Support Vector Machine Regression

Greg Grudic
(Notes borrowed from Bernhard Schölkopf)
What are the Support Vectors in Classification?

Maximized Margin
Learning Regression Models

- Collect Training data
- Build Model: stock value = M (feature space)
- Make a prediction
SV Regression: $\varepsilon$-Insensitive Loss

Goal: generalize SV pattern recognition to regression, preserving the following properties:

- formulate the algorithm for the linear case, and then use kernel trick
- sparse representation of the solution in terms of SVs

$\varepsilon$-Insensitive Loss:

$$|y - f(x)|_\varepsilon := \max\{0, |y - f(x)| - \varepsilon\}$$

Estimate a linear regression $f(x) = \langle w, x \rangle + b$ by minimizing

$$\frac{1}{2}||w||^2 + \frac{C}{m} \sum_{i=1}^{m} |y_i - f(x_i)|_\varepsilon.$$

B. Schölkopf, Canberra, February 2002
\( \varepsilon \)-SV Regression Estimation [64]
Formulation as an Optimization Problem

Estimate a linear regression

$$f(x) = \langle w, x \rangle + b$$

with precision $\varepsilon$ by minimizing

\[
\text{minimize} \quad \tau(w, \xi, \xi^*) = \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{m} (\xi_i + \xi_i^*)
\]

subject to

\[
(\langle w, x_i \rangle + b) - y_i \leq \varepsilon + \xi_i \\
y_i - (\langle w, x_i \rangle + b) \leq \varepsilon + \xi_i^* \\
\xi_i, \xi_i^* \geq 0
\]

for all $i = 1, \ldots, m$. 

B. Schölkopf, Canberra, February 2002
Dual Problem, In Terms of Kernels

For $C > 0, \varepsilon \geq 0$ chosen a priori,

\[
\text{maximize} \quad W(\alpha, \alpha^*) = -\varepsilon \sum_{i=1}^{m} (\alpha^*_i + \alpha_i) + \sum_{i=1}^{m} (\alpha^*_i - \alpha_i) y_i \\
- \frac{1}{2} \sum_{i,j=1}^{m} (\alpha^*_i - \alpha_i)(\alpha^*_j - \alpha_j) k(x_i, x_j)
\]

subject to \quad 0 \leq \alpha_i, \alpha^*_i \leq C, \ i = 1, \ldots, m, \ \text{and} \ \sum_{i=1}^{m} (\alpha^*_i - \alpha_i) = 0.

The regression estimate takes the form

\[
f(x) = \sum_{i=1}^{m} (\alpha^*_i - \alpha_i) k(x_i, x) + b,
\]

B. Schölkopf, Canberra, February 2002
\( \nu \)-SV Regression

Again, use \( \nu \) to eliminate another parameter: Estimate \( \varepsilon \) from the data s.t. the \( \nu \)-property holds.

Primal problem: for \( 0 \leq \nu \leq 1 \), minimize

\[
\tau (w, \varepsilon) = \frac{1}{2} \|w\|^2 + C \left( \nu \varepsilon + \frac{1}{m} \sum_{i=1}^{m} |y_i - f(x_i)| \varepsilon \right)
\]
\( \nu \)-SV-Regression: Automatic Tube Tuning

Identical machine parameters (\( \nu = 0.2 \)), but different amounts of noise in the data.
$\varepsilon$-SV-Regression, Run on the Same Data

*Identical* machine parameters ($\varepsilon = 0.2$), but different amounts of noise in the data.
Boston Housing Benchmark

- 506 examples, 13-dimensional.

Results (MSE):
- Bagging regression trees: 11.7 [8]
- $\varepsilon$-SV regression: 7.6 [59]

Mean Squared Error (MSE)
Results on test data:

$$\text{TestData} : (y_1, x_1), \ldots, (y_K, x_K)$$

$$MSE = \frac{1}{K} \sum_{i=1}^{K} (y_i - f(x_i))^2$$

- 100 runs, with 25 randomly selected test points.
- training set is split into actual training set and validation set (80 points) for selecting $\varepsilon$, $C'$, and kernel parameters
## Comparison: $\nu$ vs. $\varepsilon$

<table>
<thead>
<tr>
<th>$\nu$-SVR</th>
<th>0.1</th>
<th>0.2</th>
<th>0.3</th>
<th>0.4</th>
<th>0.5</th>
<th>0.6</th>
<th>0.7</th>
<th>0.8</th>
<th>0.9</th>
<th>1.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>automatic $\varepsilon$</td>
<td>2.6</td>
<td>1.7</td>
<td>1.2</td>
<td>0.8</td>
<td>0.6</td>
<td>0.3</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td><strong>MSE</strong></td>
<td>9.4</td>
<td>8.7</td>
<td>9.3</td>
<td>9.5</td>
<td>10.0</td>
<td>10.6</td>
<td>11.3</td>
<td>11.3</td>
<td>11.3</td>
<td>11.3</td>
</tr>
<tr>
<td>Errors</td>
<td>0.0</td>
<td>0.1</td>
<td>0.2</td>
<td>0.2</td>
<td>0.3</td>
<td>0.4</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>SVs</td>
<td>0.3</td>
<td>0.4</td>
<td>0.6</td>
<td>0.7</td>
<td>0.8</td>
<td>0.9</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>$\varepsilon$-SVR</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td><strong>MSE</strong></td>
<td>11.3</td>
<td>9.5</td>
<td>8.8</td>
<td>9.7</td>
<td>11.2</td>
<td>13.1</td>
<td>15.6</td>
<td>18.2</td>
<td>22.1</td>
<td>27.0</td>
</tr>
<tr>
<td>Errors</td>
<td>0.5</td>
<td>0.2</td>
<td>0.1</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>SVs</td>
<td>1.0</td>
<td>0.6</td>
<td>0.4</td>
<td>0.3</td>
<td>0.2</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
</tr>
</tbody>
</table>

- RBF kernel, $C$ and $\sigma$ chosen as in [56]

B. Schölkopf, Canberra, February 2002