SUPPORT VECTOR MACHINES

Optimal Soft-Margin Hyperplanes

This linear classifier is the solution of

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

$$\text{s.t.} \quad y_i(w^T x_i + b) \geq 1 - \xi_i, \quad i = 1, \ldots, n$$

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

The SVM is a "kernelized" version of this method, that is, it uses inner product kernels to extend the method to a nonlinear classifier. However, before we can apply the kernel trick, we need to first express the classifier so that feature vectors are only involved via inner products.

The Soft Margin Dual

The Lagrangian is

(A) $$L(w, b, \xi, \alpha, \beta) =$$

Since the optimization problem is convex & differentiable, the KKT conditions are necessary and sufficient conditions for primal / dual optimality (with zero duality gap).

The dual is

$$\max_{\substack{\alpha,\beta \\ \alpha_i,\beta_i \geq 0}} \left(\min_{w,b,\xi} \quad L(w, b, \xi, \alpha, \beta) \right)$$

For fixed α, β, the minimizing w, b, ξ will satisfy

(B) $$\frac{\partial L}{\partial w} =$$

$$\frac{\partial L}{\partial b} =$$

$$\frac{\partial L}{\partial \xi_i} =$$

Plugging these in we obtain the dual function

(c) $L_D(\alpha, \beta) =$

Therefore, the optimal dual variables (α^*, β^*) are the solution of

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

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

$$\alpha_i + \beta_i = \frac{c}{n} \qquad i = 1, \ldots, n$$

$$\alpha_i \geq 0, \ \beta_i \geq 0, \qquad i = 1, \ldots, n$$

We can eliminate β to obtain the <u>soft-margin dual QP</u>

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

$$\text{s.t.} \quad \sum_{i=1}^{n} \alpha_i y_i = 0$$

$$0 \leq \alpha_i \leq \frac{c}{n}, \qquad i = 1, \ldots, n.$$

The dual has a number of desirable properties:

1. We may obtain w^*, b^* from α^*.

   (a) From the KKT conditions we have

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

   $\Rightarrow$ the optimal normal vector is a linear combo. of data points

   (b) Recovering b^* is a little less obvious. We'll return to this later.

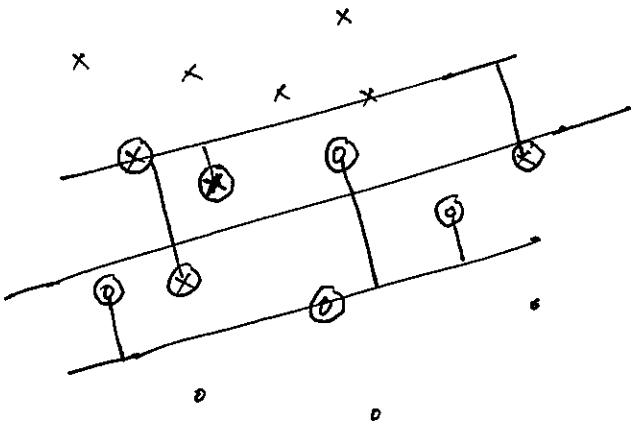
2. From the KKT conditions we have

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

Recall that x_i for which

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

(D) are called _____. These are the points on or inside the margin of separation



By the KKT condition, either x_i is a support vector, or $\alpha_i^* = 0$

extremely important fact!

Conclusion | We may write

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

It has been widely demonstrated empirically that only a small fraction of the training patterns are support vectors (those that are closest to the decision boundary).

Therefore, the soft-margin criterion produces a hyperplane with a _____ representation. This is advantageous for efficient storage and evaluation.

(E)

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

To see this recall the Lagrange multipliers β_i corresponding to the constraints $\xi_i \geq 0$. By the KKT conditions, we have that

$$\beta_i^* \cdot \xi_i^* = 0.$$

Since $\alpha_i^* + \beta_i^* = \frac{C}{n}$, the claim follows.

Exercise | Suggest a procedure for determining b^* using 2. and 3. above.

Solution | If $0 < \alpha_i^* < \frac{C}{n}$, then

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

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

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

4. The dual QP and classifier only involve

(F) feature vectors through _____ _____

$\Rightarrow$ can apply kernel trick.

Support Vector Machines

Let k be an IP 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 α_i^* is the solution of

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

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

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

and b^* is given by

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

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

Remarks

- The final classifier depends only on those x_i with $\alpha_i > 0$, i.e. the (6) _____ _____.

- The size of the dual QP is n, independent of k, $\mathcal{I}$, or $\mathcal{H}$. This is remarkable since the dimension of $\mathcal{H}$ may be _____.

- The soft-margin hyperplane was the first machine learning algorithm to be "kernelized." Since then, the idea has been applied to many other algorithms, e.g., kernel ridge regression, kernel PCA

Sequential Minimal Optimization

How can we compute the solution of

$$\max_{\alpha} \; -\frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j k_{ij} + \sum \alpha_i$$

$$\text{s.t.} \quad \sum \alpha_i y_i = 0, \quad 0 \le \alpha_i \le \frac{C}{n}, \; i = 1, \ldots, n$$

[where $k_{ij} = k(x_i, x_j)$]

efficiently? There are several general approaches to solving quadratic programs, and many have been applied to solving the SVM dual. A very efficient solver that capitalizes on the structure of the SVM dual constraints is the _____ algorithm. (Platt, 1999)

(H)

The SMO algorithm is an example of a _____ algorithm.

Algorithm

Initialize $\alpha = (\alpha_1, \ldots, \alpha_n)$

Repeat

(1) Select i, j, $1 \le i, j \le n$

(2) Update α_i and α_j by optimizing dual QP, holding all other α_k, $k \neq i, j$, fixed

Until termination criterion satisfied.

In step (2), we choose α_i, α_j to solve

(I)

where

$$c_i =$$

$$c_j =$$

$$\vdots$$

The reason for decomposing to a 2-variable subproblem is that this subproblem can be

(J) solved _____.

Several strategies have been proposed for selecting i and j, typically based on heuristics that predict which pair of variables will lead to the largest change in the objective function.

The SMO algorithm can be shown to converge to the _____ optimum.

The running time is $O(n^3)$ worst case, but often more like $O(n^2)$ in practice.

References

Platt, 1999

Osuna, Freund, and Girosi, 1997

Schölkopf and Smola, 2002.

Andrew Ng's notes

Key

A. $L(w, b, \xi, \alpha, \beta) = \frac{1}{2} w^T w + \frac{C}{n} \sum \xi_i$

$- \sum_{i=1}^{n} \alpha_i \left(y_i (w^T x_i + b) - 1 + \xi_i \right) - \sum_{i=1}^{n} \beta_i \xi_i$

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

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

$\frac{\partial L}{\partial \xi_i} = \frac{C}{n} - \alpha_i - \beta_i = 0$

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

D. support vectors E. sparse F. inner products

G. support vectors, infinite H. SMO, decomposition

I. $\max_{\alpha_i, \alpha_j} \; -\frac{1}{2} \left[\alpha_i^2 k_{ii} + \alpha_j^2 k_{jj} + 2 \alpha_i \alpha_j y_i y_j k_{ij} \right] + c_i \alpha_i + c_j \alpha_j$

s.t. $\alpha_i y_i + \alpha_j y_j = - \sum_{\ell \neq i,j} \alpha_\ell y_\ell$

$0 \leq \alpha_i, \alpha_j \leq \frac{C}{n}$

$c_i = 1 - \frac{1}{2} \sum_{\ell \neq i,j} \alpha_\ell y_\ell y_i k_{i\ell}$, c_j similar

J. exactly, global