Summary

Task: structured output learning when each input can have several correct outputs.

Examples: protein folds, graph matchings with symmetry, object locations in images, natural language translations, ...

Contribution: max-margin formulation
• convex,
• efficient by use of working sets,
• tractable even for exponentially sized label sets,
• PAC-Bayesian generalization bound.

→ essentially same as single-label SSVM [1].

Idea: build on one-versus-rest SVM instead of Crammer-Singer multiclass SVM, as SSVM does.

Common Notation
• \( \mathcal{X} \) inputs (arbitrary)
• \( \mathcal{Y} \) structured labels, e.g. graph-labelings, parse trees, ...
• \( k : (\mathcal{X} \times \mathcal{X}) \times (\mathcal{X} \times \mathcal{X}) \to \mathbb{R} \): joint kernel function,
  \( \phi : \mathcal{X} \to \mathbb{R}^k \): induced feature map,
• \( f : \mathcal{X} \to \mathbb{R}^k \): compatibility function,
  \( f(x) = (w, \phi(x))_\mathcal{H} \)

Single-label Structured Prediction: CRFs, SSVMs, ...

Single-label prediction: learn a function \( f : \mathcal{X} \to \mathcal{Y} \) from training data \( (x^1, y^1), \ldots, (x^n, y^n) \), with \( x \in \mathcal{X}, y \in \mathcal{Y} \).

\( \Delta : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+ \): task-dependent loss

\( \Delta(y, \hat{y}) \) cost of predicting \( \hat{y} \) if \( y \) is correct

SSVM training problem (slack-rescaled) [1]:
\[
(w^*, \xi^*) = \arg \min_{w \in \mathbb{R}^d, \xi \in \mathbb{R}^n} \frac{1}{2} \|w\|_2^2 + C \sum_{i=1}^n \xi_i^i,
\]
subject to, for \( i = 1, \ldots, n \), all \( y \in \mathcal{Y} \),
\( \xi_i^i + \Delta(y^i, y^i) [1 + f(x^i, y^i) - f(x^i, y^i)] \geq 0 \)

Prediction function:
\( g(x) = \text{argmax}_{y \in \mathcal{Y}} f(x, y) \)

Multi-label Structured Prediction

Multi-label prediction: learn a function \( f : \mathcal{X} \to \mathcal{P}(\mathcal{Y}) \), from training set \( (x^1, Y^1), \ldots, (x^n, Y^n) \) with \( Y^i \subseteq \mathcal{Y} \).

\( \mathcal{M} : \mathcal{P}(\mathcal{Y}) \times \mathcal{Y} \to \mathbb{R}^+\): task-dependent per-label misclassification cost

\( \mathcal{L} : \mathcal{P}(\mathcal{Y}) \to \mathbb{R}_+ \): induced feature map

First Idea: SSVM with output space \( \mathcal{P}(\mathcal{Y}) \) – P-SSVM

• output set has \( 2^{|\mathcal{Y}|} \) elements → we might need \( O(3^{|\mathcal{Y}|}) \) samples [3]
• working set training needs identifying most violated \( (x^i, Y^i) \) with \( Y \subseteq \mathcal{Y} \), where \( Y \) could be as large as all of \( \mathcal{Y} \).

Main Insight:

P-SSVM training does not scale to structured spaces with large \( |\mathcal{Y}| \).

Proposed Method: MLSP

Define indicator vectors, \( v^i \in \{\pm 1\}^{|\mathcal{Y}|} \), \( v^i_j := \begin{cases} +1 & \text{if } y^i_j \in Y^i \text{, otherwise.} \end{cases} \)

MLSP training problem (slack-rescaled):
\[
(w^*, \xi^*) = \arg \min_{w \in \mathbb{R}^d, \xi \in \mathbb{R}^n} \frac{1}{2} \|w\|_2^2 + C \sum_{i=1}^n \xi_i^i
\]
subject to, for \( i = 1, \ldots, n \), all \( y \in \mathcal{Y} \),
\( \xi_i^i + \Delta(y^i, y^i) [1 - v^i(x^i, y^i)] \geq 0 \)

Prediction function:
\( g(x) = \{ y \in \mathbb{Y} : f(x, y) > 0 \} \)

Generalization Properties

Let \( g_w(x) := \{ y \in \mathcal{Y} : f_w(x, y) > 0 \} \) for \( f_w(x, y) = \langle w, \phi(x) \rangle \).

Assume \( |\mathcal{Y}| < r \) and \( \|\phi(x)\|_2 \leq s \) for all \( x \in \mathcal{X}, y \in \mathcal{Y} \), and \( \lambda(Y, \bar{y}) \leq \Lambda \) for all \( (Y, \bar{y}) \in \mathcal{Y} \).

For any distribution \( P_q \) over weight vectors, denote by \( L(Q_w, P) \) the expected \( \Delta \)-risk for \( P \)-distributed data,
\( L(Q_w, P) = \mathbb{E}_{w \sim Q} P(\{y \in \mathcal{Y} : f_w(x, y) > 0\}) = \mathbb{E}_{w \sim Q} \phi(x)) \Delta(Y, \bar{y}) \),
where \( \Delta(Y, \bar{y}) := \max_{y \in Y, \bar{y}} \lambda(Y, \bar{y}) \) is the max-loss over sets.

Theorem 1 (slack-rescaled version): With probability at least \( 1 - \sigma \) over the sample \( S \) of size \( n \), the inequality holds simultaneously for all weight vectors \( w \).
\[
L(Q_w, P) \leq \sum_{i=1}^n \ell(x^i, y^i, \ell) + \frac{\|w\|_2^2}{n} - \frac{\|\xi\|_2^2}{n} \ln \left( \frac{n}{w^2} \right) \left( \frac{\|w\|_2^2}{n} \ln \frac{n}{w^2} + \ln \frac{2}{\min} \right) \frac{1}{2} n - 1
\]
\( \ell(x^i, y^i, \ell) := \max \lambda(Y, \bar{y}) \left| v^i(x^i, y^i) < 1 \right| \)
(=margin violation).

Experimental

Baselines: SSVM with multi-label prediction rule (\( * \)), P-SSVM, JKSE [5]

Evaluation measures:

<table>
<thead>
<tr>
<th>Measure</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \Delta_{\max}(Y, \bar{y}) = \max_{y \in Y, \bar{y}} \lambda(Y, \bar{y}) )</td>
<td>max-loss (lower is better)</td>
</tr>
<tr>
<td>( \Delta_{\sum}(Y, \bar{y}) = \sum_{y \in Y, \bar{y}} \lambda(Y, \bar{y}) )</td>
<td>sum-loss (lower is better)</td>
</tr>
<tr>
<td>mAUC: area under per-sample precision-recall curves (higher is better)</td>
<td></td>
</tr>
<tr>
<td>prec/rec@F1: precision, recall, F1-score (higher is better)</td>
<td></td>
</tr>
</tbody>
</table>

Hierarchical Classification:

Object Detection:

<table>
<thead>
<tr>
<th>Measure</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>(</td>
<td>\mathcal{Y}</td>
</tr>
<tr>
<td>( \Delta_{\max}(Y, \bar{y}) )</td>
<td>prec/rec@F1</td>
</tr>
<tr>
<td>( \Delta_{\max}(Y, \bar{y}) )</td>
<td>prec/rec@F1</td>
</tr>
</tbody>
</table>

Main Insight:

MLSPs achieve similar results to P-SSVMs, but scale to large \( |\mathcal{Y}| \).

Summary

| Maximum margin framework for predicting multiple structured labels, |
| Efficient through re-use of techniques/insights from single-label SSVM. |