Structured SVM for Automatic Speech Recognition

S.-X. Austin Zhang and Mark Gales

February 1, 2013

Cambridge University Engineering Department
Outline

SVM for Speech Recognition

Structured SVM for Speech Recognition
  Joint features
  Training
  Decoding

Implementation and Experiments
Outline

SVM for Speech Recognition

Structured SVM for Speech Recognition
  Joint features
  Training
  Decoding

Implementation and Experiments
SVMs for Speech Recognition

SVMs generate boundaries between classes

---

![Diagram of SVMs generating boundaries between classes.](attachment:svm_diagram.png)

+ Class +1
O Class -1

Margin

Support Vectors

Optimal Hyperplane
SVMs for Speech Recognition

SVMs generate boundaries between classes

- For isolated digit recognition, no problem (10 classes)
- For continuous speech? too many classes! (6 digits $\Rightarrow 10^6$ classes)
SVMs for Speech Recognition

SVMs generate boundaries between classes

- For isolated digit recognition, no problem (10 classes)
- For continuous speech? too many classes! (6 digits \(\Rightarrow 10^6\) classes)

Simplest approach

**Step 1** Using HMM-based segmentation

**Step 2** For each segment, isolated classification
**SVMs for Speech Recognition**

**SVMs generate boundaries between classes**

- For isolated digit recognition, no problem (10 classes)
- For continuous speech? too many classes! (6 digits ⇒ $10^6$ classes)

**Simplest approach**

1. Using HMM-based segmentation
2. For each segment, isolated classification

**Problem:** Restricted to one fixed segmentation
SVM ⇒ Structured SVM

Incorporate structures into SVM classes ⇒ Structured SVM!

What is the structures?
Incorporate structures into SVM classes ⇒ Structured SVM!

What is the structures?

- Sequence structure in Handwriting Recognition
Where does the Structured SVM sit?

**Summary of generative, discriminative and discriminant models**

<table>
<thead>
<tr>
<th>Training</th>
<th>Unstructured $w$ → Structured $w$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ML</td>
<td></td>
</tr>
<tr>
<td>↓</td>
<td></td>
</tr>
<tr>
<td>CML</td>
<td></td>
</tr>
<tr>
<td>↓</td>
<td></td>
</tr>
<tr>
<td>Large Margin</td>
<td></td>
</tr>
</tbody>
</table>
Where does the Structured SVM sit?

Summary of generative, discriminative and discriminant models

<table>
<thead>
<tr>
<th>Training</th>
<th>Unstructured $w$ $\rightarrow$ Structured $w$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ML</td>
<td>Naive Bayes $P(O, w)$ $\rightarrow$ HMM $P(O, w)$</td>
</tr>
<tr>
<td>↓</td>
<td></td>
</tr>
<tr>
<td>CML</td>
<td></td>
</tr>
<tr>
<td>↓</td>
<td></td>
</tr>
<tr>
<td>Large Margin</td>
<td></td>
</tr>
</tbody>
</table>
Where does the Structured SVM sit?

Summary of generative, discriminative and discriminant models

<table>
<thead>
<tr>
<th>Training</th>
<th>Unstructured $w$ $\rightarrow$</th>
<th>Structured $w$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ML</td>
<td>Naive Bayes $P(O, w)$ $\rightarrow$ HMM $P(O, w)$</td>
<td></td>
</tr>
<tr>
<td>↓</td>
<td>↓</td>
<td>↓</td>
</tr>
<tr>
<td>CML</td>
<td>Logistic $P(w</td>
<td>O)$ $\rightarrow$ CRF $P(w</td>
</tr>
<tr>
<td>↓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Large Margin</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Where does the Structured SVM sit?

Summary of generative, discriminative and discriminant models

<table>
<thead>
<tr>
<th>Training</th>
<th>Unstructured $w$ $\rightarrow$ Structured $w$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ML</td>
<td>Naive Bayes $P(O, w)$ $\rightarrow$ HMM $P(O, w)$</td>
</tr>
<tr>
<td>CML</td>
<td>Logistic $P(w</td>
</tr>
<tr>
<td>Large Margin</td>
<td>SVM $\alpha_w^T \phi(O)$ $\rightarrow$ Structured SVM $\alpha^T \phi(O, w)$</td>
</tr>
</tbody>
</table>
Outline

SVM for Speech Recognition

**Structured SVM for Speech Recognition**
- Joint features
- Training
- Decoding

Implementation and Experiments
Model whole utterance – **observations** $O$, **word sequence** $w$

- *joint features* $\phi(O, w)$  How to build?
Structured SVM for Speech Recognition

Joint Features

Model whole utterance – observations $O$, word sequence $w$

- joint features $\phi(O, w)$

\[
\begin{bmatrix}
\log P(o; \lambda^{one}) \\
\vdots \\
\log P(o; \lambda^{zero})
\end{bmatrix}
\]

generative models

\[
\phi(O, w)
\]

- combine generative and discriminative
- model compensation
- language model features
Structured SVM for Speech Recognition

**Joint Features**

Model whole utterance – observations $O$, word sequence $w$

- **Joint features** $\phi(O, w)$

$$
\begin{bmatrix}
\log P(o; \lambda^{one}) \\
\vdots \\
\log P(o; \lambda^{zero})
\end{bmatrix}
$$

- Decoding by match score: $\arg\max_w \alpha^T \phi(O, w)$
Structured SVM for Speech Recognition

Training

Training $\alpha$, Maximize the Margin
Subject to: score of correct $w_{\text{ref}}$ $\geq$ all competing $w$

$$\begin{align*}
\min_{\alpha, \xi} & \quad \frac{1}{2} \|\alpha\|^2 \\
\text{s.t.} & \quad \alpha^T \phi(\text{Sample 1}, "1 2 3") \geq \alpha^T \phi(\text{Sample 1}, "0 0 0") + 1, \\
& \quad \alpha^T \phi(\text{Sample 1}, "1 2 3") \geq \alpha^T \phi(\text{Sample 1}, "0 0 1") + 1, \\
& \quad \alpha^T \phi(\text{Sample 1}, "1 2 3") \geq \alpha^T \phi(\text{Sample 1}, "9 9 9") + 1,
\end{align*}$$

\[
\vdots
\]

$$\begin{align*}
\min_{\alpha, \xi} & \quad \frac{1}{2} \|\alpha\|^2 \\
\text{s.t.} & \quad \alpha^T \phi(\text{Sample n}, "4 5 6") \geq \alpha^T \phi(\text{Sample n}, "0 0 0") + 1, \\
& \quad \alpha^T \phi(\text{Sample n}, "4 5 6") \geq \alpha^T \phi(\text{Sample n}, "0 0 1") + 1, \\
& \quad \alpha^T \phi(\text{Sample n}, "4 5 6") \geq \alpha^T \phi(\text{Sample n}, "9 9 9") + 1,
\end{align*}$$

\[
\vdots
\]
To generalize the training

- Replace “0-1 loss” as $\mathcal{L}(\mathbf{w}_{\text{ref}}, \mathbf{w})$
- Introduce slack variable $\xi_i$

$$\begin{align*}
\min_{\alpha, \xi} & \quad \frac{1}{2} \|\alpha\|^2 + C \sum_{i=1}^{n} \xi_i \\
\text{s.t.} & \quad \alpha^T \phi(\text{“1 2 3”}) \geq \max_{\mathbf{w} \neq \text{“1 2 3”}} \left\{ \alpha^T \phi(\text{“1 2 3”}, \mathbf{w}) + \mathcal{L}(\text{“1 2 3”}, \mathbf{w}) \right\} - \xi_1
\end{align*}$$

Unconstrained form

$$\begin{align*}
\frac{1}{2} \|\alpha\|^2 + C \sum_{i=1}^{n} \left[ - \alpha^T \phi(\mathbf{O}^{(i)}, \mathbf{w}^{(i)}_{\text{ref}}) \right] + \max_{\mathbf{w} \neq \mathbf{w}_{\text{ref}}} \left\{ \mathcal{L}(\mathbf{w}, \mathbf{w}^{(i)}_{\text{ref}}) + \alpha^T \phi(\mathbf{O}^{(i)}, \mathbf{w}) \right\} + \text{one - one \mathbf{w}}
\end{align*}$$
Structured SVM for Speech Recognition

Training

Training $\alpha$ in QP form

$$\min_{\alpha, \xi} \frac{1}{2} \|\alpha\|^2 + \frac{C}{n} \sum_{i=1}^{n} \xi_i$$

s.t. for every utterance $i$, for ALL the competing labels $w^{(i)}$

$$\alpha^T \phi(O^{(i)}, w_{\text{ref}}^{(i)}) - \alpha^T \phi(O^{(i)}, w^{(i)}) \geq \mathcal{L}(w_{\text{ref}}^{(i)}, w^{(i)}) - \xi_i$$

Handling infinite number of constraints - Cutting Plane Algorithm
Structured SVM for Speech Recognition

Training

Cutting Plane Algorithm

repeat
    for $i=1,\ldots,n$ do
        ① constraint set ← Generate a new constraint ($w_\ast^{(i)}$)
        ② $\alpha$ ← Solving QP with current constraint set
    end for
until no new constraints
Structured SVM for Speech Recognition

Decoding

- Decoding by match score: \[ \arg \max_w \alpha^T \phi(O, w) \]

- Search based on lattices
Structured SVM ≡ Large Margin Log Linear Model

Example of log-linear models

\[ P(w|O; \alpha) = \frac{1}{Z} \exp (\alpha^T \phi(O, w)) \]

- **Same Decoding:** \( \arg \max_w P(w|O; \alpha) = \arg \max_w \alpha^T \phi(O, w) \)

- **Becomes Structured SVM in Large Margin Training**

  - Margin: minimum distance between correct \( w_{\text{ref}} \) and competing \( w \)
    \[
    \min_{w \neq w_{\text{ref}}} \left\{ \log \left( \frac{P(w_{\text{ref}}|O; \alpha)}{P(w|O; \alpha)} \right) \right\}
    \]

  - Incorporate log \( P(\alpha) \propto \frac{1}{2} ||\alpha||^2 \)
Outline

SVM for Speech Recognition

Structured SVM for Speech Recognition
  Joint features
  Training
  Decoding

Implementation and Experiments
Practical Issues

Optimize segmentation

- introduce latent variables – Concave-Convex optimisation

Joint feature space very large

- Parameter tying – context-dependent phones
- Kernelization

Efficiency

- $n$-slack $\Rightarrow$ 1-slack ($\xi = \sum_{i=1}^{n} \xi_i$)
- Modifying Prior
- Parallelization
- Caching and Pruning
Implementation

Training Phase

Decoding Phase

Generating Numerator Lattice

Generating Denominator Lattice

Generating Denominator Lattice

Search \textit{reference segmentation}

Search \textit{competing $W$ and segmentation}

Search \textit{best $W$ and segmentation}

Large Margin Training $\alpha$

$W^{(i)}_{\text{ref}} \rightarrow \text{Generating Numerator Lattice}$

$O^{(i)} \rightarrow \text{Generating Denominator Lattice}$

$O \rightarrow \text{Generating Denominator Lattice}$

$\hat{W}$

$\alpha$

Updated $\alpha$

Updated $\alpha$

Parallelized

Parallelized
Experiments

Aurora 2: Noise corrupted continuous digit task.

<table>
<thead>
<tr>
<th>Model</th>
<th>Parameter</th>
<th>Criteria</th>
<th>Err. %</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMM</td>
<td>46,732</td>
<td>ML</td>
<td>9.5</td>
</tr>
<tr>
<td>SVM</td>
<td>+144</td>
<td>LM</td>
<td>8.3</td>
</tr>
<tr>
<td>Log Linear</td>
<td>+144</td>
<td>CML</td>
<td>8.1</td>
</tr>
<tr>
<td>Struct SVM</td>
<td>+144</td>
<td>LM</td>
<td>7.4</td>
</tr>
</tbody>
</table>

Aurora 4: Noise corrupted medium to large vocabulary task.

<table>
<thead>
<tr>
<th>Model</th>
<th>Parameter</th>
<th>Criteria</th>
<th>Err. %</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMM</td>
<td>3.98M</td>
<td>ML</td>
<td>17.8</td>
</tr>
<tr>
<td>Log Linear</td>
<td>+2210</td>
<td>MPE</td>
<td>17.4</td>
</tr>
<tr>
<td>Struct SVM</td>
<td>+2210</td>
<td>LM</td>
<td>16.8</td>
</tr>
</tbody>
</table>

• Gain over HMMs: 22% in Aurora 2, 6% in Aurora 4
Structured SVMs ≡ Large Margin Log Linear Models

Structured SVM for ASR

- Joint Features
- Training and Decoding
- Practical Issues

Good performance