lagrangian is

$$L(w,b,\xi,\alpha,\beta) = \frac{1}{2}\|w\|^2 + \frac{C}{n}\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i\left(y_i\left(w^T x_i + b\right) - 1 + \xi_i\right) - \sum_{i=1}^{n}\beta_i\xi_i$$

The dual problem is

$$\max_{\alpha\geq 0,\ \beta\geq 0} \quad L_D(\alpha,\beta)$$

where

$$L_D(\alpha,\beta) = \min_{w,b,\xi} \quad L(w,b,\xi,\alpha,b)$$

This is an unconstrained minimization with a convex, differentiable objective function. Therefore, for fixed α,β, the minimizing w,b,ξ satisfy

$$\frac{\partial L}{\partial w} = w - \sum \alpha_i y_i x_i = 0$$

$$\frac{\partial L}{\partial w} = w - \sum \alpha_i y_i x_i = 0$$

$$\frac{\partial L}{\partial b} = \sum \alpha_i y_i = 0$$

$$\frac{\partial L}{\partial \xi_i} = \frac{c}{n} - \alpha_i - \beta_i = 0 \qquad \forall i$$

Therefore

$$L_D(\alpha, \beta) = -\frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j + \sum_i \alpha_i$$

and the dual problem is equivalent to

$$\max_{\alpha, \beta} \quad -\frac{1}{2} \sum \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle + \sum_{i=1}^{n} \alpha_i$$

$$\text{s.t.} \quad \sum \alpha_i y_i = 0$$

$$\alpha_i + \beta_i = \frac{c}{n} \qquad \forall i$$

$$\alpha_i \geq 0, \ \beta_i \geq 0 \qquad \forall i$$

We can further eliminate β to obtain

$$\text{(OSM-D)} \quad \max_{\alpha} \quad -\frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle + \sum_i \alpha_i$$

$$\text{s.t.} \quad \sum \alpha_i y_i = 0$$

$$0 \le \alpha_i \le \frac{C}{n} \qquad \forall i$$

Note that the dual is another quadratic program (QP).

Problem (OSM) is convex with affine constraints, and therefore strong duality holds. Since the objective and constraint functions are differentiable, the KKT necessity theorem says that the solutions of the primal and dual problems are related by the KKT conditions.

This allows us to make the following observations. Let w^*, b^*, ξ^* denote a primal solution, and α^*, β^* a dual solution.

1. We can obtain w^* and b^*, which define the classifier, from α^*.

   (a) From the first KKT condition,

$$w^* = \sum_{i=1}^{n} \alpha_i^* y_i x_i$$

so the optimal normal vector is a linear combination of data points.

(b) We'll discuss recovery of b^* shortly.

2. From complementary slackness,

$$\alpha_i^* \left(1 - \xi_i^* - y_i (w^{*T} x_i + b^*) \right) = 0.$$

If x_i satisfies

$$y_i (w^{*T} x_i + b^*) = 1 - \xi_i^*,$$

we call x_i a support vector. Therefore, if x_i is not a support vector, then $\alpha_i^* = 0$.

This means w^* depends only on the SVs:

$$w^* = \sum_{\substack{\text{support} \\ \text{vectors}}} \alpha_i^* y_i x_i$$

There is a geometric interpretation:

(a) If $y_i (w^{*T} x_i + b^*) > 1$, then $\xi_i^* = 0$ and x_i is not a support vector.

(b) If $y_i(w^{*T}x_i + b) = 1$, then $\xi_i^* = 0$ and x_i is a SV.

(c) If $0 \leq y_i(w^{*T}x_i + b^*) < 1$, then $\xi_i^* > 0$ and x_i is a SV.

(d) If $y_i(w^{*T}x_i + b) < 0$, then $\xi_i^* > 0$ and x_i is a SV.

Based on which case occurs, we say

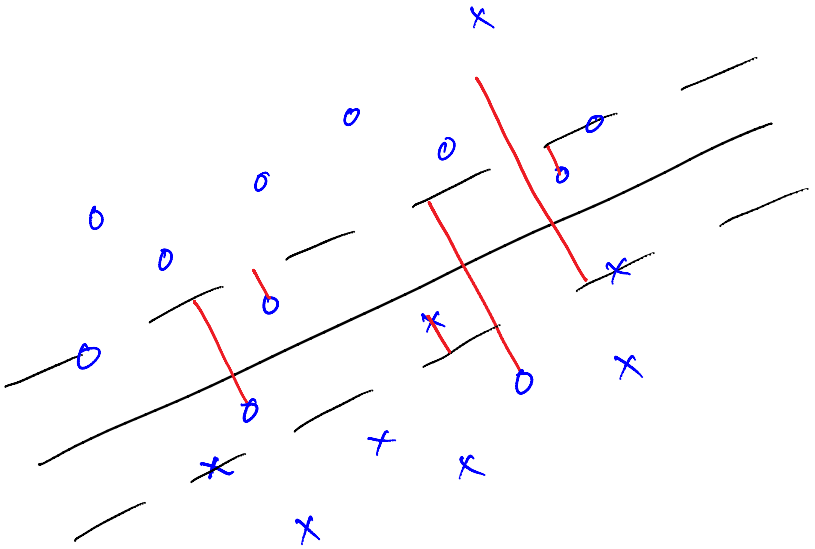
(a) x_i is outside the margin

(b) x_i is on the margin

(c) x_i is within the margin

(d) x_i is misclassified

Cases (b) – (d) are called margin errors, and here we have $\xi_i^* = 1 - y_i(w^{*T}x_i + b^*)$. It can be shown (homework problem) that $\xi_i^* \propto$ distance from x_i to the margin. This leads to the following picture.



The stems correspond to margin errors/support vectors.

3. If $\alpha_i^* < \frac{C}{n}$, then $\xi_i^* = 0$. This follows from applying complementary slackness to the constraint $\xi_i \geq 0$, which tells us $\beta_i^* \xi_i^* = 0$.

If $\alpha_i^* < \frac{C}{n}$, then $\beta_i^* > 0$, so $\xi_i^* = 0$.

This insight gives us a way to determine b^*.

Consider any i s.t. $0 < \alpha_i^* < \frac{C}{n}$. Then

$$y_i (w^{*T} x_i + b^*) = 1 - \xi^* = 1$$

Solve for b^* (recall $y_i = \pm 1$ so $y_i^2 = 1$) to get

$$b^* = y_i - w^{*T} x_i.$$

In practice, it is common to average over several such i to counter numerical errors.

Support Vector Machines

The dual QP and final classifier only involve the data via inner products. Therefore we can kernelize the optimal soft margin hyperplane. The resulting classifier is known as a support vector machine.

Let k be an inner product kernel.

The SVM classifier is

$$f(x) = \text{sign}\left\{ \sum_{i=1}^{n} \alpha_i^* y_i k(x, x_i) + b^* \right\}$$

where $\alpha^* = (\alpha_1^*, \ldots, \alpha_n^*)$ is the solution of

$$\max_{\alpha} \; -\frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j k(x_i, x_j) + \sum \alpha_i$$

$$\text{s.t.} \quad \sum \alpha_i y_i = 0$$

$$0 \le \alpha_i \le \frac{C}{n}, \quad i = 1, \ldots, n$$

and b^* is given by

$$b^* = y_i - \sum_{j=1}^{n} \alpha_j^* y_j k(x_j, x_i)$$

for some i such that $0 < \alpha_i^* < \frac{C}{n}$.

Remarks

- The final classifier depends only on those training data points that are support vectors.
- The size of the dual (i.e., the number of variables) is n, and in particular it is independent of the output dimension of the feature map Φ associated with k, which could be infinite.
- The soft-margin hyperplane was the first machine learning algorithm to be kernelized.

The SMO Algorithm

SMO stands for sequential minimal optimization. In the

context of SVMs, it is an algorithm to efficiently solve the SVM dual.

The basic conceptual strategy is that of coordinate ascent.

For the moment consider a general constrained opt. problem

$$\max_{\alpha} \quad f(\alpha)$$
$$\text{s.t.} \quad \alpha \in S$$

where S is the feasible set. Coordinate ascent does

Initialize $\alpha^0 = (\alpha_1^0, \ldots, \alpha_n^0) \in S$

$t \leftarrow 0$

Repeat

$t \leftarrow t+1$

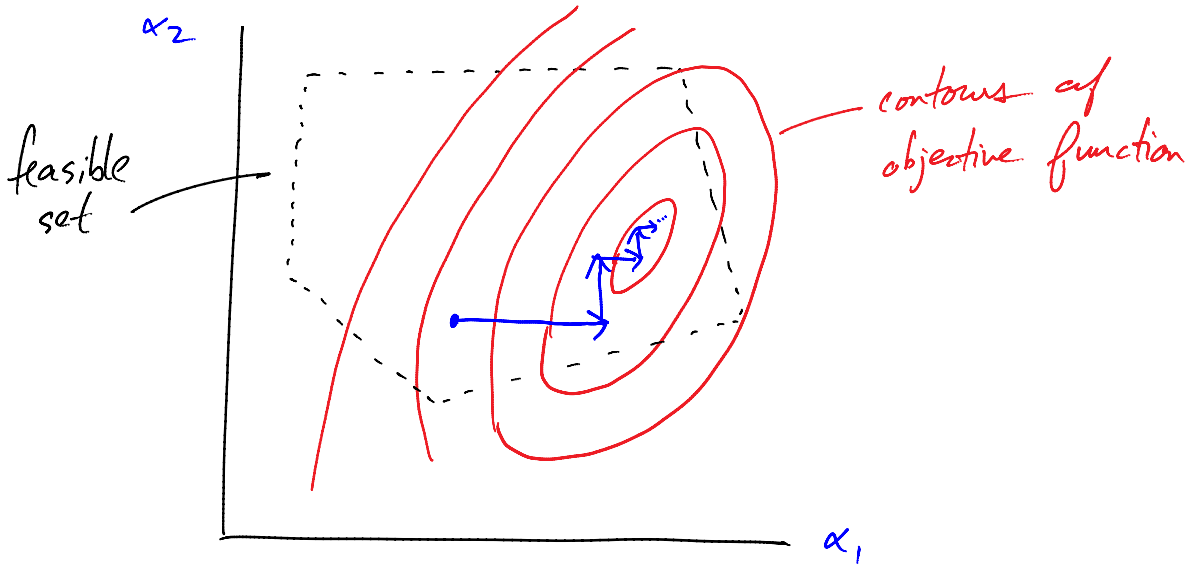
For $i = 1, \ldots, n$

$$\alpha_i^t = \arg\max_{\alpha'} f(\alpha_1^t, \ldots, \alpha_{i-1}^t, \alpha', \alpha_{i+1}^{t-1}, \ldots, \alpha_n^{t-1})$$

End

Until converged

If $n=2$ we have the following picture:



However, we can't apply CA to the SVM as stated above. The reason is that the constraint $\sum \alpha_i y_i = 0$ determines each α_i as a function of the others, namely

$$\alpha_i = y_i \left(-\sum_{j \neq i} \alpha_j y_j \right)$$

where we used $y_i^2 = 1$. So instead, SMO updates two variables at a time:

Initialize α^0 (any feasible pt, e.g., the zero vector)

Repeat

- Select α_i, α_j to be updated
- Update α_i, α_j by solving the SVM dual QP,

- Update α_i, α_j by solving the SVM dual QP, holding all other α_k, $k \neq i, j$, fixed.

Until termination criterion satisfied.

The first step is usually accomplished with a heuristic that estimates which two variables will lead to the largest increase in the objective.

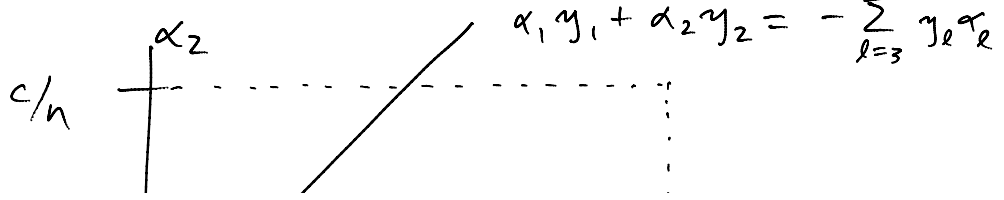
The whole point of SMO is that the second step can be performed efficiently. Suppose $\alpha_3, \ldots, \alpha_n$ are fixed. We need to solve

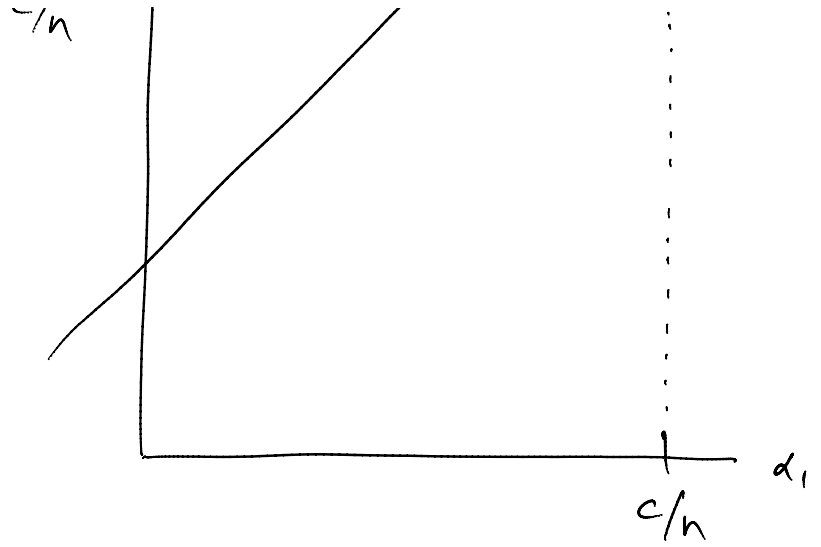
$$\max_{\alpha_1, \alpha_2} \; -\frac{1}{2}\left[\alpha_1^2 k_{11} + \alpha_2^2 k_{22} + 2\alpha_1\alpha_2 y_1 y_2 k_{12}\right] + c_1\alpha_1 + c_2\alpha_2$$

$$\text{s.t.} \quad \alpha_1 y_1 + \alpha_2 y_2 = -\sum_{l=3}^{n} y_l \alpha_l$$

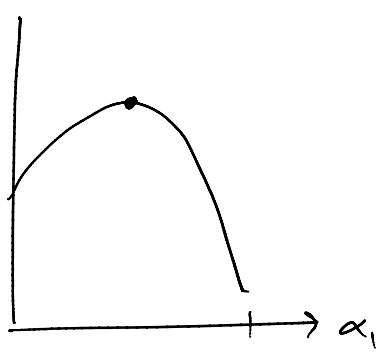
$$0 \leq \alpha_1, \alpha_2 \leq \frac{C}{n}$$

where c_1 and c_2 are constants depending on $\alpha_3, \ldots, \alpha_n$, and $k_{ij} = k(x_i, x_j)$.

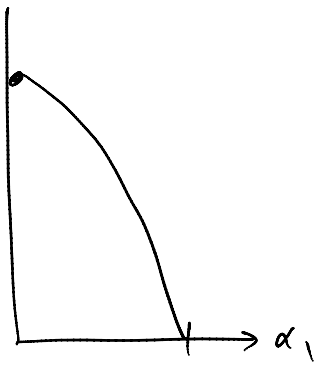




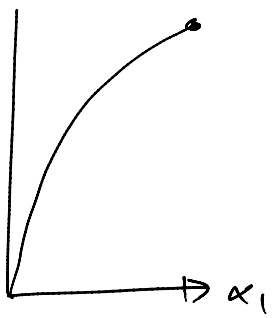
We can use the linear equation to solve for α_2 in terms of α_1, and then solve the resulting QP for α_1. The objective is just a parabola, so it's easy: The max occurs either at the critical point of the parabola, or the boundary of the feasible range for α_1.



Case 1



Case 2



Case 3

The SMO algorithm converges to the global optimum.

The computational complexity is $O(n^3)$ worst case,

but often more like $O(n^2)$ typically.

For more on SMO, see the paper by Platt (1999).