Loss-based Learning with Latent Variables

M. Pawan Kumar
École Centrale Paris
École des Ponts ParisTech
INRIA Saclay, Île-de-France

Joint work with
Ben Packer, Daphne Koller, Kevin Miller and Danny Goodman
Image Classification

Images $x_i$, Boxes $h_i$, Labels $y_i$, Image $x$

- Bison
- Deer
- Elephant
- Giraffe
- Llama
- Rhino

$y = \text{“Deer”}$
Image Classification

Feature $\Psi(x,y,h)$ (e.g. HOG)

Score $f : \Psi(x,y,h) \rightarrow (-\infty, +\infty)$

$$f(\Psi(x,y_1,h))$$

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>0.00</td>
<td>0.75</td>
<td>0.00</td>
</tr>
<tr>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>
Image Classification

Feature $\Psi(x, y, h)$ (e.g. HOG)

Score $f : \Psi(x, y, h) \rightarrow (-\infty, +\infty)$

Prediction $y(f)$

Learn $f$

\[
\begin{array}{ccc}
0.00 & 0.00 & 0.00 \\
0.00 & 0.75 & 0.00 \\
0.00 & 0.00 & 0.00 \\
0.00 & 0.00 & 0.00 \\
\end{array}
\]

\[
\begin{array}{ccc}
0.00 & 0.23 & 0.00 \\
0.00 & 0.00 & 0.01 \\
0.01 & 0.00 & 0.00 \\
\end{array}
\]
Loss-based Learning

User defined loss function $\Delta(y, y(f))$

$$f^* = \arg\min_f \sum_i \Delta(y_i, y_i(f))$$

Minimize loss between predicted and ground-truth output

No restriction on the loss function

General framework (object detection, segmentation, …)
Outline

• Latent SVM

• Max-Margin Min-Entropy Models

• Dissimilarity Coefficient Learning

Andrews et al., NIPS 2001; Smola et al., AISTATS 2005; Felzenszwalb et al., CVPR 2008; Yu and Joachims, ICML 2009
Image Classification

Images $x_i$  Boxes $h_i$  Labels $y_i$  Image $x$

Bison

Deer

Elephant

Giraffe

Llama

Rhino

$y = \text{“Deer”}$
Latent SVM

Scoring function

Parameters \( w^T \Phi(x,y,h) \)

Features

Prediction

\[
y(w), h(w) = \arg\max_{y,h} w^T \Phi(x,y,h)
\]
Learning Latent SVM

Training data \{(x_i, y_i), i = 1,2,...,n\}

\[ w^* = \arg \min_w \sum_i \Delta(y_i, y_i(w)) \]

Highly non-convex in \( w \)

Cannot regularize \( w \) to prevent overfitting
Learning Latent SVM

Training data \{(x_i, y_i), i = 1,2,...,n\}

\[
\begin{align*}
    w^T\Psi(x,y_i(w),h_i(w)) + \Delta(y_i,y_i(w)) & \leq w^T\Psi(x,y_i(w),h_i(w)) \\
    w^T\Psi(x,y_i(w),h_i(w)) + \Delta(y_i,y_i(w)) & \leq \max_{h_i} w^T\Psi(x,y,h_i) \\
    \max_y \max_{h_i} \{w^T\Psi(x,y,h) + \Delta(y,y_i)\} & \leq \max_{h_i} w^T\Psi(x,y_i,h_i)
\end{align*}
\]
Learning Latent SVM

Training data \( \{(x_i, y_i), i = 1,2,\ldots,n\} \)

\[
\min_w \|w\|^2 + C \sum_i \xi_i \\
w^T\Psi(x_i,y,h) + \Delta(y_i,y) - \max_{h_i} w^T\Psi(x,y_i,h_i) \leq \xi_i
\]

Difference-of-convex program in \( w \)

Local minimum or saddle point solution (CCCP)

Self-Paced Learning, NIPS 2010
Recap

Scoring function

\[ w^T \Psi(x, y, h) \]

Prediction

\[ y(w), h(w) = \text{argmax}_{y, h} w^T \Psi(x, y, h) \]

Learning

\[ \min_w ||w||^2 + C \sum_i \xi_i \]

\[ w^T \Psi(x_i, y, h) + \Delta(y_i, y) - \max_{h_i} w^T \Psi(x, y_i, h_i) \leq \xi_i \]
Image Classification

Images $x_i$, Boxes $h_i$, Labels $y_i$, Image $x$

Bison
Deer
Elephant
Giraffe
Llama
Rhino

$y = \text{“Deer”}$
Image Classification

Score $w^T \Psi(x, y, h) \rightarrow (-\infty, +\infty)$
Image Classification

Score $w^T \Psi(x, y, h) \rightarrow (-\infty, +\infty)$

Only maximum score used

No other useful cue?

Uncertainty in $h$

<table>
<thead>
<tr>
<th>$w^T \Psi(x, y_1, h)$</th>
<th>$w^T \Psi(x, y_2, h)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00 0.00 0.25</td>
<td>0.00 0.24 0.00</td>
</tr>
<tr>
<td>0.00 0.25 0.00</td>
<td>0.00 0.00 0.00</td>
</tr>
<tr>
<td>0.25 0.00 0.00</td>
<td>0.01 0.00 0.00</td>
</tr>
</tbody>
</table>
Outline

- Latent SVM
- Max-Margin Min-Entropy (M3E) Models
- Dissimilarity Coefficient Learning

Miller, Kumar, Packer, Goodman and Koller, AISTATS 2012
M3E

Scoring function

\[ P_w(y,h|x) = \frac{\exp(w^T \Psi(x,y,h))}{Z(x)} \]

Prediction

\[ y(w) = \text{argmin}_y H_\alpha(P_w(h|y,x)) - \log P_w(y|x) \]

Rényi Entropy

Marginalized Probability

Rényi Entropy of Generalized Distribution

\[ G_\alpha(y;x,w) \]
Rényi Entropy

\[ G_\alpha(y;x,w) = \frac{1}{1-\alpha} \log \left( \frac{\sum_h P_w(y,h|x)^\alpha}{\sum_h P_w(y,h|x)} \right) \]

\(\alpha = 1\). Shannon Entropy of Generalized Distribution

\[- \sum_h P_w(y,h|x) \log(P_w(y,h|x)) \]

\[ \frac{\sum_h P_w(y,h|x)}{\sum_h P_w(y,h|x)} \]
Rényi Entropy

\[ G_\alpha(y; x, w) = \frac{1}{1-\alpha} \log \left( \frac{\sum_h P_w(y, h|x)^\alpha}{\sum_h P_w(y, h|x)} \right) \]

\( \alpha = \text{Infinity}. \) Minimum Entropy of Generalized Distribution

\[- \max_h \log(P_w(y, h|x)) \]
Rényi Entropy

\[ G_\alpha(y;x,w) = \frac{1}{1-\alpha} \log \left( \frac{\sum_h P_w(y,h|x)^\alpha}{\sum_h P_w(y,h|x)} \right) \]

\( \alpha = \text{Infinity. Minimum Entropy of Generalized Distribution} \)

\[- \max_h w^\top \Psi(x,y,h) \]

Same prediction as latent SVM
Learning M3E

Training data \{ (x_i, y_i), \ i = 1, 2, \ldots , n \} 

$$w^* = \ \text{argmin}_w \ \sum_i \Delta (y_i, y_i(w))$$

Highly non-convex in \( w \)

Cannot regularize \( w \) to prevent overfitting
Learning M3E

Training data \{(x_i, y_i), i = 1,2,...,n\}

\[ G_\alpha(y_i(w);x_i,w) + \Delta(y_i,y_i(w)) - G_\alpha(y_i(w);x_i,w) \]

\[ \leq G_\alpha(y_i(x_i,w) + \Delta(y_i,y_i(w)) - G_\alpha(y_i(w);x_i,w) \]

\[ \leq G_\alpha(y_i(x_i,w) + \max_y\{\Delta(y_i,y) - G_\alpha(y;x_i,w)\} \]
Learning M3E

Training data \{ (x_i, y_i), i = 1, 2, ..., n \}

\[
\min_w \|w\|^2 + C \sum_i \xi_i \\
G_\alpha(y_i; x_i, w) + \Delta(y_i, y) - G_\alpha(y; x_i, w) \leq \xi_i
\]

When \( \alpha \) tends to infinity, M3E = Latent SVM

Other values can give better results
Image Classification

Mammals Dataset

271 images, 6 classes

90/10 train/test split

5 folds

0/1 loss
Image Classification

Giraffe

HOG-Based Model. Dalal and Triggs, 2005
Motif Finding

UniProbe Dataset

~ 40,000 sequences

Binding vs. Not-Binding

50/50 train/test split

5 Proteins, 5 folds
Motif + Markov Background Model. Yu and Joachims, 2009
Recap

Scoring function

\[ P_w(y,h|x) = \exp(w^T\Psi(x,y,h))/Z(x) \]

Prediction

\[ y(w) = \text{argmin}_y G_\alpha(y;x,w) \]

Learning

\[ \min_w ||w||^2 + C \sum_i \xi_i \]

\[ G_\alpha(y_i;x_i,w) + \Delta(y_i,y) - G_\alpha(y;x_i,w) \leq \xi_i \]
Outline

• Latent SVM

• Max-Margin Min-Entropy Models

• Dissimilarity Coefficient Learning

Kumar, Packer and Koller, ICML 2012
Object Detection

Images $x_i$  Boxes $h_i$  Labels $y_i$  Image $x$

- Bison
- Deer
- Elephant
- Giraffe
- Llama
- Rhino

$y = \text{“Deer”}$
Minimizing General Loss

\[
\min_w \sum_i \Delta(y_i, h_i, y_i(w), h_i(w)) \quad \text{Supervised Samples}
\]

\[
+ \sum_i \Delta'(y_i, y_i(w), h_i(w)) \quad \text{Weakly Supervised Samples}
\]

Unknown latent variable values
Minimizing General Loss

$$\min_w \sum_i \sum_{h_i} \Delta(y_i, h_i, y_i(w), h_i(w)) P_w(h_i|x_i, y_i)$$

A single distribution to achieve two objectives
Problem

Model Uncertainty in Latent Variables

Model Accuracy of Latent Variable Predictions
Solution

Use two different distributions for the two different tasks

Model Uncertainty in Latent Variables

Model Accuracy of Latent Variable Predictions
Solution

Use two different distributions for the two different tasks

\[ P_{\theta}(h_i|y_i, x_i) \]
Solution

Use two different distributions for the two different tasks

\[ P_{\theta}(h_i|y_i,x_i) \]

\[ P_{w}(y_i,h_i|x_i) \]
The Ideal Case

No latent variable uncertainty, correct prediction

\[ P_\theta(h_i|y_i, x_i) \]

\[ P_w(y_i, h_\theta|\mathbf{x}_i) \]
In Practice

Restrictions in the representation power of models

$P_{\theta}(h_i|y_i,x_i)$

$P_w(y_i,h_i|x_i)$
Our Framework

Minimize the dissimilarity between the two distributions

$$P_\theta(h_i|y_i,x_i)$$

User-defined dissimilarity measure

$$P_w(y_i,h_i|x_i)$$

$$(y_i(w),h_i(w))$$

$$(y_i,h_i)$$
Our Framework

Minimize Rao’s Dissimilarity Coefficient

\[ P_\theta(h_i|y_i, x_i) \]

\[ \sum_h \Delta(y_i, h, y_i(w), h_i(w))P_\theta(h|y_i, x_i) \]

\[ P_w(y_i, h_i|x_i) \]
Our Framework

Minimize Rao’s Dissimilarity Coefficient

\[ P_{\theta}(h_i|y_i, x_i) \]

\[ H_i(w, \theta) \]

\[ -\beta \sum_{h, h'} \Delta(y_i, h, y_i, h')P_{\theta}(h|y_i, x_i)P_{\theta}(h'|y_i, x_i) \]

\[ \mathcal{P}_w(y_i, h_i|x_i) \]
Our Framework

Minimize Rao’s Dissimilarity Coefficient

\[ P_\theta(h_i|y_i, x_i) \]

\[ H_i(w, \theta) - \beta H_i(\theta, \theta) \]

\[ - (1-\beta) \Delta(y_i(w), h_i(w), y_i(w), h_i(w)) \]

\[ P_w(y_i, h_i|x_i) \]
Our Framework

Minimize Rao’s Distinctilarity Coefficient

\[ P_\theta(h_i|y_i,x_i) \]

\[ \min_{w,\theta} \sum_i H_i(w,\theta) - \beta H_i(\theta,\theta) \]

\[ P_w(y_i,h_i|x_i) \]

\[ (y_i(w),h_i(w)) \]

\[ (y_i,h_i) \]
Optimization

\[
\min_{\mathbf{w}, \theta} \sum_i H_i(\mathbf{w}, \theta) - \beta H_i(\theta, \theta)
\]

Initialize the parameters to \( \mathbf{w}_0 \) and \( \theta_0 \)

Repeat until convergence

   Fix \( \mathbf{w} \) and optimize \( \theta \)

   Fix \( \theta \) and optimize \( \mathbf{w} \)

End
Optimization of $\theta$

$$\min_{\theta} \sum_{i} \sum_{h} \Delta(y_i, h, y_i(w), h_i(w)) P_{\theta}(h|y_i, x_i) - \beta H_i(\theta, \theta)$$

Case I: $y_i(w) = y_i$
Optimization of $\theta$

$$\min_{\theta} \sum_i \sum_h \Delta(y_i, h, y_i(w), h_i(w)) P_{\theta}(h|y_i, x_i) - \beta H_i(\theta, \theta)$$

Case I: $y_i(w) = y_i$
Optimization of $\theta$

$$\min_\theta \sum_i \sum_h \Delta(y_i, h, y_i(w), h_i(w)) P_\theta(h|y_i, x_i) - \beta H_i(\theta, \theta)$$

Case II: $y_i(w) \neq y_i$
Optimization of $\theta$

$$\min_{\theta} \sum_i \sum_h \Delta(y_i, h, y_i(w), h_i(w))P_\theta(h|y_i, x_i) - \beta H_i(\theta, \theta)$$

Stochastic subgradient descent

Case II: $y_i(w) \neq y_i$
Optimization of $w$

$$\min_w \sum_i \sum_h \Delta(y_i, h, y_i(w), h_i(w))P_\theta(h|y_i, x_i)$$

Expected loss, models uncertainty

Form of optimization similar to Latent SVM

Concave-Convex Procedure (CCCP)

$\Delta$ independent of $h$, implies latent SVM
Object Detection

Train Input $x_i$

Output $y_i$

- Bison
- Deer
- Elephant
- Giraffe
- Llama
- Rhino

Input $x$

Output $y = \text{“Deer”}$

Latent Variable $h$

Mammals Dataset

60/40 Train/Test Split

5 Folds
Results – 0/1 Loss

Statistically Significant

Average Test Loss

Fold 1     Fold 2     Fold 3     Fold 4     Fold 5

LSVM       Our
Results – Overlap Loss

Average Test Loss

Fold 1  Fold 2  Fold 3  Fold 4  Fold 5

LSVM  Our

0.6  0.5  0.4  0.3  0.2  0.1  0
# Action Detection

<table>
<thead>
<tr>
<th>Train Input $x_i$</th>
<th>Output $y_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Jumping</td>
</tr>
<tr>
<td></td>
<td>Phoning</td>
</tr>
<tr>
<td></td>
<td>Playing Instrument</td>
</tr>
<tr>
<td></td>
<td>Reading</td>
</tr>
<tr>
<td></td>
<td>Riding Bike</td>
</tr>
<tr>
<td></td>
<td>Riding Horse</td>
</tr>
<tr>
<td></td>
<td>Running</td>
</tr>
<tr>
<td></td>
<td>Taking Photo</td>
</tr>
<tr>
<td></td>
<td>Using Computer</td>
</tr>
<tr>
<td></td>
<td>Walking</td>
</tr>
</tbody>
</table>

Output $y_i = \text{“Using Computer”}$

Latent Variable $h$

PASCAL VOC 2011

60/40 Train/Test Split

5 Folds
Results – 0/1 Loss

Statistically Significant

Average Test Loss

- LSVM
- Our

Fold 1 | Fold 2 | Fold 3 | Fold 4 | Fold 5
Results – Overlap Loss

Statistically Significant

Average Test Loss

Fold 1  Fold 2  Fold 3  Fold 4  Fold 5

0.63  0.64  0.65  0.66  0.67  0.68  0.69  0.7  0.71  0.72  0.73  0.74

LSVM  Our
Conclusions

- Latent SVM ignores latent variable uncertainty
- M3E for latent variable independent loss
- DISC for latent variable dependent loss
- Strict generalizations of Latent SVM
- Code available online
Questions?

http://cvc.centrale-ponts.fr/personnel/pawan