Marginal Structured SVM with Hidden Variables

Wei Ping*, Qiang Liu, Alexander Ihler

Department of Computer Science
University of California, Irvine
June 22nd, 2014
Outline

- Structured Prediction
  - Hidden / latent variables
- Marginal Structured SVM
- A Unified Framework
- Empirical Results
Structured Prediction

- Supervised learning
  - Learn predictor $f: X \rightarrow Y$, using $(x_1, y_1)...(x_n, y_n)$
  - Structured prediction: $Y = \{1, \ldots, k\}^d$
    - Key: correlated output variables
- Image segmentation: classify pixels into semantic categories
Structured Prediction: State-of-the-art

- Conditional random fields (CRFs) (Lafferty et al., ICML 2001)
  - Maximum likelihood; asymptotically consistent

- Structured SVM (SSVM) (Tsochantaridis et al., JMLR 2005)
  - Empirical loss minimization
    - Task specific loss
  - Often empirically better, especially
    - training set is relatively small
    - model assumptions are violated
Hidden Variables: Missing Values

- Image segmentation
  - Label every pixel… **Expensive!**
  - Label ambiguous regions… **Difficult!**
  - Partially labeled datasets, e.g. MSRC dataset (Winn et al., ICCV 2005)

![Image of cow](image.png)

![Graph representation](graph.png)
Hidden Variables: Capturing Dependence

- Adding hidden variables can enrich the model’s flexibility

Efficiently introduces higher order dependencies
Model Setting

- Undirected conditional model with hidden variables

\[ p(y, h|x; w) = \frac{1}{Z(x; w)} \exp \left[ w^T \phi(x, y, h) \right] \]

- \( x \): input variables
- \( y \): output variables
- \( h \): hidden variables
- \( \phi(x, y, h) \): joint feature map; relationships between \( x, y, h \)
- \( w \): weights, try to estimate in learning
- \( Z(x; w) \): partition function
Previous Learning Methods

- Hidden CRFs (HCRF) (Quattoni et al., PAMI 2006)
  - Extension of CRFs; maximum marginal likelihood
  - Smooth

- Latent Structured SVM (LSSVM) (Yu et al., ICML 2009)
  - Extension of SSVM; surrogate loss
  - Often empirically better
  - Piece-wise linear
    - Gradient based optimizations are slow
  - Relies on joint MAP inference during learning…
Inference with Hidden Variables

- During learning, inference is used to “fill in” hidden variables.

- Joint MAP: \( [\hat{y}(w), \hat{h}(w)] = \arg\max_{y, h} w^T \phi(x, y, h) \)
  - Impute ambiguous region \((h)\) with most probable assignment.
  - Does not maintain uncertainty of \(h\)!

- Marginal MAP: \( \hat{y}(w) = \arg\max_{y \in \mathcal{Y}} \log \sum_{h} \exp [w^T \phi(x, y, h)] \)
  - More robust when the uncertainty in \(h\) is large (Liu & Ihler JMLR 2013)
Overview

- Marginal Structured SVM (MSSVM):
  - Properly handle the uncertainty of h
    - incorporating marginal MAP in max-margin learning
  - Significantly outperforms LSSVM
    - uncertainty of h is high
  - Consistently outperforms HCRF
  - Much smoother objective than LSSVM
    - gradient-based optimization are more efficient
Empirical loss can’t be directly minimized
- Discontinuous with w (hamming loss)

Surrogate loss given marginal MAP prediction

\[
\hat{y}_i(w) = \arg \max_{y \in \mathcal{Y}} \log \sum_h \exp \left[ w^T \phi(x_i, y, h) \right]
\]

\[
\Delta(y_i, \hat{y}_i(w)) \leq \Delta(y_i, \hat{y}_i(w)) + \log \sum_h \exp \left[ w^T \phi(x_i, \hat{y}_i(w), h) \right]
- \log \sum_h \exp \left[ w^T \phi(x_i, y_i, h) \right]
\]

\[
\leq \max_y \left\{ \Delta(y_i, y) + \log \sum_h \exp \left[ w^T \phi(x_i, y, h) \right] \right\}
- \log \sum_h \exp \left[ w^T \phi(x_i, y_i, h) \right]
\]
Objective Function

- **MSSVM:**

\[
\frac{1}{2} \| w \|^2 + C \sum_{i=1}^{n} \max_{y} \left\{ \Delta(y_i, y) + \log \sum_{h} \exp[w^T \phi(x_i, y, h)] \right\} - C \sum_{i=1}^{n} \log \sum_{h} \exp \left[ w^T \phi(x_i, y_i, h) \right]
\]

- **Compared to LSSVM:**

\[
\frac{1}{2} \| w \|^2 + C \sum_{i=1}^{n} \max_{y} \max_{h} \left\{ \Delta(y_i, y) + [w^T \phi(x_i, y, h)] \right\} - C \sum_{i=1}^{n} \max_{h} \left[ w^T \phi(x_i, y_i, h) \right]
\]

- “hard” prediction of h
- “soft-max” (log-sum-exp) over h
A Unified Framework

- Introduce a “temperature” parameter that smooth between max and soft-max operator

\[
\epsilon \log \sum_h \exp \left[ \frac{f(h)}{\epsilon} \right]
\]

- Approach to max operator if \( \epsilon \to 0^+ \)
- Reduces to log-sum-exp if \( \epsilon = 1 \)
A Unified Framework

- Two “temperature” $\epsilon_y$ and $\epsilon_h$ for $y$ and $h$,

$$\frac{1}{2} \| w \|^2 + C \cdot \sum_{i=1}^{n} \epsilon_y \log \sum_y \exp \left\{ \frac{1}{\epsilon_y} \left[ \Delta(y_i, y) + \epsilon_h \log \sum_h \exp \left( \frac{w^T \phi(x_i, y, h)}{\epsilon_h} \right) \right] \right\}$$

$$- C \cdot \sum_{i=1}^{n} \epsilon_h \log \sum_h \exp \left( \frac{w^T \phi(x_i, y_i, h)}{\epsilon_h} \right)$$

<table>
<thead>
<tr>
<th>Model</th>
<th>$\epsilon_h \to 0^+ (\max_h)$</th>
<th>$\epsilon_h = 1 (\sum_h)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\epsilon_y \to 0^+ (\max_y)$</td>
<td>LSSVM</td>
<td>MSSVM</td>
</tr>
<tr>
<td>$\epsilon_y = 1 (\sum_y)$</td>
<td>N/A</td>
<td>HCRF</td>
</tr>
<tr>
<td>$\epsilon_h = \epsilon_y \in (0, 1)$</td>
<td>$\epsilon$-extension model (Schwing et al., ICML 2012)</td>
<td></td>
</tr>
</tbody>
</table>
Training Algorithms

- **MSSVM:**

\[
\frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} \max_y \left\{ \Delta(y_i, y) + \log \sum_h \exp[w^T \phi(x_i, y, h)] \right\} - C \sum_{i=1}^{n} \log \sum_h \exp \left[ w^T \phi(x_i, y_i, h) \right]
\]

- **Sub-gradient decent (SGD):**

\[
\nabla_w M = w + C \sum_{i=1}^{n} \mathbb{E}_{p(h|x_i, \hat{y}_i)}[\phi(x_i, \hat{y}_i, h)] - C \sum_{i=1}^{n} \mathbb{E}_{p(h|x_i, y_i)}[\phi(x_i, y_i, h)]
\]

where, \( \hat{y}_i = \arg\max_{y \in Y} \left\{ \Delta(y_i, y) + \log \sum_h \exp[w^T \phi(x_i, y, h)] \right\} \)

- 1\(^{st}\) expectation: marginals from mixed-product BP (Liu & Ihler JMLR 2013)
- 2\(^{nd}\) expectation: marginals from sum-product BP (Kschischang, et al., TIT 2001)
Concave-convex procedure

- **MSSVM**: differences of two convex functions,

\[
\frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} \max_{y} \left\{ \Delta(y_i, y) + \log \sum_{h} \exp[w^T \phi(x_i, y, h)] \right\} - C \sum_{i=1}^{n} \log \sum_{h} \exp \left[w^T \phi(x_i, y_i, h) \right]
\]

\[f^+(w)\]

\[f^-(w)\]

- **CCCP updating rule**:
  - minimize convex surrogate, where \( f^-(w) \) is linearized:

\[
w^{t+1} \leftarrow \arg \min_w \{ f^+(w) - w^T \nabla f^-(w^t) \}
\]

\[
\nabla f^-(w^t) = C \sum_{i=1}^{n} \mathbb{E}_{p(h|x_i,y_i)}[\phi(x_i,y_i,h)] \text{ is the gradient at } w^t
\]
Experiments: Simulated Data

- Simulated data from pair-wise MRF on graph \( G = (V, E) \)

  ![Diagram of hidden chain and 2-D grid]

- hidden chain
- 2-D grid

- discrete variables
- mixed interactions
- log-potentials sampled from Gaussian
Simulated Data

Averaged accuracy(%) 

<table>
<thead>
<tr>
<th>Hidden Chain</th>
<th>MSSVM</th>
<th>LSSVM</th>
<th>HCRFs</th>
</tr>
</thead>
<tbody>
<tr>
<td>SGD</td>
<td>69.20</td>
<td>66.87</td>
<td>68.75</td>
</tr>
<tr>
<td>CCCP</td>
<td>69.63</td>
<td>67.91</td>
<td>69.03</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>2D-grid graph</th>
<th>MSSVM</th>
<th>LSSVM</th>
<th>HCRFs</th>
</tr>
</thead>
<tbody>
<tr>
<td>SGD</td>
<td>74.12</td>
<td>71.96</td>
<td>73.51</td>
</tr>
<tr>
<td>CCCP</td>
<td>74.08</td>
<td>73.38</td>
<td>73.62</td>
</tr>
</tbody>
</table>

- MSSVM has highest prediction accuracy
- CCCP is better than SGD for LSSVM
  - SGD’s difficult to converge on non-smooth LSSVM
- Both SGD and CCCP perform well for MSSVM
Uncertainty of Hidden Variables

- Adjust uncertainty of hidden variables

MSSVM is significant better than LSSVM as uncertainty increases.

<table>
<thead>
<tr>
<th></th>
<th>MSSVM</th>
<th>LSSVM</th>
<th>HCRF₁</th>
<th>M3E</th>
<th>ModLat</th>
</tr>
</thead>
<tbody>
<tr>
<td>79.30</td>
<td>79.46</td>
<td>78.68</td>
<td>79.04</td>
<td>77.16</td>
<td></td>
</tr>
<tr>
<td>70.00</td>
<td>70.07</td>
<td>69.88</td>
<td>68.53</td>
<td>67.91</td>
<td></td>
</tr>
<tr>
<td>67.24</td>
<td>65.98</td>
<td>66.66</td>
<td>66.05</td>
<td>65.15</td>
<td></td>
</tr>
<tr>
<td>69.63</td>
<td>67.91</td>
<td>69.03</td>
<td>65.19</td>
<td>67.96</td>
<td></td>
</tr>
<tr>
<td>73.88</td>
<td>71.38</td>
<td>72.58</td>
<td>67.21</td>
<td>71.52</td>
<td></td>
</tr>
<tr>
<td>72.08</td>
<td>69.24</td>
<td>70.88</td>
<td>65.48</td>
<td>66.54</td>
<td></td>
</tr>
</tbody>
</table>

3/11/2014
Image Segmentation

- **Grid model**
  - 20 × 40 pixels, $y_i$ in \{1, ..., 5\}
  - $x$ is obtained by adding noise $N(0, 5)$ to $y$
  - Some pixels marked as missing ($h$), at random
Experiment: Object Categorization

- MSRC data: partial labeling of pixels to categories
  - Missing labels: usually undefined categories
  - ambiguous regions
  - boundaries

- Important to maintain uncertainty during learning!

3/11/2014

Features etc. modeled from Verbeek et al., NIPS 2007.

<table>
<thead>
<tr>
<th>MSRC Data</th>
<th>MSSVM</th>
<th>LSSVM</th>
<th>HCRFs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Building</td>
<td>72.4</td>
<td>70.7</td>
<td>71.7</td>
</tr>
<tr>
<td>Grass</td>
<td>89.7</td>
<td>88.9</td>
<td>88.3</td>
</tr>
<tr>
<td>Sky</td>
<td>88.3</td>
<td>85.6</td>
<td>88.2</td>
</tr>
<tr>
<td>Tree</td>
<td>71.9</td>
<td>71.0</td>
<td>70.1</td>
</tr>
<tr>
<td>Car</td>
<td>70.8</td>
<td>69.4</td>
<td>70.2</td>
</tr>
</tbody>
</table>
Thanks!

谢谢！
Experiment: Empirical Convergence

- LSSVM’s hard-max makes the objective piecewise linear
  - MSSVM are smoother by marginalizing hidden variables
  - SGD on LSSVM need much smaller learning rate than MSSVM, and converge much slower
  - Set appropriate learning rates, which make them tend to converge

- CCCP on LSSVM converging faster than SGD,
  - transform complex piecewise linear into a sequence of easier convex sub-problems
Experiment: Training Sample Size

- HCRF is MLE, implies
  - Asymptotically “consistent”;
  - Statistically efficient

- Given a correct model,

  MSSVM is the best under reasonable large data sizes (1024 training examples and ~150 free parameters)

  HCRF continue to improve as training size increase.

  High dimensional parameters & few training sample in structured prediction