Rigorous Support Vector Machine and Feature Selection

Student: Jinbo Bi

Supervisor: Dr. Vladimir N. Vapnik

bij2@rpi.edu

vlad@research.nj.nec.com

A VC BOUND ON GENERALIZATION ERROR (SLT P148)

Theorem 1 *With high probability, the bound*

$$R(\alpha_\ell) \leq R_{emp}(\alpha_\ell) + \frac{h}{\ell}(1 - \ln\frac{h}{2\ell})\left(1 + \sqrt{1 + \frac{4R_{emp}(\alpha_\ell)}{\frac{h(1-\ln(h/2\ell))}{\ell}}}\right)$$

holds true, where

$R_{emp}(\alpha_\ell)$ *is the percentage of training errors,*

h *is the VC dimension of the set of hypothesis functions.*

Separating Hyperplanes

Separating hyperplanes: $h = n + 1$

$$y = \begin{cases} 1, & \text{if } (\mathbf{w} \cdot \mathbf{x}) - b \geq 0, \\ -1, & \text{if } (\mathbf{w} \cdot \mathbf{x}) - b < 0, \end{cases} \qquad ||\mathbf{w}|| = 1.$$

Δ-margin separating hyperplanes:

$$y = \begin{cases} 1, & \text{if } (\mathbf{w} \cdot \mathbf{x}) - b \geq \Delta, \\ c, & \text{if } -\Delta < (\mathbf{w} \cdot \mathbf{x}) - b < \Delta, \\ -1, & \text{if } (\mathbf{w} \cdot \mathbf{x}) - b \leq -\Delta, \end{cases} \qquad ||\mathbf{w}|| = 1.$$

BOUND ON VC DIMENSION (SLT P408)

Theorem 2 *If input vectors belong to a sphere of radius R, then the set of Δ-margin separating hyperplanes has the VC dimension h bounded by*

$$h \leq \min\left\{\left[\frac{R^2}{\Delta^2}\right], n\right\} + 1.$$

If data are uniformly distributed on the surface of the sphere, then the bound is tight.

STRUCTURAL RISK MINIMIZATION

- Minimize the right hand side of the inequality

$$R(\alpha_\ell) \leq R_{emp}(\alpha_\ell) + \Phi(\frac{h}{\ell}).$$

- Fix the VC dimension h, and minimize the empirical risk, which means fix the ratio $\frac{R^2}{\Delta^2}$ and minimize the $R_{emp}(\alpha)$.
- Use tighter bound by normalizing data ($R = 1$).
- Choose the h which gives the best bound.

SRM OF HYPERPLANES WITH MARGIN

The primal

$$\begin{aligned} \min_{\mathbf{w},b,\xi} \quad & \textstyle\sum_{i=1}^{\ell} \xi_i \\ \text{s.t.} \quad & y_i((\mathbf{w}\cdot\mathbf{x}_i)-b) \geq 1-\xi_i, \quad i=1,\cdots,\ell, \\ & \xi_i \geq 0, \quad i=1,\cdots,\ell, \\ & (\mathbf{w}\cdot\mathbf{w}) \leq H. \end{aligned}$$

The dual

$$\begin{aligned} \min_{\alpha} \quad & \sqrt{H\sum_{i,j=1}^{\ell}\alpha_i\alpha_j y_i y_j(\mathbf{x}_i\cdot\mathbf{x}_j)} - \sum_{i=1}^{\ell}\alpha_i \\ \text{s.t.} \quad & \textstyle\sum_{i=1}^{\ell} y_i\alpha_i = 0, \\ & 0 \leq \alpha_i \leq 1, \quad i=1,\cdots,\ell. \end{aligned}$$

NEW IDEA

- Until now we did not specify the inner product. Now we introduce the inner product with scaling factors

$$\mathbf{w}' S \mathbf{x} = \sum_{i=1}^{n} w_i x_i s_i, \ s_i \geq 0.$$

- Let us optimize the bound over both $\mathbf{w}$ and S.

THE OPTIMIZATION PROBLEM

$$\min_{\mathbf{w},S,b,\xi} \quad \sum_{i=1}^{\ell} \xi_i + \gamma \sum_{j=1}^{n} s_j$$

$$\text{s.t.} \quad \left.\begin{array}{l} y_i(\mathbf{w}' S \mathbf{x}_i - b) \geq 1 - \xi_i, \\ \xi_i \geq 0, \quad i = 1, \cdots, \ell, \end{array}\right\} \text{empirical error}$$

$$\left.\begin{array}{l} \mathbf{w}' S \mathbf{w} \leq H, \\ \mathbf{x}_i' S \mathbf{x}_i \leq 1, \quad i = 1, \cdots, \ell, \\ s_j \geq 0, \quad j = 1, \cdots, . \end{array}\right\} \text{capacity control}$$

The algorithm:

(1) Finds the $\mathbf{w}$ with fixed S in the dual space

(2) Finds the S with fixed $\mathbf{w}$ in the primal space

(3) Re-normalize data, and go to step (1)

EXPERIMENTS

- Synthetic Data
- Digit Recognition

SYNTHETIC DATA

- Data are i.i.d. drew, uniformly distributed on $[-5, 5]^{52}$
- The classification rule is: sgn$(\frac{1}{\sqrt{2}}x_1 - \frac{1}{\sqrt{2}}x_2)$
- The parameters are: $H = 9$, $\gamma = 0.02$.
  (H is about $10\% - 20\%$ of data)

Synthetic Data (I)

$\ell_1 = \ell_2 = 50$, $(\ell = \ell_1 + \ell_2 = 100)$

Iter	#Feat.	R_{trn}	R_{tst}	Upd.H
1	52	2	12	9
2	12	1	1	9
3	8	0	0.5	9
4	8	0	0	7
5	6	0	0	6
6	5	0	0	6

Feat.	$\mathbf{w}_{opt}$	$\mathbf{w}_{est}$
1	0.7	0.67
2	-0.7	-0.73
11	0	0.06
19	0	-0.16
34	0	-0.05

$||\mathbf{w}_{opt} - \mathbf{w}_{est}|| = 0.18$

Synthetic Data (II)

$\ell_1 = 37, \ell_2 = 13, (\ell = \ell_1 + \ell_2 = 50)$

Iter	#Feat.	R_{trn}	R_{tst}	Corr. h
1	52	10	40	9
2	22	2	20	9
3	18	0	19	9
4	14	0	15	9
5	13	0	14	9
6	10	0	13.5	9
7	8	0	12	8
8	8	0	8	7
9	7	0	5	6
10	5	0	4.5	5

Feat.	$\mathbf{w}_{opt}$	$\mathbf{w}_{est}$
1	0.7	0.48
2	-0.7	-0.83
9	0	0.15
14	0	-0.22
49	0	-0.03

$||\mathbf{w}_{opt} - \mathbf{w}_{est}|| = 0.37$

HANDWRITTEN DIGIT DATA

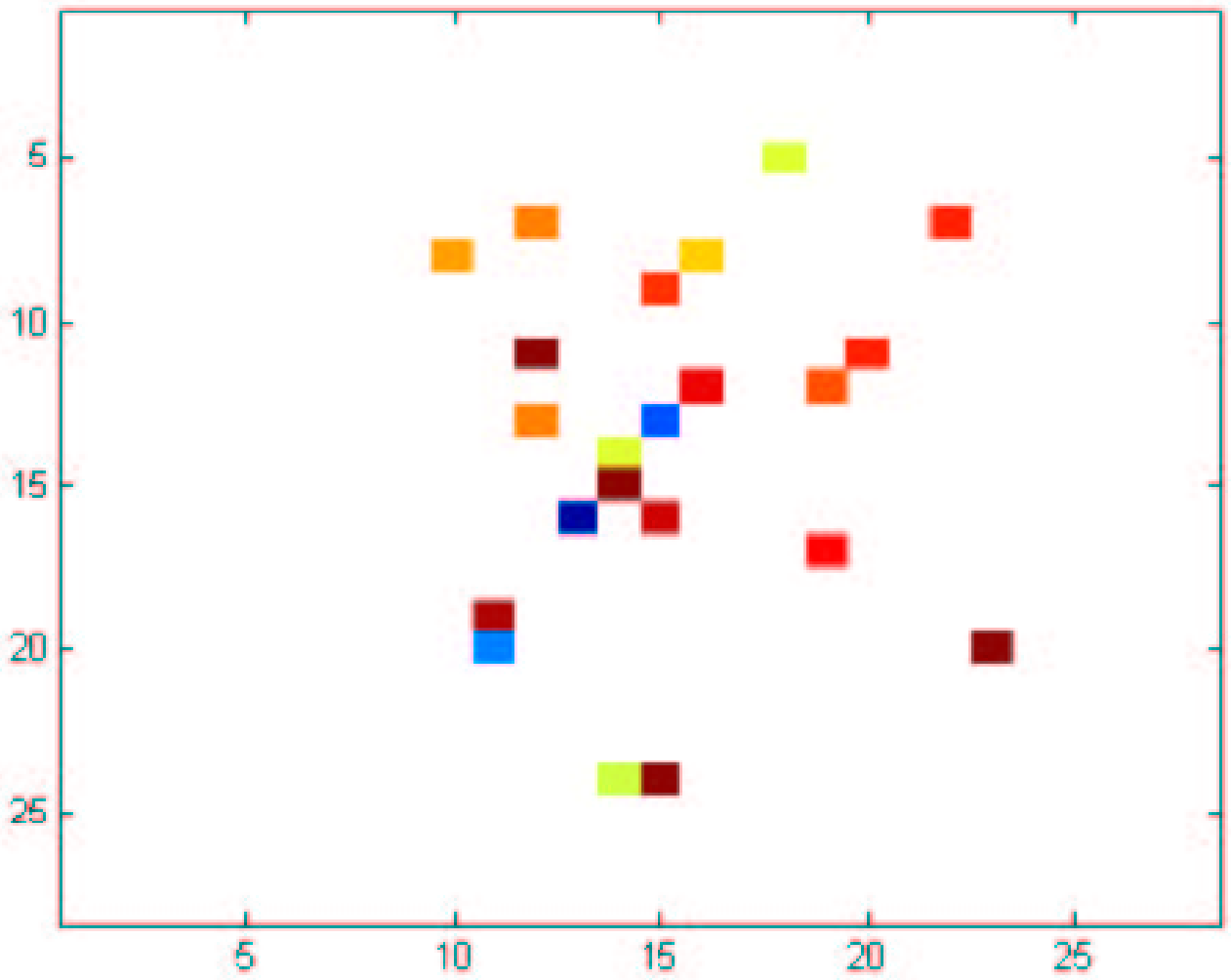
Distinguish $\{1,\ 2,\ 3,\ 4,\ 5\}$ **from** $\{6,\ 7,\ 8,\ 9,\ 0\}$

$n = 784 = 28 \times 28$, $H = 16$, $\gamma = 0.00001$

Train: 100, Test: 1000

Iter	#Feat.	R_{trn}	R_{tst}	Upd.H
1	784	9	22.5	16
2	37	6	21.9	16
3	23	6	21.6	16
4	23	6	21.7	16
5	22	5	21.9	16

Selected Features in Digit Recognition



CONCLUSIONS

- The theorems work.
- Using our algorithm, we can control the generalization risk.
- Rigorous SVM allows us to perform feature selection.