SVMs and Kernels

10701/15781 recitation
10/22/09
From a linear classifier to ...

\[ \sum_i w_i x_i + b \geq 0 \]

\[ \sum_i w_i x_i + b \leq 0 \]

*One of the most famous slides you will see, ever!
Maximum margin

Maximum possible separation between positive and negative training examples

\[
\text{minimize}_{w, b} \quad w \cdot w \quad (wx_j + b)y_j \geq 1, \forall j
\]

*One of the most famous slides you will see, ever!
Number of support vectors

SVMs:

\[ \text{minimize}_{w,b} \quad w \cdot w \]
\[ (wx_j + b)y_j \geq 1, \forall j \]

where \( m \) is dimension of the input vector

examples...
Number of support vectors

SVMs:

\[
\begin{align*}
\text{minimize}_{w,b} & \quad w \cdot w \\
(wx_j + b)y_j & \geq 1, \forall j
\end{align*}
\]

At most! \[m + 1\]

where \(m\) is dimension of the input vector

Except for degenerate cases!
Multi-class SVM example
Multi-class SVM example

Rule?
Multi-class SVM example

\[
\mathbf{w}^{(y_j)} \cdot \mathbf{x}_j + b^{(y_j)} \geq \mathbf{w}^{(y')} \cdot \mathbf{x}_j + b^{(y')} + 1,
\]

\[\forall y' \neq y_j, \forall j\]
Multi-class SVM example

\[ w^{(y_j)} \cdot x_j + b^{(y_j)} \geq w^{(y')} \cdot x_j + b^{(y')} + 1, \]

\[ \forall y' \neq y_j, \forall j \]
$\Phi(\vec{x}) = \left(x_1^2, \sqrt{2}x_1x_2, x_2^2\right)^T$

$K(\vec{x}, \vec{z}) = \left(\vec{x}^T \vec{z}\right)^2$

$\vec{x} = (x_1, x_2)$

$\vec{z} = (z_1, z_2)$
Kernels

\[ \Phi(\vec{x}) = (x_1^2, \sqrt{2}x_1x_2, x_2^2)^T \]

\[ K(\vec{x}, \vec{z}) = (\vec{x}^T \vec{z})^2 \]

\[ \vec{x} = (x_1, x_2) \]

\[ \vec{z} = (z_1, z_2) \]

\[ K(\vec{x}, \vec{z}) = (\vec{x}^T \vec{z})^2 \]

\[ = (x_1z_1 + x_2z_2)^2 \]

\[ = (x_1^2z_1^2 + 2x_1z_1x_2z_2 + x_2^2z_2^2) \]

\[ = (x_1^2, \sqrt{2}x_1x_2, x_2^2)^T (z_1^2, \sqrt{2}z_1z_2, z_2^2) \]

\[ = \phi(\vec{x})^T \phi(\vec{z}) \]
Kernels

<table>
<thead>
<tr>
<th>Type of Kernel</th>
<th>Inner product kernel</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Polynomial Kernel</td>
<td>$K(\bar{x}, \bar{x}_i) = (\bar{x}^T \bar{x}_i + \theta)^d$</td>
<td>Power $p$ and threshold $\theta$ is specified a priori by the user</td>
</tr>
<tr>
<td>Gaussian Kernel</td>
<td>$K(\bar{x}, \bar{x}_i) = e^{-\frac{1}{2\sigma^2}</td>
<td></td>
</tr>
<tr>
<td>Sigmoid Kernel</td>
<td>$K(\bar{x}, \bar{x}_i) = tanh(\eta \bar{x}^T \bar{x}_i + \theta)$</td>
<td>Mercer’s Theorem is satisfied only for some values of $\eta$ and $\theta$</td>
</tr>
<tr>
<td>Kernels for Sets</td>
<td>$K(\chi, \chi') = \sum_{i=1}^{N_\chi} \sum_{j=1}^{N_{\chi'}} k(x_i, x'_j)$</td>
<td>Where $k(x_i, x'_j)$ is a kernel on elements in the sets $\chi, \chi'$</td>
</tr>
<tr>
<td>Spectrum Kernel for strings</td>
<td>count number of substrings in common</td>
<td>It is a kernel, since it is a dot product between vectors of indicators of all the substrings.</td>
</tr>
</tbody>
</table>

Table 1: Summary of Inner-Product Kernels [Hay98]
Kernels

\[ K(x, z) = \Phi(x)^T \Phi(z) \]

Infinite dimensions?

Complexity of the optimization problem remains only dependent on the dimensionality of the input space and not of the feature space!
Finding the margin by hand
Finding the margin by hand
Finding the margin by hand
How many SVs now?

But this is 2D data?

\[ w = \sum_{i} \alpha_i y_i \Phi(x_i) \]
How many SVs now?

The worst case is the VC dimension
For a given algorithm, the largest set of points that the algorithm can shatter.
VC dimension

For a given algorithm, the largest set of points that the algorithm can shatter.
For a given algorithm, the largest set of points that the algorithm can shatter.
VC dimension

For a linear classifier in m dimensions, VC dimension is \((m+1)\)

So, the worst case for the number of SVs is \((m+1)\)
How many SVs now?

Dimensionality:

\[ w = \sum_{i} \alpha_i y_i \Phi(x_i) \]
\[ b = y_k - w \Phi(x_k), \]
for any \( k \) where \( \alpha_k > 0 \)

\( w \) depends only on the alpha’s, not on the dim of \( x \)!
Quiz

\[(wx_j + b) \geq 1, \text{ for } y_i = +1\]
\[(wx_j + b) \leq -1, \text{ for } y_i = -1\]

Why 1 and -1?
Quiz

Can we apply a kernel to any algorithm?

SVM: \[
\text{minimize}_{w,b} \quad w \cdot w \\
(wx_j + b)y_j \geq 1, \forall j
\]

LR: \[
w^* = \arg \min_w \sum_j (y_j - \sum_i w_i h_i(x_j))^2
\]

Decision trees?
Boosting?
Quiz

Computing the b’s:

\[ b = y_k - w x_k, \text{ for any } k \text{ where } \alpha_k > 0 \]

Which k do we choose?
K-NN and homework problem

Cross-validation error
Training error
Testing error
Questions?