Rigorous Support Vector Machine and Feature Selection

Student: Jinbo Bì
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*

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

*holds true, where*

- \( R_{emp}(\alpha) \) is the percentage of training errors,
- \( h \) is the VC dimension of the set of hypothesis functions.
Separating hyperplanes: \( h = n + 1 \)

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

\( \Delta \)-margin separating hyperplanes:

\[
y = \begin{cases} 
1, & \text{if } (w \cdot x) - b \geq \Delta, \\
c, & \text{if } -\Delta < (w \cdot x) - b < \Delta, \\
-1, & \text{if } (w \cdot x) - b \leq -\Delta,
\end{cases}
\quad ||w|| = 1.
\]
**Theorem 2** If input vectors belong to a sphere of radius $R$, then the set of $\Delta$-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
  \[ R(\alpha_\ell) \leq R_{emp}(\alpha_\ell) + \Phi\left(\frac{h}{\ell}\right). \]

- 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

\[ \min_{w,b,\xi} \sum_{i=1}^{\ell} \xi_i \]

s.t. \[ y_i ((w \cdot x_i) - b) \geq 1 - \xi_i, \quad i = 1, \ldots, \ell, \]
\[ \xi_i \geq 0, \quad i = 1, \ldots, \ell, \]
\[ (w \cdot w) \leq H. \]

The dual

\[ \min_{\alpha} \sqrt{H \sum_{i,j=1}^{\ell} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) - \sum_{i=1}^{\ell} \alpha_i} \]

s.t. \[ \sum_{i=1}^{\ell} y_i \alpha_i = 0, \]
\[ 0 \leq \alpha_i \leq 1, \quad i = 1, \ldots, \ell. \]
NEW IDEA

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

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

• Let us optimize the bound over both \( w \) and \( S \).
The optimization problem

\[
\begin{align*}
\min_{w, S, b, \xi} & \quad \sum_{i=1}^{\ell} \xi_i + \gamma \sum_{j=1}^{n} s_j \\
\text{s.t.} & \quad y_i (w' S x_i - b) \geq 1 - \xi_i, \\
& \quad \xi_i \geq 0, \quad i = 1, \ldots, \ell, \\
& \quad w' S w \leq H, \\
& \quad x_i' S x_i \leq 1, \quad i = 1, \ldots, \ell, \\
& \quad s_j \geq 0, \quad j = 1, \ldots.
\end{align*}
\]

\( \gamma \) \text{ empirical error} \\
\( \ell \) \text{ capacity control}

The algorithm:
(1) Finds the \( w \) with fixed \( S \) in the dual space
(2) Finds the \( S \) with fixed \( w \) 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 $[-5, 5]^2$.
- The classification rule is: $\text{sgn}(\frac{1}{\sqrt{2}}x_1 - \frac{1}{\sqrt{2}}x_2)$.
- The parameters are: $H = 9$, $\gamma = 0.02$.  
  ($H$ is about 10% – 20% of data)
## Synthetic Data (I)

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

<table>
<thead>
<tr>
<th>Iter</th>
<th>#Feat.</th>
<th>( R_{trn} )</th>
<th>( R_{tst} )</th>
<th>Upd. ( H )</th>
<th>Feature</th>
<th>( w_{opt} )</th>
<th>( w_{est} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>52</td>
<td>2</td>
<td>12</td>
<td>9</td>
<td>1</td>
<td>0.7</td>
<td>0.67</td>
</tr>
<tr>
<td>2</td>
<td>12</td>
<td>1</td>
<td>1</td>
<td>9</td>
<td>2</td>
<td>-0.7</td>
<td>-0.73</td>
</tr>
<tr>
<td>3</td>
<td>8</td>
<td>0</td>
<td>0.5</td>
<td>9</td>
<td>11</td>
<td>0</td>
<td>0.06</td>
</tr>
<tr>
<td>4</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>7</td>
<td>19</td>
<td>0</td>
<td>-0.16</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>34</td>
<td>0</td>
<td>-0.05</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[ \| w_{opt} - w_{est} \| = 0.18 \]
SYNTHETIC DATA (II)

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

<table>
<thead>
<tr>
<th>Iter</th>
<th>#Feat.</th>
<th>( R_{trn} )</th>
<th>( R_{tst} )</th>
<th>Corr.</th>
<th>( h )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>52</td>
<td>10</td>
<td>40</td>
<td></td>
<td>9</td>
</tr>
<tr>
<td>2</td>
<td>22</td>
<td>2</td>
<td>20</td>
<td></td>
<td>9</td>
</tr>
<tr>
<td>3</td>
<td>18</td>
<td>0</td>
<td>19</td>
<td></td>
<td>9</td>
</tr>
<tr>
<td>4</td>
<td>14</td>
<td>0</td>
<td>15</td>
<td></td>
<td>9</td>
</tr>
<tr>
<td>5</td>
<td>13</td>
<td>0</td>
<td>14</td>
<td></td>
<td>9</td>
</tr>
<tr>
<td>6</td>
<td>10</td>
<td>0</td>
<td>13.5</td>
<td></td>
<td>9</td>
</tr>
<tr>
<td>7</td>
<td>8</td>
<td>0</td>
<td>12</td>
<td></td>
<td>8</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
<td>0</td>
<td>8</td>
<td></td>
<td>7</td>
</tr>
<tr>
<td>9</td>
<td>7</td>
<td>0</td>
<td>5</td>
<td></td>
<td>6</td>
</tr>
<tr>
<td>10</td>
<td>5</td>
<td>0</td>
<td>4.5</td>
<td></td>
<td>5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Feat.</th>
<th>( w_{opt} )</th>
<th>( w_{est} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.7</td>
<td>0.48</td>
</tr>
<tr>
<td>2</td>
<td>-0.7</td>
<td>-0.83</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>0.15</td>
</tr>
<tr>
<td>14</td>
<td>0</td>
<td>-0.22</td>
</tr>
<tr>
<td>49</td>
<td>0</td>
<td>-0.03</td>
</tr>
</tbody>
</table>

\[ \| w_{opt} - w_{est} \| = 0.37 \]
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

<table>
<thead>
<tr>
<th>Iter</th>
<th>#Feat.</th>
<th>( R_{trn} )</th>
<th>( R_{tst} )</th>
<th>Upd.( H )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>784</td>
<td>9</td>
<td>22.5</td>
<td>16</td>
</tr>
<tr>
<td>2</td>
<td>37</td>
<td>6</td>
<td>21.9</td>
<td>16</td>
</tr>
<tr>
<td>3</td>
<td>23</td>
<td>6</td>
<td>21.6</td>
<td>16</td>
</tr>
<tr>
<td>4</td>
<td>23</td>
<td>6</td>
<td>21.7</td>
<td>16</td>
</tr>
<tr>
<td>5</td>
<td>22</td>
<td>5</td>
<td>21.9</td>
<td>16</td>
</tr>
</tbody>
</table>
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.