Improved Natural Language Learning via Variance-Regularization Support Vector Machines

Shane Bergsma
University of Alberta

Dekang Lin
Google, Inc.

Dale Schuurmans
University of Alberta

1. Introduction

Multi-class classifiers with real-valued features (counts, probabilities, scores, etc.)

Often, good feature weights all have fairly similar values (just like some good unsupervised systems)

We’ll regularize (bias) an SVM so that it prefers weights that are similar (i.e. have low variance). It’ll learn faster with fewer examples!

2. Preposition Selection


Output: from (not during, in, on, etc.)

Features: \( x \): Counts of 34 different prepositions filling slot

Classifier: \( H : x \rightarrow y \) \( N \) features \( K \) classes

3. Features: N-gram counts

Ask the web! (Bergsma, Lin, Goebel, UCI AI 2009)

\[ x_{\text{from}} := (\text{Freq(worked in Sweden from)}, \text{Freq(worked in Sweden from 1997)}, \text{Freq(Sweden from 1997 to)}, \text{Freq(Sweden from 1997 to 2001)}), \ldots) \]

\( x_{\text{during}} := (\text{Freq(worked in Sweden during)}, \text{Freq(worked in Sweden during 1997)}, \text{Freq(worked in Sweden during 1997 to)}, \ldots) \)

\( x_{\text{in}} := (\text{Freq(worked in Sweden in)}, \text{Freq(worked in Sweden in 1997)}, \text{Freq(worked in Sweden in 1997 to)}, \ldots) \)

\[ x = (x_{\text{from}}, x_{\text{during}}, x_{\text{in}}, \ldots) \]

E.g.: \( x = (9.8, 4.1, 12.1, 9.3, 2.7, 8.3, 1.3, \ldots) \)

4. Unsupervised Classification

\[ H(\bar{x}) = \underset{r=1}{\overset{K}{\operatorname{arg\ max}}} \{ar{1} \cdot \bar{x}_r \} \]

5. Results

Preposition Selection

Unsupervised Accuracy is 73.7%

Outputs/fillers: (about, above, across, after, against, along, among, around, as, at, before, behind, beneath, beside, between, by, down, during, for, from, in, inside, into, like, of, off, on, onto, over, round, through, to, towards, with,)

Outputs/fillers: {about, above, across, after, against, along, among, around, as, at, before, behind, beneath, beside, between, by, down, during, for, from, in, inside, into, like, of, off, on, onto, over, round, through, to, towards, with,}

Context-Sensitive Spelling Correction

Unsupervised Accuracy is 94.8%

Example: “If you cite/site/sight/site me, then I’ll cite/site/sight/site you.”

Outputs/fillers: {among, between, amount, number, cite/site/sight/site, peace, piece, raise, rise}

Non-Referential Pronoun Detection

Unsupervised Accuracy is 80.1%

Example: “(It) looked as if the ball was in the net.”

Outputs = {referential/non-referential}, Fillers={it/its, they/their/their, all pronouns, *, <UNK>}

4. Multi-Class SVM Models

A) Standard k-SVM:

(Crammer & Singer, JMLR 2001)

\[ H_W(\bar{x}) = \underset{r=1}{\overset{K}{\operatorname{arg\ max}}} \{ \bar{W}_r \cdot \bar{x} \} \]

\[ \bar{W} \text{ from: } (+1, +2, +1, 0, 0, 0.1, -0.2, 0, \ldots) \]

\[ \bar{W} : (9.8, 4.1, 12.1, 9.3, 2.7, 8.3, 1.3, \ldots) \]

Training:

\[ \min_{\bar{W}, \xi^1, \ldots, \xi^M} \frac{1}{2} \sum_{i=1}^{K} ||\bar{W}_i||^2 + C \sum_{i=1}^{M} \xi^i \]

subject to:

\[ \forall r \neq y, \quad \bar{W}_y \cdot \bar{x}_r - \bar{W}_r \cdot \bar{x}_r \geq 1 - \xi^r \]

B) SVM with Class-Specific Features (CS-SVM):

\[ H_W(\bar{x}) = \underset{r=1}{\overset{K}{\operatorname{arg\ max}}} \{ \bar{W}_r \cdot \bar{x} \} \]

\[ \text{N weights: } \{ \bar{W}_{y^1}, \ldots, \bar{W}_{y^M} \} \]

\[ \min_{\bar{W}, \xi^1, \ldots, \xi^M} \frac{1}{2} \bar{w}^T \bar{w} + C \sum_{i=1}^{M} \xi^i \]

subject to:

\[ \forall r \neq y, \quad \bar{w}_{y^r} \cdot \bar{x}_r^r - \bar{w}_y \cdot \bar{x}_r^y \geq 1 - \xi^r \]

C) SVM with variance regularization (VR-SVM):

\[ \min_{\bar{w}, \xi^1, \ldots, \xi^M} \frac{1}{2} \bar{w}^T C \bar{w} + C \sum_{i=1}^{M} \xi^i \]

subject to:

\[ \forall r \neq y, \quad \bar{w}_{y^r} \cdot \bar{x}_r^r - \bar{w}_y \cdot \bar{x}_r^y \geq 1 - \xi^r \]

\[ C = \text{diag}(\bar{p}) - \bar{p} \bar{p}^T \]

\[ \text{Var}([W] = E[|W|^2] - E[|W|]^2) \]

5. Results

Number of Training Examples

<table>
<thead>
<tr>
<th>System</th>
<th>10</th>
<th>100</th>
<th>1000</th>
<th>10,000</th>
<th>100,000</th>
</tr>
</thead>
<tbody>
<tr>
<td>One-vs.-all SVM</td>
<td>16</td>
<td>50.6</td>
<td>66.1</td>
<td>71.1</td>
<td>73.5</td>
</tr>
<tr>
<td>Multi-class SVM</td>
<td>13.7</td>
<td>50.0</td>
<td>65.8</td>
<td>72.0</td>
<td>74.7</td>
</tr>
<tr>
<td>CS-SVM</td>
<td>27.1</td>
<td>58.8</td>
<td>69.0</td>
<td>73.5</td>
<td>74.2</td>
</tr>
<tr>
<td>VR-SVM</td>
<td>73.8</td>
<td>74.2</td>
<td>74.7</td>
<td>74.9</td>
<td>74.9</td>
</tr>
</tbody>
</table>

Number of Training Examples

<table>
<thead>
<tr>
<th>System</th>
<th>10</th>
<th>100</th>
<th>1000</th>
<th>10,000</th>
<th>100,000</th>
</tr>
</thead>
<tbody>
<tr>
<td>CS-SVM</td>
<td>86.0</td>
<td>93.5</td>
<td>95.1</td>
<td>95.7</td>
<td>95.7</td>
</tr>
<tr>
<td>VR-SVM</td>
<td>94.9</td>
<td>95.3</td>
<td>95.6</td>
<td>95.7</td>
<td>95.8</td>
</tr>
</tbody>
</table>

Example: “If you cite/site/sight/site me, then I’ll cite/site/sight/site you.”

Outputs/fillers: {among, between, amount, number, cite/site/sight/site, peace, piece, raise, rise}

Number of Training Examples

<table>
<thead>
<tr>
<th>System</th>
<th>10</th>
<th>100</th>
<th>1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>CS-SVM</td>
<td>59.0</td>
<td>71.0</td>
<td>84.3</td>
</tr>
<tr>
<td>VR-SVM</td>
<td>70.2</td>
<td>76.2</td>
<td>84.5</td>
</tr>
</tbody>
</table>