5.6 Relaxation Procedures

- broader class of criterion functions
- minimization method

Descent Algorithm

$$J_q(a) = \sum_{y \in Y} (a^t y)^2 \quad \text{like } J_p(a) = \sum_{y \in Y} -a^t y$$

↑ misclassified samples

Note: J_q has a continuous gradient (unlike J_p)

↑ smoother surface

problem 1: smoothness of J_q near boundary → converge to point on boundary

problem 2: prone to domination by longest vectors

Soln: $$J_r(a) = \frac{1}{2} \sum_{y \in Y} \frac{(a^t y - b)^2}{\|y\|^2}$$

← add in a boundary term

↑ normalization for vector length

Gradient $$\nabla J_r = \sum_{y \in Y} \frac{a^t y - b}{\|y\|^2} y$$

update $a(1)$ arbitrary

$$a(k+1) = a(k) + \eta(k) \sum_{y \in \mathcal{Y}} \frac{b - a^t y}{\|y\|^2} y$$

Relaxation algorithm Alg. 8

↖ batch alg.

Note: single sample alg shown in Alg 9

note: superscript to denote k^{th} misclassified sample

$$\text{eg } a(k+1) = a(k) + \eta \frac{b - a^t(k) y^k}{\|y^k\|^2} y^k$$

where $a^t(k) y^k \leq b$ for all k

also assume fixed learning rate η

Alg is called "single-sample relaxation rule with margin"

geometrical interpretation:

$$r(k) = \frac{b - a^t(k) y^k}{\|y^k\|}$$

is the distance from $a(k)$ to $a^t y^k = b$

see Fig. 5.14

moved

note: $a(k)$ is moved a fraction η of the distance to the hyperplane $a^t y^k = b$

$\rightarrow$ if $\eta = 1 \Rightarrow$ move exactly to the hyperplane

idea: "tension" of inequality $a^t(k) y^k \leq b$ is "relaxed"

note: After updating $a(k+1)$

we have $a(k+1) y^k - b = (1-\eta)(a^t(k) y^k - b)$

↑ update proportional to η

if $\eta > 1 \Rightarrow a^t(k+1) y^k > b$ <u>over relaxation</u>

if $\eta < 1 \Rightarrow a^t(k+1) y^k < b$ <u>under relaxation</u>

note: generally $0 < \eta < 2$

over relaxation $\Longrightarrow$ over shooting fig 5.15

under relaxation $\Longrightarrow$ undershooting $\Rightarrow$ slow descent fig 5.15

Convergence Proof

obs. if # of corrections is finite $\Rightarrow$ $a(k)$ is a soln vector

if not finite $\Rightarrow$ $a(k)$ converges to a limit vector on the boundary of the soln region

note region $a^t y \geq b$ is contained in region $a^t y > 0$ if $b > 0$

$\Rightarrow$ $a(k)$ will enter larger region & remain for all $k >$ some k_0

why? if $\hat{a}$ is any vector in the soln region, i.e. $\hat{a}^t y_i > b$ $\forall i$

$\Rightarrow$ $a(k)$ get closer to $\hat{a}$ at each step

How can we see this?

Consider the update: $a(k+1) = a(k) + \eta \frac{b - a^t(k) y^k}{\|y^k\|^2} y^k$

then $\Rightarrow$ $\|a(k+1) - \hat{a}\|^2 = \|a(k) - \hat{a}\|^2 - 2\eta \frac{(b - a^t(k) y^k)}{\|y^k\|^2} (\hat{a} - a(k))^t y^k + \eta^2 \frac{(b - a^t(k) y^k)^2}{\|y^k\|^2}$

and $(\hat{a} - a(k))^t y^k > b - a^t(k) y^k \geq 0$

+ $0 < \eta < 2$

so $\|a(k+1) - \hat{a}\| \leq \|a(k) - \hat{a}\|$

5.7 Nonseparable Behavior

obs: Life is wonderful when the data is linearly separable

⟹ we can use perceptron & ~~estat~~ relaxation procedures

these are "error-correcting" procedures

⟹ we modify "a" when a misclassification occurs

obs: any set of fewer than $2\hat{d}$ samples is likely to be linearly separable

note: a sufficiently large sample is likely to ~~be~~ not to be linearly separable

⟹ problem "error-correction" never terminates

⟹ need rules for terminating correction procedure

Empirical rules: based on tendency for "a" to fluctuate near some limiting value

goal: obtain acceptable performance in nonseparable cases & find separating vector in separable cases.

approach: variable learning rate $\eta(k)$

eg $\eta(k) \to 0$ as $k \to \infty$

rate of change in η is important (simulated annealing ch