Softmax-Margin CRFs:
Training Log-Linear Models with Cost Functions

Kevin Gimpel and Noah A. Smith
Risk Perceptron

Minimum Error Rate Training

Conditional Likelihood

Max-Margin

MIRA

Boosting

Latent Variable

Conditional Likelihood

Based on probabilistic inference

Uses a cost function

Is convex

Minimum Error Rate Training
**Risk Perceptron**
- **Minimum Error Rate Training**
- **Conditional Likelihood**
- **Max-Margin**
- **MIRA**
- **Boosting**
- **Is convex**
- **Based on probabilistic inference**
- **Uses a cost function**
- **Softmax-Margin**
- **Latent Variable Conditional Likelihood**
- **Minimum Error Rate Training**

**Inference**

**Carnegie Mellon**

**Ark**

**lti**
Risk Perceptron

Minimum Error Rate Training

Conditional Likelihood

Max-Margin

Latent Variable Conditional Likelihood

Softmax-Margin

Jensen Risk Bound

Minimum Error Rate Training

Based on probabilistic inference

Uses a cost function

Is convex

Boosting

Perceptron
Linear Models for Structured Prediction

\[ \hat{y} = \arg\max_{y \in \mathcal{Y}(x)} \theta^\top f(x, y) \]

For probabilistic interpretation, exponentiate and normalize:

\[
p_\theta(y|x) = \frac{\exp\{\theta^\top f(x, y)\}}{\sum_{y' \in \mathcal{Y}(x)} \exp\{\theta^\top f(x, y')\}}
\]
Training

- Standard approach is to maximize conditional likelihood:

\[
\min_{\theta} \sum_{i=1}^{n} \left( -\theta^T f(x^{(i)}, y^{(i)}) + \log \sum_{y \in \mathcal{Y}(x^{(i)})} \exp\{\theta^T f(x^{(i)}, y)\} \right)
\]

- Another approach maximizes margin (Taskar et al., 2003):

\[
\min_{\theta} \sum_{i=1}^{n} \left( -\theta^T f(x^{(i)}, y^{(i)}) + \max_{y \in \mathcal{Y}(x^{(i)})} \left( \theta^T f(x^{(i)}, y) + \text{cost}(y^{(i)}, y) \right) \right)
\]

\textit{task-specific cost function}
Training

[Box] Standard approach is to maximize conditional likelihood:

$$\min_{\theta} \sum_{i=1}^{n} \left( -\theta^\top f(x^{(i)}, y^{(i)}) + \log \sum_{y \in \mathcal{Y}(x^{(i)})} \exp\{\theta^\top f(x^{(i)}, y)\} \right)$$

[Box] Another approach maximizes margin (Taskar et al., 2003):

$$\min_{\theta} \sum_{i=1}^{n} \left( -\theta^\top f(x^{(i)}, y^{(i)}) + \max_{y \in \mathcal{Y}(x^{(i)})} \left( \theta^\top f(x^{(i)}, y) + \text{cost}(y^{(i)}, y) \right) \right)$$

(cost-augmented decoding)
Training

■ Standard approach is to maximize conditional likelihood:

\[
\min_{\theta} \sum_{i=1}^{n} \left( -\theta^\top f(x^{(i)}, y^{(i)}) + \log \sum_{y \in \mathcal{Y}(x^{(i)})} \exp\{\theta^\top f(x^{(i)}, y)\} \right)
\]

■ Another approach maximizes margin (Taskar et al., 2003):

\[
\min_{\theta} \sum_{i=1}^{n} \left( -\theta^\top f(x^{(i)}, y^{(i)}) + \max_{y \in \mathcal{Y}(x^{(i)})} \left( \theta^\top f(x^{(i)}, y) + \text{cost}(y^{(i)}, y) \right) \right)
\]

■ Softmax-margin: replace “max” with “softmax”

\[
\min_{\theta} \sum_{i=1}^{n} \left( -\theta^\top f(x^{(i)}, y^{(i)}) + \log \sum_{y \in \mathcal{Y}(x^{(i)})} \exp\{\theta^\top f(x^{(i)}, y) + \text{cost}(y^{(i)}, y)\} \right)
\]

“cost-augmented summing”
Training

- Standard approach is to maximize conditional likelihood:

\[
\min_\theta \sum_{i=1}^{n} \left( -\theta^\top f(x^{(i)}, y^{(i)}) + \log \sum_{y \in \mathcal{Y}(x^{(i)})} \exp\{\theta^\top f(x^{(i)}, y)\} \right)
\]

- Another approach maximizes margin (Taskar et al., 2003):

\[
\min_\theta \sum_{i=1}^{n} \left( -\theta^\top f(x^{(i)}, y^{(i)}) + \max_{y \in \mathcal{Y}(x^{(i)})} \left( \theta^\top f(x^{(i)}, y) + \text{cost}(y^{(i)}, y) \right) \right)
\]

- Softmax-margin: replace “max” with “softmax”

\[
\min_\theta \sum_{i=1}^{n} \left( -\theta^\top f(x^{(i)}, y^{(i)}) + \log \sum_{y \in \mathcal{Y}(x^{(i)})} \exp\{\theta^\top f(x^{(i)}, y) + \text{cost}(y^{(i)}, y)\} \right)
\]

Sha and Saul (2006), Povey et al. (2008)
Properties of Softmax-Margin

- Has a probabilistic interpretation in the minimum divergence framework (Jelinek, 1997)
  - Details in technical report

- Is a bound on:
  - Max-margin
  - Conditional likelihood
  - Risk
Properties of Softmax-Margin

- Has a probabilistic interpretation in the minimum divergence framework (Jelinek, 1997)
  - Details in technical report

- Is a bound on:
  - Max-margin (because “softmax” bounds “max”) ✓
  - Conditional likelihood
  - Risk
Risk?

- **Risk** is the expected value of the cost function (Smith and Eisner, 2006; Li and Eisner, 2009):

\[
\min_{\theta} \sum_{i=1}^{n} \mathbb{E}_{p_{\theta}(\cdot|x^{(i)})}[\text{cost}(y^{(i)}, \cdot)]
\]
Bounding Conditional Likelihood and Risk

Softmax-margin:

\[
\sum_{i=1}^{n} \left( -\theta^\top f(x^{(i)}, y^{(i)}) + \log \sum_{y \in \mathcal{Y}(x^{(i)})} \exp\{\theta^\top f(x^{(i)}, y) + \text{cost}(y^{(i)}, y)\} \right)
\]

\[
= \sum_{i=1}^{n} \left( -\theta^\top f(x^{(i)}, y^{(i)}) + \log Z_i \right) + \sum_{i=1}^{n} \log \mathbb{E}_{p_i} [\exp\{\text{cost}(y^{(i)}, \cdot)\}]
\]

- Conditional likelihood
- Bound on risk via Jensen’s inequality
Bounding Conditional Likelihood and Risk

Softmax-margin:

\[
\sum_{i=1}^{n} \left( -\theta^\top f(x^{(i)}, y^{(i)}) + \log \sum_{y \in Y(x^{(i)})} \exp\{\theta^\top f(x^{(i)}, y) + \text{cost}(y^{(i)}, y)\} \right)
\]

\[
= \sum_{i=1}^{n} \left( -\theta^\top f(x^{(i)}, y^{(i)}) + \log Z_i \right) + \sum_{i=1}^{n} \log \mathbb{E}_{p_i}[^{\exp\{\text{cost}(y^{(i)}, \cdot)\}}]
\]

- Conditional likelihood
- Bound on risk via Jensen’s inequality

Softmax-margin is a convex bound on max-margin, conditional likelihood, and risk
Bounding Conditional Likelihood and Risk

- **Softmax-margin:**

\[
\sum_{i=1}^{n} \left( -\theta^T f(x^{(i)}, y^{(i)}) + \log \sum_{y \in y(x^{(i)})} \exp \{ \theta^T f(x^{(i)}, y) + \text{cost}(y^{(i)}, y) \} \right)
\]

\[
= \sum_{i=1}^{n} \left( -\theta^T f(x^{(i)}, y^{(i)}) + \log Z_i \right) + \sum_{i=1}^{n} \log \mathbb{E}_{p_i}[\exp \{ \text{cost}(y^{(i)}, \cdot) \}]
\]

**Conditional likelihood**

**Jensen Risk Bound**

Easier to optimize than risk (cf. Li and Eisner, 2009)
Implementation

- Conditional likelihood → Softmax-margin
  - If cost function factors the same way as the features, it’s easy:
    - Add additional features for the cost function
    - Keep their weights fixed
  - If not, use a simpler cost function or use approximate inference
Experiments

- English named-entity recognition (CoNLL 2003)

- Compared softmax-margin and Jensen risk bound with five baselines:
  - Perceptron (Collins, 2002)
  - 1-best MIRA with cost-augmented decoding (Crammer et al., 2006)
  - Max-margin via subgradient descent (Ratliff et al., 2006)
  - Conditional likelihood (Lafferty et al., 2001)
  - Risk (Xiong et al., 2009)

- For risk and Jensen risk bound, initialized using output of conditional likelihood training

- Used Hamming cost for cost function
## Results

<table>
<thead>
<tr>
<th>Method</th>
<th>Test $F_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Perceptron</td>
<td>83.98*</td>
</tr>
<tr>
<td>MIRA</td>
<td>85.72</td>
</tr>
<tr>
<td>Max-Margin</td>
<td>85.28*</td>
</tr>
<tr>
<td>Conditional Likelihood</td>
<td>85.46*</td>
</tr>
<tr>
<td>Risk</td>
<td>85.59</td>
</tr>
<tr>
<td>Jensen Risk Bound</td>
<td>85.65</td>
</tr>
<tr>
<td>Softmax-Margin</td>
<td>85.84</td>
</tr>
</tbody>
</table>

* Indicates significance (compared with softmax-margin)
## Results

<table>
<thead>
<tr>
<th>Method</th>
<th>Test F₁</th>
</tr>
</thead>
<tbody>
<tr>
<td>Perceptron</td>
<td>83.98*</td>
</tr>
<tr>
<td>MIRA</td>
<td>85.72</td>
</tr>
<tr>
<td>Max-Margin</td>
<td>85.28*</td>
</tr>
<tr>
<td><strong>Conditional Likelihood</strong></td>
<td><strong>85.46</strong>*</td>
</tr>
<tr>
<td>Risk</td>
<td>85.59</td>
</tr>
<tr>
<td>Jensen Risk Bound</td>
<td>85.65</td>
</tr>
<tr>
<td>Softmax-Margin</td>
<td>85.84</td>
</tr>
</tbody>
</table>

* Indicates significance (compared with softmax-margin)

Significant improvement with equal training time and implementation difficulty
## Results

<table>
<thead>
<tr>
<th>Method</th>
<th>Test F₁</th>
</tr>
</thead>
<tbody>
<tr>
<td>Perceptron</td>
<td>83.98*</td>
</tr>
<tr>
<td>MIRA</td>
<td>85.72</td>
</tr>
<tr>
<td>Max-Margin</td>
<td>85.28*</td>
</tr>
<tr>
<td>Conditional Likelihood</td>
<td>85.46*</td>
</tr>
<tr>
<td>Risk</td>
<td>85.59</td>
</tr>
<tr>
<td>Jensen Risk Bound</td>
<td>85.65</td>
</tr>
<tr>
<td>Softmax-Margin</td>
<td>85.84</td>
</tr>
</tbody>
</table>

* Indicates significance (compared with softmax-margin)

Comparable performance with half the training time
Risk
Conditional Likelihood
Max-Margin
MIRA
Uses a cost function
Is convex
Based on probabilistic inference
Perceptron
Softmax-Margin
Jensen Risk Bound
Risk
Based on probabilistic inference
Uses a cost function
Is convex
Thank you!

- See extended technical report for:
  
  - Probabilistic interpretation for softmax-margin in minimum divergence framework (Jelinek, 1997)
  
  - Softmax-margin training with hidden variables
  
  - Additional experiments
Loss Functions for Binary Classification

- **Softmax-Margin**: $\log(1 + \exp(m - z))$
- **Max-Margin**: $\max(0, m - z)$
- **Conditional Likelihood**: $\log(1 + \exp(-z))$
- **0-1**: $m \times I(z \leq 0)$
- **Jensen Risk Bound**: $\log\left(\frac{1 + \exp(m - z)}{1 + \exp(-z)}\right)$
- **Risk**: $\frac{m \exp(-z)}{1 + \exp(-z)}$
<table>
<thead>
<tr>
<th>Training Method</th>
<th>Requirements</th>
<th>Cost Function</th>
<th>Convex</th>
<th>Prob. Interp.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Perceptron</td>
<td>decoding</td>
<td></td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>MIRA</td>
<td>cost-augmented decoding</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Max-Margin</td>
<td>cost-augmented decoding</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Conditional Likelihood</td>
<td>summing</td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Risk</td>
<td>expectations of products</td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td><strong>Jensen Risk Bound</strong></td>
<td>cost-augmented summing</td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td><strong>Softmax-Margin</strong></td>
<td>cost-augmented summing</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>