Recitation: Week 5

Language Technologies Institute
CMU School of Computer Science
Agenda

- Review of SVM
- Review of kernel
- Compare Classifiers
- Q&A of HW4
Central Limit Theorem

$$\sqrt{n} \left( \left( \frac{1}{n} \sum_{i=1}^{n} X_i \right) - \mu \right) \rightarrow \mathcal{N}(0, \sigma^2)$$
Central Limit Theorem

\[ \sqrt{n} \left( \left( \frac{1}{n} \sum_{i=1}^{n} X_i \right) - \mu \right) \rightarrow \mathcal{N}(0, \sigma^2) \]

i.i.d.

can be ANY distribution
Example – Exponential
Exponential $\rightarrow n=100$

\[
\sqrt{n} \left( \left( \frac{1}{n} \sum_{i=1}^{n} X_i \right) - \mu \right) \to \mathcal{N}(0, \sigma^2)
\]
For Binary Classifiers

$X_i$ is either 0 or 1  classification is either correct or incorrect
For Binary Classifiers

\[ X_i \text{ is either } 0 \text{ or } 1 \quad \rightarrow \quad \text{classification is either correct or incorrect} \]

classification for the classifier \( h \) can be considered as a

binary distribution with parameter \( p = \text{error}(h) \)
For Binary Classifiers

That’s the origin of

\[ p \pm z_n \sqrt{\frac{p(1-p)}{n}} \]

\[ E[\text{Binary}] = p \]

\[ \text{Var}[\text{Binary}] = p(1-p) \]
Example Question

You test two distinct classifiers ($h_1$ and $h_2$) on a dataset of 200 points.
The observed error rate for $h_1$ is 0.15
The observed error rate for $h_2$ is 0.19
Is $h_1$ statistically better than $h_2$ for a confidence level of 95%?
What about for 98%?

<table>
<thead>
<tr>
<th>$N%$</th>
<th>50%</th>
<th>68%</th>
<th>80%</th>
<th>90%</th>
<th>95%</th>
<th>98%</th>
<th>99%</th>
</tr>
</thead>
<tbody>
<tr>
<td>$z_N$</td>
<td>0.67</td>
<td>1.00</td>
<td>1.28</td>
<td>1.64</td>
<td>1.96</td>
<td>2.33</td>
<td>2.58</td>
</tr>
</tbody>
</table>
SVM
An intuitive example

max $d$
Separate points (cont’d)

\[ \{(x_i, y_i)\}_{i=1}^{N}, \quad x \in \mathbb{R}^m, \quad y \in \{-1, 1\} \]

\[ w^T x + b = 1 \]
\[ w^T x + b = 0 \]
\[ w^T x + b = -1 \]

\[ y(\mathbf{w}^T \mathbf{x} + b) \geq 1 \]
Margin

\[ 2d = \left( \frac{w}{\|w\|_2^2} \right)^T (x_1 - x_2) \]

\[ w^T x_1 + b = 1 \quad w^T x_2 + b = 1 \]

\[ 2d = \frac{2}{\|w\|_2} \]

\[ \text{margin} = \frac{2}{\|w\|_2} \]
SVM Formulation

\[ \text{min} \quad \frac{1}{2} \| \mathbf{w} \|_2^2 \]

\[ \text{s.t.} \quad \forall i, \quad y_i (\mathbf{w}^T \mathbf{x}_i + b) \geq 1 \]
SVM as a QP problem

\[
\begin{align*}
\min & \quad \frac{1}{2} u^T Ru + d^T u + c \\
\text{subject to n inequality constraints:} & \quad a_{11}u_1 + a_{12}u_2 + \ldots \leq b_1 \\
& \quad \vdots \quad \vdots \\
& \quad a_{n1}u_1 + a_{n2}u_2 + \ldots \leq b_n \\
\text{and k equivalency constraints:} & \quad a_{n+1,1}u_1 + a_{n+1,2}u_2 + \ldots = b_{n+1} \\
& \quad \vdots \\
& \quad a_{n+k,1}u_1 + a_{n+k,2}u_2 + \ldots = b_{n+k}
\end{align*}
\]

Min \( (w^T w)/2 \)

subject to the following inequality constraints:

For all \( x \) in class + 1
\[ w^Tx + b \geq 1 \]

For all \( x \) in class - 1
\[ w^Tx + b \leq -1 \]
Final optimization for non-linearly separable case

The new optimization problem is:

$$
\min_w \frac{1}{2} w^T w + \sum_{i=1}^{n} C \varepsilon_i
$$

subject to the following inequality constraints:

For all $x_i$ in class +1

$$w^T x + b \geq 1 - \varepsilon_i$$

For all $x_i$ in class -1

$$w^T x + b \leq -1 + \varepsilon_i$$

For all $i$

$$\varepsilon_i \geq 0$$

A total of $n$ constraints

Another $n$ constraints
Dual SVM for linearly separable case

Substituting $w$ into our target function and using the additional constraint we get:

\[
\max_{\alpha} \sum_{i} \alpha_i \left( \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j x_i x_j \right) \]

\[
\sum_{i} \alpha_i y_i = 0
\]

\[
\alpha_i \geq 0 \quad \forall i
\]

\[
\min_{w,b} \frac{w^T w}{2} - \sum_{i} \alpha_i [(w^T x_i + b)y_i - 1]
\]

\[
\alpha_i \geq 0 \quad \forall i
\]

\[
w = \sum_{i} \alpha_i x_i y_i
\]

\[
b = y_i - w^T x_i
\]

for $i$ s.t. $\alpha_i > 0$

\[
\sum_{i} \alpha_i y_i = 0
\]
Kernels
Why Kernel?

Figure 1: Transforming the data can make it linearly separable
Understanding Kernels

\[ K_{\text{lin}}(x_i, x_j) = x_i \cdot x_j \]

\[ K_{\text{poly}}(x_i, x_j) = (x_i \cdot x_j + 1)^p \]

\[ K_{\text{rbf}}(x_i, x_j) = \exp\left(-\frac{(x_i - x_j)^2}{s^2}\right) \]

\[ K_{\text{sig}}(x_i, x_j) = \tanh(s(x_i \cdot x_j) + c) \]

Can we use any function for kernel?
Mercer’s Theorem

A symmetric function $K(x, y)$ can be expressed as an inner product

$$K(x, y) = \langle \phi(x), \phi(y) \rangle$$

for some $\phi$ if and only if $K(x, y)$ is positive semi-definite, i.e.

$$\int K(x, y)g(x)g(y)dx\,dy \geq 0 \quad \forall g$$

Or, in other words

$$\begin{bmatrix}
K(x_1, x_1) & K(x_1, x_2) & \cdots \\
K(x_2, x_1) & \ddots & \\
\vdots & & \ddots
\end{bmatrix}
\text{ is psd for any collection } \{x_1 \ldots x_n\}$$
Demo of Polynomial Kernel

http://www.youtube.com/watch?v=3liCbRZPrZA
• Slides courtesy
  Prof. Aarti Singh and Eric Xing
  William Yang

• http://www.cs.cmu.edu/~epxing/Class/10701/Lecture/lecture9-svmII.pdf
  Page 40-41, 44-54

http://www.cs.cmu.edu/~epxing/Class/10701/Lecture/lecture8-svm.pdf
Page 14-15