Constrained Optimization – Dual Problem

$\frac{w^T x + b}{\|w\|} \geq \gamma$

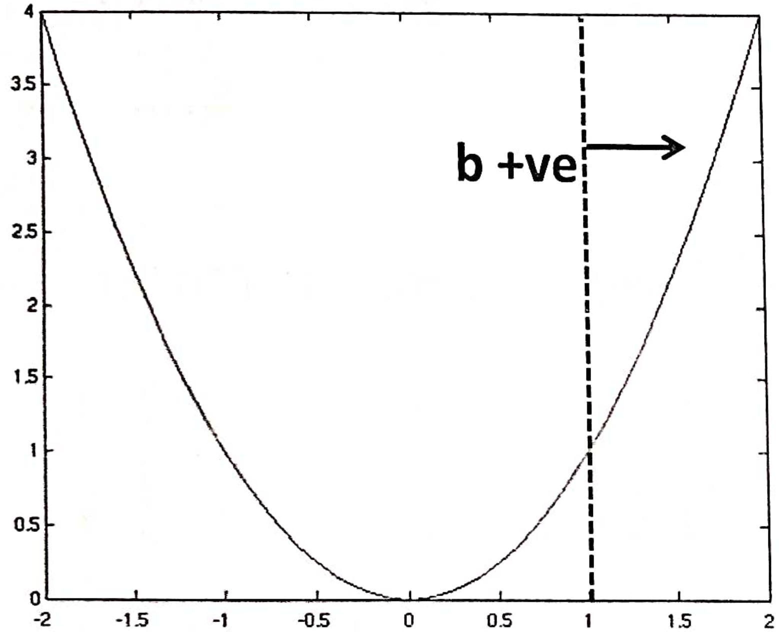
$\Rightarrow \frac{w^T x + b}{\|w\|} \leq -\gamma$

$w^T x + b \leq -\gamma\|w\| \leq -1$

$-\gamma\|w\| \leq -1$

$\gamma\|w\| \geq 1$

$\gamma \geq \frac{1}{\|w\|}$



$$x^* = b$$

Primal problem:

$$\min_x \; x^2$$
$$\text{s.t.} \quad x \geq b$$

Moving the constraint to objective function
Lagrangian:

$$L(x, \alpha) = x^2 - \alpha(x - b)$$
$$\text{s.t.} \quad \alpha \geq 0$$

Dual problem:

$$\max_\alpha \; d(\alpha) \rightarrow \min_x L(x, \alpha)$$
$$\text{s.t.} \quad \alpha \geq 0$$

Connection between Primal and Dual

Primal problem: p* = $\min_x \; x^2$
s.t. $x \geq b$

Dual problem: d* = $\max_\alpha \; d(\alpha)$
s.t. $\alpha \geq 0$

- **Weak duality:** The dual solution d* lower bounds the primal solution p* i.e. d* ≤ p*

Duality gap = p*-d*

- **Strong duality:** d* = p* holds often for many problems of interest e.g. if the primal is a feasible convex objective with linear constraints (Slater's condition)

Solving the dual

Solving:

$$\max_\alpha \min_x \overbrace{x^2 - \alpha(x-b)}^{L(x,\alpha)}$$
$$\text{s.t.} \quad \alpha \geq 0$$

Find the dual: Optimization over x is unconstrained.

$$\frac{\partial L}{\partial x} = 2x - \alpha = 0 \Rightarrow x^* = \frac{\alpha}{2} \qquad L(x^*, \alpha) = \frac{\alpha^2}{4} - \alpha\left(\frac{\alpha}{2} - b\right)$$
$$= -\frac{\alpha^2}{4} + b\alpha$$

Solve: Now need to maximize L(x*,α) over α ≥ 0

Solve unconstrained problem to get α′ and then take max(α′,0)

$$\frac{\partial}{\partial \alpha} L(x^*, \alpha) = -\frac{\alpha}{2} + b \ \Rightarrow \alpha' = 2b$$

$$\Rightarrow \alpha^* = \max(2b, 0) \qquad \Rightarrow x^* = \frac{\alpha^*}{2} = \max(b, 0)$$

α = 0 constraint is inactive, α > 0 constraint is active (tight)

Dual SVM – linearly separable case

n training points, d features $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ where x_i is a d-dimensional vector

- Primal problem: minimize$_{\mathrm{w},b}$ $\frac{1}{2}\mathrm{w}.\mathrm{w}$

  $\left(\mathrm{w}.\mathrm{x}_j + b\right) y_j \geq 1, \ \forall j$

w - weights on features (d-dim problem)

- Dual problem (derivation):

$$L(\mathrm{w}, b, \alpha) = \frac{1}{2}\mathrm{w}.\mathrm{w} - \sum_j \alpha_j \left[\left(\mathrm{w}.\mathrm{x}_j + b\right) y_j - 1\right]$$
$$\alpha_j \geq 0, \ \forall j$$

α - weights on training pts (n-dim problem)

Dual SVM – linearly separable case

- Dual problem (derivation):

$$\max_{\alpha} \min_{\mathbf{w},b} L(\mathbf{w}, b, \alpha) = \frac{1}{2}\mathbf{w}.\mathbf{w} - \sum_j \alpha_j \left[\left(\mathbf{w}.\mathbf{x}_j + b\right) y_j - 1\right]$$

$\alpha_j \geq 0, \ \forall j$

$$\frac{\partial L}{\partial \mathbf{w}} = 0 \qquad \Rightarrow \mathbf{w} = \sum_j \alpha_j y_j \mathbf{x}_j$$

$$\frac{\partial L}{\partial b} = 0 \qquad \Rightarrow \sum_j \alpha_j y_j = 0$$

If we can solve for αs (dual problem), then we have a solution for **w**,b (primal problem)

Dual SVM – linearly separable case

- Dual problem:

$$\max_{\alpha} \min_{\mathbf{w},b} L(\mathbf{w}, b, \alpha) = \frac{1}{2}\mathbf{w}.\mathbf{w} - \sum_j \alpha_j \left[\left(\mathbf{w}.\mathbf{x}_j + b\right) y_j - \right.$$

$$\alpha_j \geq 0, \ \forall j$$

$$\Rightarrow \mathbf{w} = \sum_j \alpha_j y_j \mathbf{x}_j \qquad\qquad \Rightarrow \sum_j \alpha_j y_j = 0$$

Dual SVM – linearly separable case

$$\text{maximize}_\alpha \quad \sum_i \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i\alpha_j y_i y_j \mathbf{x}_i.\mathbf{x}_j$$

$$\sum_i \alpha_i y_i = 0$$
$$\alpha_i \geq 0$$

Dual problem is also QP

Solution gives α_js $\longrightarrow$

$$\boxed{\mathbf{w} = \sum_i \alpha_i y_i \mathbf{x}_i}$$

What about b?

1. x_i with non-zero α_i are called SV.
2. The decision boundary is determined only by the SV.

Dual SVM – linearly separable case

3. Let t_j ($j=1,2,\cdots,s$) be the indices of the SVs. Then $w = \sum_{j=1}^{s} \alpha_{t_j} y_{t_j} \cdot x_{t_j}$

$$\text{maximize}_\alpha \quad \sum_i \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i\alpha_j y_i y_j \mathbf{x}_i.\mathbf{x}_j$$

$$\sum_i \alpha_i y_i = 0$$

$$\alpha_i \geq 0$$

$y = w^T x + b$
$\Rightarrow b = y - w^T x$
For any $\alpha_k > 0$, $(x_k, y_k) \Rightarrow$
$b = y_k - w^T x_k$

Dual problem is also QP

Solution gives α_js $\longrightarrow$

$$\boxed{\begin{gathered}\mathbf{w} = \sum_i \alpha_i y_i \mathbf{x}_i \\[1em] b = y_k - \mathbf{w}.\mathbf{x}_k \\ \text{for any } k \text{ where } \alpha_k > 0\end{gathered}}$$

Use any one of support vectors with α_k>0 to compute b since constraint is tight $(w.x_k + b)y_k = 1$

& If $y_k = -1$ then $w.x_k + b = -1$
If $y_k = 1$ then $w.x_k + b = 1$
$\Rightarrow w.x_k + b = y_k \Rightarrow b = y_k - w.x_k$ for any k where $\alpha_k > 0$

Dual formulation only depends on dot-products, not on w!

$\Phi(x_i)$ map x_i to $\Phi(x_i)$

$K(x_i, x_j) = \Phi(x_i) \cdot \Phi(x_j)$.

$$\text{maximize}_\alpha \quad \sum_i \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j \underbrace{\mathbf{x}_i . \mathbf{x}_j}$$

$$\sum_i \alpha_i y_i = 0$$

$$C \geq \alpha_i \geq 0 \longrightarrow \text{Regularazition Parameter } C$$

$$\Downarrow$$

$$\text{maximize}_\alpha \quad \sum_i \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j \underbrace{K(\mathbf{x}_i, \mathbf{x}_j)}$$

$$K(\mathbf{x}_i, \mathbf{x}_j) = \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j)$$

$$\sum_i \alpha_i y_i = 0$$

$$C \geq \alpha_i \geq 0$$

$\Phi(\mathbf{x})$ – High-dimensional feature space, but never need it explicitly as long as we can compute the dot product fast using some Kernel K

Dot Product of Polynomials

$\Phi(\mathbf{x}) =$ polynomials of degree exactly d

$K(x,z) = \Phi(x) \cdot \Phi(z)$

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad \mathbf{z} = \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}$$

d=1 $\Phi(\mathbf{x}) \cdot \Phi(\mathbf{z}) = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \cdot \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = x_1 z_1 + x_2 z_2 = \mathbf{x} \cdot \mathbf{z}$

$X = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$; $\Phi(X) = \begin{bmatrix} x_1^2 \\ \sqrt{2}\, x_1 x_2 \\ x_2^2 \end{bmatrix}$

$x \xrightarrow{\Phi} \Phi(x)$

d=2 $\Phi(\mathbf{x}) \cdot \Phi(\mathbf{z}) = \begin{bmatrix} x_1^2 \\ \sqrt{2} x_1 x_2 \\ x_2^2 \end{bmatrix} \cdot \begin{bmatrix} z_1^2 \\ \sqrt{2} z_1 z_2 \\ z_2^2 \end{bmatrix} = x_1^2 z_1^2 + x_2^2 z_2^2 + 2 x_1 x_2 z_1 z_2$

$$= (x_1 z_1 + x_2 z_2)^2$$
$$= (\mathbf{x} \cdot \mathbf{z})^2$$

d $\Phi(\mathbf{x}) \cdot \Phi(\mathbf{z}) = K(\mathbf{x}, \mathbf{z}) = (\mathbf{x} \cdot \mathbf{z})^d$

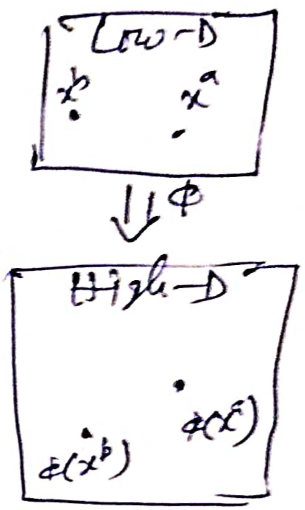
For many mappings from a low-D space to a high-D space, there is a simple operation on two vectors in the low-D space that can be used to compute the scalar product of their two images in the High-D space. 48

Finally: The Kernel Trick!

$K(x^a, x^b) = \Phi(x^a) \cdot \Phi(x^b)$

letting the kernel do the work

doing the scalar product in the obvious way.



$$\text{maximize}_\alpha \quad \sum_i \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j)$$

$$K(\mathbf{x}_i, \mathbf{x}_j) = \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j)$$
$$\sum_i \alpha_i y_i = 0$$
$$C \geq \alpha_i \geq 0$$

$$\mathbf{w} = \sum_i \alpha_i y_i \Phi(\mathbf{x}_i)$$

$$b = y_k - \mathbf{w}.\Phi(\mathbf{x}_k)$$

for any k where $C > \alpha_k > 0$

- Never represent features explicitly
  - Compute dot products in closed form

- Constant-time high-dimensional dot-products for many classes of features

Common Kernels

Performance of

1) SVM works very we The user must choo the kernel function & its parameters, but the rest is automatic.

2) They can be expensive in time & space for big data-sets. The computation of the max^m margin hyper-plane depends on the square of the number of training cases. We need to store all the support vectors.

3) SVM's are very good if you have no idea about what structure to impose on the task.

- Polynomials of degree d

$$K(\mathbf{u},\mathbf{v}) = (\mathbf{u}\cdot\mathbf{v})^d$$

- Polynomials of degree up to d

$$K(\mathbf{u},\mathbf{v}) = (\mathbf{u}\cdot\mathbf{v}+1)^d$$

- Gaussian/Radial kernels (polynomials of all orders – recall series expansion of exp)

$$K(\mathbf{u},\mathbf{v}) = \exp\left(-\frac{||\mathbf{u}-\mathbf{v}||^2}{2\sigma^2}\right)$$

- Sigmoid

$$K(\mathbf{u},\mathbf{v}) = \tanh(\eta\mathbf{u}\cdot\mathbf{v}+\nu)$$

$\beta\ \vec{u}\cdot\vec{v} + \gamma$

53

Q: Consider the following dataset:

X	Y
1	1
2	1
4	-1
5	-1
6	1

where, X is the conditional feature and Y is the decision feature (class) of the objects. Answer the following:
1) Graphically demonstrate that the objects are not linearly separable.
2) Apply the SVM and polynomial kernel function $K(u,v) = (uv+1)^2$ to generate the discriminant function. Assume that the Lagrangian multipliers corresponding to the objects are
$\alpha_1 = 0$, $\alpha_2 = 2.5$, $\alpha_3 = 0$, $\alpha_4 = 7.3$, $\alpha_5 = 4.8$
3) Use the discriminant function to predict the class label of object with $X = 3$.

Example

- Suppose we have 5 one-dimensional data points
  - $x_1=1$, $x_2=2$, $x_3=4$, $x_4=5$, $x_5=6$, with 1, 2, 6 as class 1 and 4, 5 as class 2 $\Rightarrow$ $y_1=1$, $y_2=1$, $y_3=-1$, $y_4=-1$, $y_5=1$
- We use the polynomial kernel of degree 2
  - $K(x,y) = (xy+1)^2$
  - C is set to 100
- We first find α_i ($i=1, \ldots, 5$) by

	X	Y	
$O_1 = x_1$	1	1 y_1	α_1
$O_2 = x_2$	2	1 y_2	α_2
$O_3 = x_3$	4	-1 y_3	α_3
$O_4 = x_4$	5	-1 y_4	α_4
$O_5 = x_5$	6	1 y_5	α_5

$$\max. \sum_{i=1}^{5} \alpha_i - \frac{1}{2} \sum_{i=1}^{5} \sum_{i=1}^{5} \alpha_i \alpha_j y_i y_j (x_i x_j + 1)^2$$

$$\text{subject to } \boxed{100 \geq} \alpha_i \geq 0, \sum_{i=1}^{5} \alpha_i y_i = 0$$

$f(z) = wz + b = \sum_i \alpha_i y_i K(x_i, z) + b$

$= \sum \alpha_i y_i (x_i z + 1)^2 + b$

$= 2.5(1)(2z+1)^2 + 7.333(-1)(5z+1)^2 + 4.833(1)(6z+1)^2 + b$

x	y	α
1	1	→ 0
2	1	→ 2.5
4	-1	→ 0
5	-1	→ 7.333
6	1	→ 4.833

Example

- By using a QP solver, we get
  - $\alpha_1=0$, $\alpha_2=2.5$, $\alpha_3=0$, $\alpha_4=7.333$, $\alpha_5=4.833$
  - Note that the constraints are indeed satisfied
  - The support vectors are $\{x_2=2, x_4=5, x_5=6\}$
- The discriminant function is

$W = \sum \alpha_i y_i \phi(x_i)$
$= 2.5\,\phi(2)$
$- 7.333\,\phi(5)$
$+ 4.833\,\phi(6)$

$$
\begin{aligned}
& f(z) \\
&= 2.5(1)(2z+1)^2 + \underbrace{4.833}_{\alpha_5}\underbrace{(1)}_{y_5}\underbrace{(6z+1)^2}_{K(z,x_5)} + 7.333(-1)(5z+1)^2 + b \\
&= 0.6667z^2 - 5.333z + b
\end{aligned}
$$

- b is recovered by solving f(2)=1 or by f(5)=-1 or by f(6)=1, as x_2 and x_5 lie on the line $\phi(\mathbf{w})^T\phi(\mathbf{x}) + b = 1$ and x_4 lies on the line $\phi(\mathbf{w})^T\phi(\mathbf{x}) + b = -1$
- All three give b=9 $\Longrightarrow f(z) = 0.6667z^2 - 5.333z + 9$

$\cancel{W = \alpha_2 y_2 \phi(x_2)}$

$\cancel{= 2.5(1)\,\phi(2)}$

$b = y_K - W\phi(x_K)$

$x_2 = 2,\ y_2 = 1$, $\alpha_2 = 2.5$

$\cancel{\phi(W)^T\phi(X) + b = 1}$

for $f(2)=1 \Rightarrow x = 2, y = 1$

$\therefore \cancel{b = 1 - \phi(W)^T\phi(X) = 1}$

$\Rightarrow b = 1 - [2.5\,\phi(x_2) - 7.333\,\phi(x_4) + 4.833\,\phi(x_5)]\cdot\phi(x_2)$

$= 1 - [2.5 \times \phi(x_2)\cdot\phi(x_2) - 7.333\,\phi(x_4)\cdot\phi(x_2) + 4.833\,\phi(x_5)\cdot\phi(x_2)]$

$= 1 - [2.5\,K(x_2,x_2) - 7.333\,K(x_4,x_2) + 4.833\,K(x_5,x_2)]$

$= 1 - [2.5 \times (x_2^2+1)^2 - 7.333\,(x_4 x_2+1)^2 + 4.833\,(x_5 x_2+1)^2$

$\Rightarrow b = 1 - [2.5(5^2) - 7.333(11)^2 + 4.833(13)^2]$

54

Example

$= 1 - [62.5 - 887.293 + 816.777]$

$= 1 - [879.277 - 887.293] = 1 - [-8.016] = 9.016 \approx 9$

α
→ 0
2.5
0
7.333
4.833

$f(z) = 0.6667z^2 - 5.333z + 9$

$f(3) = 0.6667(3)^2 - 5.333(3) + 9$

$= 6.0003 - 15999 + 9$

$= 15.0003 - 15.999 < 0$

$\Rightarrow x = 3$ lies in class 2