7. Sparse Kernel Machines

22.03.2007
Maximum Margin Classifiers

- \( y(x) = w^T \phi(x) + b \)
- \( x_1, \ldots, x_N \)  N input vectors
- \( t_1, \ldots, t_N \in \{-1, 1\} \) target values
- \( \phi(x) \) fixed feature space transformation
- bias parameter \( b \)
- new data points \( x \) classified according to sign of \( y(x) \)

Assume: training set is linearly separable in feature space

\( w, b : y(x_n) > 0 \) for points having \( t_n = +1 \) and \( y(x_n) < 0 \) for points having \( t_n = -1 \)

\( t_n y(x_n) > 0 \) for all training data points

\( y = 1 \)
\( y = 0 \)
\( y = -1 \)
Maximum Margin Classifiers

- $y(x) = w^T \phi(x) + b$
- $x_1, \ldots, x_N$ N input vectors
- $t_1, \ldots, t_N \in \{-1, 1\}$ target values
- $\phi(x)$ fixed feature space transformation
- bias parameter $b$
- new data points $x$ classified according to sign of $y(x)$
- Assume: training set is linearly separable in feature space
- $w, b : y(x_n) > 0$ for points having $t_n = +1$ and $y(x_n) < 0$ for points having $t_n = -1$
- $t_n y(x_n) > 0$ for all training data points
**Maximum Margin Classifiers**

- Perpendicular distance of \( \mathbf{x} \) from hyperplane \( y(\mathbf{x}) = 0 \) is given by \( \frac{|y(\mathbf{x})|}{||\mathbf{w}||} \)

![Diagram showing perpendicular distance and margin](image)
Maximum Margin Classifiers

- Perpendicular distance of \( x \) from hyperplane \( y(x) = 0 \) is given by \( \frac{|y(x)|}{||w||} \)

- Distance of \( x_n \) to decision surface

\[
\frac{t_n y(x_n)}{||w||} = \frac{t_n (w^T \phi(x_n) + b)}{||w||}
\]
Maximum Margin Classifiers

- Perpendicular distance of \( x \) from hyperplane \( y(x) = 0 \) is given by \( \frac{|y(x)|}{||w||} \)

- Distance of \( x_n \) to decision surface

\[
\frac{t_n y(x_n)}{||w||} = \frac{t_n (w^T \phi(x_n) + b)}{||w||}
\]

- Margin given by perpend. distance to closest point \( x_n \)
Perpendicular distance of $\mathbf{x}$ from hyperplane $y(\mathbf{x}) = 0$ is given by $|y(\mathbf{x})|/||\mathbf{w}||$

Distance of $\mathbf{x}_n$ to decision surface

$$\frac{t_n y(\mathbf{x}_n)}{||\mathbf{w}||} = \frac{t_n (\mathbf{w}^T \phi(\mathbf{x}_n) + b)}{||\mathbf{w}||}$$

Margin given by perpend. distance to closest point $\mathbf{x}_n$

optimize $\mathbf{w}, b$ in order to maximize this distance, maximum margin solution:
Maximum Margin Classifiers

- Perpendicular distance of $\mathbf{x}$ from hyperplane $y(\mathbf{x}) = 0$ is given by $|y(\mathbf{x})|/||\mathbf{w}||$

- Distance of $\mathbf{x}_n$ to decision surface

$$\frac{t_n y(\mathbf{x}_n)}{||\mathbf{w}||} = \frac{t_n (\mathbf{w}^T \phi(\mathbf{x}_n) + b)}{||\mathbf{w}||}$$

- Margin given by perpend. distance to closest point $\mathbf{x}_n$

- Optimize $\mathbf{w}, b$ in order to maximize this distance, maximum margin solution:

$$\arg \max_{\mathbf{w}, b} \left\{ \frac{1}{||\mathbf{w}||} \min_n \left[ t_n (\mathbf{w}^T \phi(\mathbf{x}_n) + b) \right] \right\}$$
equivalent problem, easier to solve

rescaling $w \rightarrow \kappa w$, $b \rightarrow \kappa b$ does not change $t_n y(x_n)/||w||$
equivalent problem, easier to solve

rescaling $w \rightarrow \kappa w$, $b \rightarrow \kappa b$ does not change $t_n y(x_n)/\|w\|

t_n(w^T \phi(x_n) + b) = 1$ for the point closest to surface
equivalent problem, easier to solve

rescaling \( w \rightarrow \kappa w, b \rightarrow \kappa b \) does not change \( t_n y(x_n)/||w|| \)

\( t_n (w^T \phi(x_n) + b) = 1 \) for the point closest to surface

\( t_n (w^T \phi(x_n) + b) \geq 1 \) \( n = 1, \ldots, N \) (*)
Maximum Margin Classifiers

- equivalent problem, easier to solve
- rescaling $w \rightarrow \kappa w, b \rightarrow \kappa b$ does not change $t_n y(x_n)/||w||$
- $t_n(w^T \phi(x_n) + b) = 1$ for the point closest to surface
- $t_n(w^T \phi(x_n) + b) \geq 1 \quad n = 1, \ldots, N$ (*)
- Maximize $||w||^{-1}$ resp. minimize $||w||^2$
equivalent problem, easier to solve
rescaling $w \rightarrow \kappa w$, $b \rightarrow \kappa b$ does not change $t_n y(x_n)/||w||$

$t_n(w^T \phi(x_n) + b) = 1$ for the point closest to surface
$t_n(w^T \phi(x_n) + b) \geq 1$ for $n = 1, \ldots, N$ (*)
Maximize $||w||^{-1}$ resp. minimize $||w||^2$

$\arg \min_{w,b} \frac{1}{2}||w||^2$, subject to the constraints (*)
Lagrange multipliers $a_n \geq 0$

$$L(w, b, a) = \frac{1}{2}||w||^2 - \sum_{n=1}^{N} a_n \left\{ t_n(w^T \phi(x_n) + b) - 1 \right\}$$
Lagrange multipliers $a_n \geq 0$

$$L(w, b, a) = \frac{1}{2}||w||^2 - \sum_{n=1}^{N} a_n \left\{ t_n (w^T \phi(x_n) + b) - 1 \right\}$$

$$\frac{\partial L}{\partial w} = 0 \implies w = \sum_{n=1}^{N} a_n t_n \phi(x_n)$$

$$\frac{\partial L}{\partial b} = 0 \implies 0 = \sum_{n=1}^{N} a_n t_n$$
Lagrange multipliers $a_n \geq 0$

$$L(w, b, a) = \frac{1}{2}||w||^2 - \sum_{n=1}^{N} a_n \left\{ t_n(w^T \phi(x_n) + b) - 1 \right\}$$

$$\frac{\partial L}{\partial w} = 0 \implies w = \sum_{n=1}^{N} a_n t_n \phi(x_n)$$

$$\frac{\partial L}{\partial b} = 0 \implies 0 = \sum_{n=1}^{N} a_n t_n$$

$$\tilde{L}(a) = \sum_{n=1}^{N} a_n - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} a_n a_m t_n t_m k(x_n, x_m)$$

constraints: $a_n \geq 0$, $\sum_{n=1}^{N} a_n t_n = 0$

kernel function $k(x, x') = \phi(x)^T \phi(x)$. 

Sparse Kernel Machines
Maximum Margin Classifiers

Classify new data points $\mathbf{x}$ using the trained model

$$y(\mathbf{x}) = \sum_{n=1}^{N} a_n t_n k(\mathbf{x}, \mathbf{x}_n) + b$$
Maximum Margin Classifiers

Classify new data points $\mathbf{x}$ using the trained model

$$y(\mathbf{x}) = \sum_{n=1}^{N} a_n t_n k(\mathbf{x}, \mathbf{x}_n) + b$$

Synthetic data, two classes, 2 D, Gaussian kernel, contours of constant $y(\mathbf{x})$
data points allowed to be on the 'wrong' side of the margin, with a penalty that increases with the distance from the boundary.

\[ \xi_n \geq 0, \quad n = 1, \ldots, N \]

\[ \xi_n = 0, \quad \text{for points that are 'inside'} \]

\[ \xi_n = |t_n - y(x_n)| \quad \text{for other points} \]

\[ y = 1 \]

\[ y = 0 \]

\[ y = -1 \]

\[ \xi > 1 \]

\[ \xi < 1 \]

\[ \xi = 0 \]

\[ \sum_{n=1}^{N} \xi_n + \frac{1}{2} ||w||^2 \]

\[ L(w, b, a) = \frac{1}{2} ||w||^2 + C \sum_{n=1}^{N} \xi_n - \sum_{n=1}^{N} \mu_n \xi_n \]
Overlapping class distributions

- Data points allowed to be on the ’wrong’ side of the margin, with a penalty that increases with the distance from the boundary
- Slack variables $\xi_n \geq 0$, $n = 1, \ldots, N$
- $\xi_n = 0$, for points that are ’inside’
- $\xi_n = |t_n - y(x_n)|$ for other points

\[ y = -1 \]
\[ y = 0 \]
\[ y = 1 \]
\[ \xi > 1 \]
\[ \xi < 1 \]
\[ \xi = 0 \]
Overlapping class distributions

- Data points allowed to be on the 'wrong' side of the margin, with a penalty that increases with the distance from the boundary.
- Slack variables $\xi_n \geq 0$, $n = 1, \ldots, N$
- $\xi_n = 0$, for points that are 'inside'
- $\xi_n = |t_n - y(x_n)|$ for other points

New constraints: $t_n y(x_n) \geq 1 - \xi_n$, $n = 1, \ldots, N$
Overlapping class distributions

- data points allowed to be on the 'wrong' side of the margin, with a penalty that increases with the distance from the boundary
- \textit{slack variables} \( \xi_n \geq 0, \ n = 1, \ldots, N \)
- \( \xi_n = 0 \), for points that are 'inside'
- \( \xi_n = |t_n - y(x_n)| \) for other points

New constraints: \( t_n y(x_n) \geq 1 - \xi_n, \ n = 1, \ldots, N \)

Minimize \[ C \sum_{n=1}^{N} \xi_n + \frac{1}{2} ||w||^2 \]
Overlapping class distributions

- data points allowed to be on the 'wrong' side of the margin, with a penalty that increases with the distance from the boundary
- slack variables $\xi_n \geq 0$, $n = 1, \ldots, N$
- $\xi_n = 0$, for points that are 'inside'
- $\xi_n = |t_n - y(x_n)|$ for other points

New constraints: $t_n y(x_n) \geq 1 - \xi_n$, $n = 1, \ldots, N$

Minimize $C \sum_{n=1}^{N} \xi_n + \frac{1}{2} ||w||^2$

$L(w, b, a) = \frac{1}{2} ||w||^2 + C \sum_{n=1}^{N} \xi_n - \sum_{n=1}^{N} a_n \{ t_n (w^T \phi(x_n) + b) - 1 + \xi_n \} - \sum_{n=1}^{N} \mu_n \xi_n$
Minimize \( \frac{1}{2} \| w \|^2 + C \sum_{n=1}^{N} \xi_n \)

Constraints: \( t_n y(x_n) \geq 1 - \xi_n, \quad n = 1, \ldots, N \)

\[ \tilde{L}(a) = \sum_{n=1}^{N} a_n - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} a_n a_m t_n t_m k(x_n, x_m) \]

\[ 0 \leq a_n \leq C \]

\[ \sum_{n=1}^{N} a_n t_n = 0 \]
Minimize \( \frac{1}{2} ||\mathbf{w}||^2 - \nu \rho + \frac{1}{N} \sum_{n=1}^{N} \xi_n \)

Constraints: \( t_n y(x_n) \geq \rho - \xi_n, \quad n = 1, \ldots, N \)
and \( \xi_n \geq 0, \rho \geq 0. \)

\[ \tilde{L}(\mathbf{a}) = -\frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} a_n a_m t_n t_m k(x_n, x_m) \]

\( 0 \leq a_n \leq 1/N, \sum_{n=1}^{N} a_n t_n = 0, \sum_{n=1}^{N} a_n \geq \nu \)

\( \nu \) can be interpreted as both an upper bound on the fraction of margin errors (points for which \( \xi_n \geq 0 \)) and a lower bound on the fraction of support vectors.
Fig. 4. Toy problem (task: to separate circles from disks) solved using $\nu$-SV classification, with parameter values ranging from $\nu = 0.1$ (top left) to $\nu = 0.8$ (bottom right). The larger we make $\nu$, the more points are allowed to lie inside the margin (depicted by dotted lines). Results are shown for a Gaussian kernel, $k(x, x') = \exp(-||x - x'||^2)$ (from [31]).

Table 1. Fractions of errors and SVs, along with the margins of class separation, for the toy example in Figure 4.

Note that $\nu$ upper bounds the fraction of errors and lower bounds the fraction of SVs, and that increasing $\nu$, i.e., allowing more errors, increases the margin.

<table>
<thead>
<tr>
<th>$\nu$</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>
</tr>
</thead>
<tbody>
<tr>
<td>fraction of errors</td>
<td>0.00</td>
<td>0.07</td>
<td>0.25</td>
<td>0.32</td>
<td>0.39</td>
<td>0.50</td>
<td>0.61</td>
<td>0.71</td>
</tr>
<tr>
<td>fraction of SVs</td>
<td>0.29</td>
<td>0.36</td>
<td>0.43</td>
<td>0.46</td>
<td>0.57</td>
<td>0.68</td>
<td>0.79</td>
<td>0.86</td>
</tr>
<tr>
<td>$\rho/|w|$</td>
<td>0.005</td>
<td>0.018</td>
<td>0.115</td>
<td>0.156</td>
<td>0.364</td>
<td>0.419</td>
<td>0.461</td>
<td>0.546</td>
</tr>
</tbody>
</table>
Fig. 5. The relation between $\nu$ and $C$ (using the RBF kernel on the problem australian from the Statlog collection [25])
**Multiclass SVM**

- **one-versus-the-rest**
  Construct $K$ separate SVM’s, using data from class $C_k$ as positive examples, and the data from remaining $K - 1$ classes as negative examples.
  
  $y(x) = \max_k y_k(x)$

  Problems: different classifiers trained on different tasks ($y_k(x)$?), training sets are imbalanced.
**Multiclass SVM**

- **one-versus-the-rest**
  Construct $K$ separate SVM’s, using data from class $C_k$ as positive examples, and the data from remaining $K - 1$ classes as negative examples.
  \[ y(x) = \max_k y_k(x) \]
  Problems: different classifiers trained on different tasks ($y_k(x)$?), training sets are imbalanced.

- **one-versus-one**
  Train $K(K - 1)/2$ different 2-class SVM’s on all possible pairs of classes, classify test points according to which class has highest number of ’votes’. Problems: ambiguities, more training time
SVM’s for regression

Minimize \[
\frac{1}{2} \sum_{n=1}^{N} (y_n - t_n)^2 + \frac{\lambda}{2} \|w\|^2
\]

\[\epsilon\] - insensitive error function

\[E_\epsilon(y(x) - t) = \begin{cases} 0, & \text{if } |y(x) - t| < \epsilon \\ |y(x) - t| - \epsilon, & \text{otherwise} \end{cases}\]
SVM’s for regression

Minimize $\frac{1}{2} \sum_{n=1}^{N} (y_n - t_n)^2 + \frac{\lambda}{2} ||w||^2$

$\epsilon$-insensitive error function

$E_{\epsilon}(y(x) - t) = \begin{cases} 0, & \text{if } |y(x) - t| < \epsilon; \\ |y(x) - t| - \epsilon, & \text{otherwise} \end{cases}$
SVM’s for regression

Minimize \( \frac{1}{2} \sum_{n=1}^{N} \{y_n - t_n\}^2 + \frac{\lambda}{2} \|w\|^2 \)

\( \epsilon \)-insensitive error function

\( E_\epsilon(y(x)-t) = \begin{cases} 0, & \text{if } |y(x) - t| < \epsilon; \\ |y(x) - t| - \epsilon, & \text{otherwise} \end{cases} \)

Minimize \( C \sum_{n=1}^{N} E_\epsilon(y(x_n) - t_n) + \frac{\lambda}{2} \|w\|^2 \)
Slack variables $\xi_n \geq 0, \hat{\xi}_n \geq 0$

$\xi_n > 0 \iff t_n > y(x_n) + \epsilon$

$\hat{\xi}_n > 0 \iff t_n < y(x_n) - \epsilon$
SVM’s for regression

Slack variables $\xi_n \geq 0$, $\hat{\xi}_n \geq 0$

$\xi_n > 0 \iff t_n > y(x_n) + \epsilon$

$\hat{\xi}_n > 0 \iff t_n < y(x_n) - \epsilon$

Inside the $\xi$-tube:

$y(x_n) - \epsilon \leq t_n \leq y(x_n) + \epsilon$
SVM’s for regression

Slack variables $\xi_n \geq 0, \hat{\xi}_n \geq 0$

$\xi_n > 0 \iff t_n > y(x_n) + \epsilon$

$\hat{\xi}_n > 0 \iff t_n < y(x_n) - \epsilon$

Inside the $\xi$-tube:

$y(x_n) - \epsilon \leq t_n \leq y(x_n) + \epsilon$

Slack variables allow points to lie outside tube

$t_n \leq y(x_n) + \epsilon + \xi_n$

$t_n \leq y(x_n) + \epsilon + \hat{\xi}_n$
SVM’s for regression

Slack variables $\xi_n \geq 0, \hat{\xi}_n \geq 0$

$\xi_n > 0 \iff t_n > y(x_n) + \epsilon$

$\hat{\xi}_n > 0 \iff t_n < y(x_n) - \epsilon$

Inside the $\xi$-tube:

$y(x_n) - \epsilon \leq t_n \leq y(x_n) + \epsilon$

Slack variables allow points to lie outside tube

$t_n \leq y(x_n) + \epsilon + \xi_n$

$t_n \leq y(x_n) + \epsilon + \hat{\xi}_n$

Minimize $C \sum_{n=1}^{N} (\xi_n + \hat{\xi}_n) + \frac{1}{2} ||w||^2$

$\xi_n \geq 0, \hat{\xi}_n \geq 0$
SVM’s for regression

SVM for regression applied to sinusoidal synthetic data using gaussian kernels.