Entropy and Margin Maximization for Structured Output Learning

Patrick Pletscher, Cheng Soon Ong, Joachim M. Buhmann
Machine Learning and Pattern Recognition Group, ETH Zurich, Switzerland

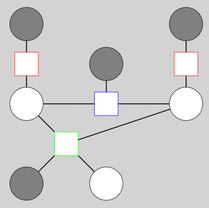
ETH Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Contributions

- Unifying view of Conditional Random Fields and Structured SVM.
- Generalized loss $\ell_\beta(\mathbf{w}, \mathbf{x}, \mathbf{y})$.
- Improved accuracy of the novel loss.

Graphical model and factor templates

- Discrete output variables $\mathbf{y} \in \mathcal{Y}$ ○ and input variables $\mathbf{x} \in \mathcal{X}$ ●.
- Factor graph structure and parameterization of the factors assumed to be given.
- Several factors can share the same parameter.



Regularized empirical risk minimization

- For given data set $\mathcal{D} = \{(\mathbf{x}^{(n)}, \mathbf{y}^{(n)})\}_{n=1}^N$ and loss $\ell(\mathbf{w}, \mathbf{x}^{(n)}, \mathbf{y}^{(n)})$

$$\mathcal{L}_\ell(\mathbf{w}, \mathcal{D}, C) = \sum_{n=1}^{N} \ell(\mathbf{w}, \mathbf{x}^{(n)}, \mathbf{y}^{(n)}) + \frac{C}{2}\|\mathbf{w}\|_2^2,$$

- Choose parameter with smallest regularized empirical risk

$$\widehat{\mathbf{w}} = \operatorname*{argmin}_{\mathbf{w}} \mathcal{L}_\ell(\mathbf{w}, \mathcal{D}, C)$$

Conditional Random Field vs. Structured SVM

- Structured SVM $\leftrightarrow$ max-margin loss:

$$\ell_{MM}(\mathbf{w}, \mathbf{x}, \mathbf{y}) = -\langle \mathbf{w}, \phi(\mathbf{x}, \mathbf{y})\rangle + \max_{\mathbf{y}' \in \mathcal{Y}} \left[\langle \mathbf{w}, \phi(\mathbf{x}, \mathbf{y}')\rangle + \Delta(\mathbf{y}', \mathbf{y})\right].$$

- Conditional Random Field $\leftrightarrow$ log-likelihood loss:

$$\ell_{LL}(\mathbf{w}, \mathbf{x}, \mathbf{y}) = -\langle \mathbf{w}, \phi(\mathbf{x}, \mathbf{y})\rangle + \log Z(\mathbf{x}, \mathbf{w}),$$

with the partition sum

$$Z(\mathbf{x}, \mathbf{w}) = \sum_{\mathbf{y}' \in \mathcal{Y}} \exp(\langle \mathbf{w}, \phi(\mathbf{x}, \mathbf{y}')\rangle).$$

- Inference: both generally intractable. Max can be simpler (submodularity).
- Optimization: both convex. Max-margin non-differentiable.

Inverse temperature β

Introduce an inverse temperature β into the CRF

$$P_\beta(\mathbf{y}|\mathbf{x}, \mathbf{w}) = \frac{1}{Z_\beta(\mathbf{x}, \mathbf{w})} \exp\left(\beta \langle \mathbf{w}, \phi(\mathbf{x}, \mathbf{y})\rangle\right).$$

In itself only changes the regularizer parameter C.

Margin term $\Delta(\mathbf{y}', \mathbf{y})$

Define posterior over outputs (large $\rightarrow$ "bad output")

$$P_\beta(\mathbf{y}'|\mathbf{y}) = \frac{1}{Z_\beta(\mathbf{y})} \exp(\beta \Delta(\mathbf{y}', \mathbf{y})).$$

Here $\Delta(\mathbf{y}', \mathbf{y})$ specifies the cost of predicting $\mathbf{y}'$ instead of $\mathbf{y}$:

$$\Delta(\mathbf{y}', \mathbf{y}) = \begin{cases} 0 & \text{if } \mathbf{y}' = \mathbf{y} \\ \geq 0 & \text{otherwise.} \end{cases}$$

Example of a joint featuremap $\phi(\mathbf{x}, \mathbf{y})$: Ising model

- $y_i \in \{0, 1\}$ and $x_i \in \{0, 1\}$.
- Energy given by

$$E(\mathbf{y}, \mathbf{x}, \mathbf{w}) = -\sum_{i \in \mathcal{V}} w^u(|y_i - x_i|) - \sum_{(i,j) \in \mathcal{E}} w^p(|y_i - y_j|)$$

- Parameters $\mathbf{w}^u = [0, a]^\mathsf{T}$ and $\mathbf{w}^p = [0, b]^\mathsf{T}$.
- Introduce ($\delta_c(z)$, 1 if $z = c$, 0 otherwise):

$$\phi(\mathbf{x}, \mathbf{y}) = \begin{bmatrix} \sum_{i \in \mathcal{V}} \delta_1(|y_i - x_i|) \\ \sum_{(i,j) \in \mathcal{E}} \delta_1(|y_i - y_j|) \end{bmatrix}, \qquad \mathbf{w} = \begin{bmatrix} a \\ b \end{bmatrix}.$$

- Energy of a configuration: $E(\mathbf{y}, \mathbf{x}, \mathbf{w}) = -\langle \mathbf{w}, \phi(\mathbf{x}, \mathbf{y})\rangle$

Softmax loss ℓ_β

- Combination

$$P_\beta(\mathbf{y}'|\mathbf{y}, \mathbf{x}, \mathbf{w}) \propto P_\beta(\mathbf{y}'|\mathbf{x}, \mathbf{w}) P_\beta(\mathbf{y}'|\mathbf{y}).$$

- Ensuring normalization

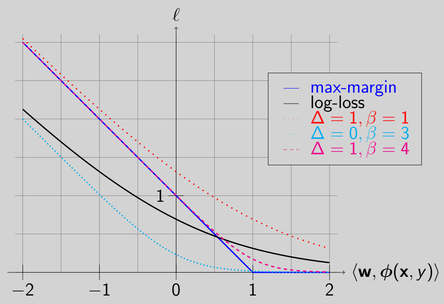
$$P_\beta(\mathbf{y}'|\mathbf{y}, \mathbf{x}, \mathbf{w}) = \frac{1}{Z_\beta(\mathbf{y}, \mathbf{x}, \mathbf{w})} \exp\Big(\beta \langle \mathbf{w}, \phi(\mathbf{x}, \mathbf{y}')\rangle + \beta \Delta(\mathbf{y}', \mathbf{y})\Big).$$

- Rescaling by $1/\beta$ and taking the logarithm:

$$\ell_\beta(\mathbf{w}, \mathbf{x}, \mathbf{y}) = -\langle \mathbf{w}, \phi(\mathbf{x}, \mathbf{y})\rangle + \frac{1}{\beta} \log \sum_{\mathbf{y}' \in \mathcal{Y}} \exp\Big(\beta \langle \mathbf{w}, \phi(\mathbf{x}, \mathbf{y}')\rangle + \beta \Delta(\mathbf{y}', \mathbf{y})\Big).$$

Connections

- SSVM: $\beta \rightarrow \infty$.
- CRF: small β and small C.
- Also applies to hidden variables.
- $P_\beta(\mathbf{y}'|\mathbf{y}, \mathbf{x}, \mathbf{w})$: loss dependent.
- Inaccurate probability estimates.
- Possibly better classification accuracy.



Special case: binary classification

Dual view

The dual minimization problem is given by

$$\min_{\mathbf{u}} \quad \frac{1}{2C}\mathbf{u}^\mathsf{T}\mathbf{A}\mathbf{u} - \mathbf{b}^\mathsf{T}\mathbf{u} + \frac{1}{\beta}\sum_{n=1}^{N}\sum_{\mathbf{y} \in \mathcal{Y}} u_{n,\mathbf{y}} \log u_{n,\mathbf{y}}$$

$$\text{s.t.} \quad u_{n,\mathbf{y}} \geq 0 \quad \text{and} \quad \sum_{\mathbf{y} \in \mathcal{Y}} u_{n,\mathbf{y}} = 1 \ \forall \mathbf{y}, n.$$

$\mathbf{A}$ is given by $A_{(n_1,\mathbf{y}),(n_2,\mathbf{y}')} = \langle \mathbf{g}_{n_1,\mathbf{y}}, \mathbf{g}_{n_2,\mathbf{y}'}\rangle$. With

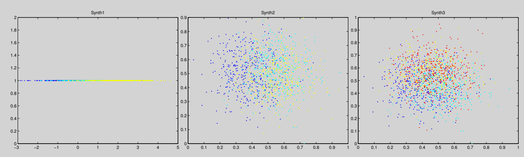
$$\mathbf{g}_{n,\mathbf{y}} = \phi(\mathbf{x}^{(n)}, \mathbf{y}^{(n)}) - \phi(\mathbf{x}^{(n)}, \mathbf{y}) \quad \text{and} \quad b_{n,\mathbf{y}} = \Delta(\mathbf{y}, \mathbf{y}^{(n)}).$$

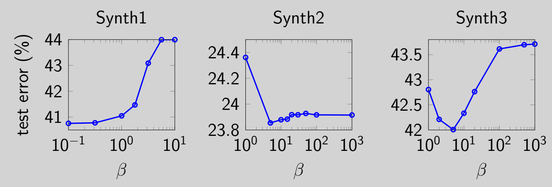
$N \cdot |\mathcal{Y}|$ dual variables are required. Primal and dual variables related by

$$\mathbf{w} = \frac{1}{C}\sum_{n=1}^{N}\sum_{\mathbf{y} \in \mathcal{Y}} u_{n,\mathbf{y}} \mathbf{g}_{n,\mathbf{y}}.$$

Synthetic multiclass experiment

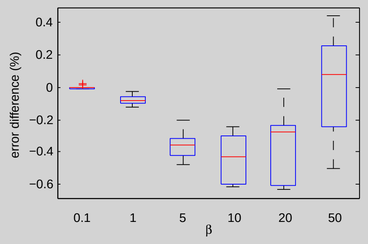
Three datasets to emphasize differences between the losses.

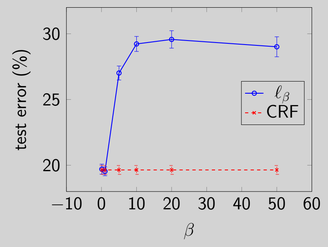




loss		test error (%)		
β	$\Delta(y', y)$	Synth1	Synth2	Synth3
1	no	$\mathbf{41.0 \pm 0.4}$	25.8 ± 0.2	43.9 ± 0.2
5	yes	44.0 ± 0.1	$\mathbf{24.2 \pm 0.1}$	$\mathbf{42.0 \pm 0.2}$
10^6	yes	44.0 ± 0.1	$\mathbf{24.2 \pm 0.1}$	43.7 ± 0.7

OCR experiment





References

- J. Lafferty, A. McCallum, and F. Pereira (2001). "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data". In: *ICML*
- I. Tsochantaridis et al. (2004). "Support vector machine learning for interdependent and structured output spaces". In: *ICML*, p. 104
- B. Taskar, C. Guestrin, and D. Koller (2003). "Max-Margin Markov Networks". In: *NIPS*
- M. Collins et al. (2008). "Exponentiated Gradient Algorithms for Conditional Random Fields and Max-Margin Markov Networks". In: *J. Mach. Learn. Res.* 9, pp. 1775–1822
- T. Zhang (2005). "Class-size Independent Generalization Analysis of Some Discriminative Multi-Category Classification". In: *NIPS*. Cambridge, MA
- K. Gimpel and N. Smith (2010). "Softmax-Margin CRFs: Training Log-Linear Models with Cost Functions". In: *HLT*, pp. 733–736