Exercise 1: Conditional Random Fields vs. Structured SVMs

Similar to probabilistic classifier chains, conditional random fields try to model the conditional distribution $\mathbb{P}(\mathbf{y} \mid \mathbf{x})$ by means of

$$\pi(\mathbf{x}, \mathbf{y}) = \frac{\exp(s(\mathbf{x}, \mathbf{y}))}{\sum_{\mathbf{y}' \in \mathcal{Y}^m} \exp(s(\mathbf{x}, \mathbf{y}'))},$$

where $x \in \mathcal{X}$ and $\mathbf{y} \in \mathcal{Y}$ with $\mathcal{Y}$ being a finite set (e.g., multi-label classification), and $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ being a scoring function. Training of a conditional random field is based on (regularized) empirical risk minimization using the negative log-loss:

$$\ell_{log}(\mathbf{x}, \mathbf{y}, s) = \log \left(\sum_{\mathbf{y}' \in \mathcal{Y}^m} \exp(s(\mathbf{x}, \mathbf{y}')) \right) - s(\mathbf{x}, \mathbf{y}).$$

Predictions are then made by means of

$$h(\mathbf{x}) = \arg\max\nolimits_{\mathbf{y} \in \mathcal{Y}^m} s(\mathbf{x}, \mathbf{y}). \tag{1}$$

Structured Support Vector Machines (Structured SVMs) are also using scoring functions for the prediction, but use the structured hinge loss for the (regularized) empirical risk minimization approach:

$$\ell_{shinge}(\mathbf{x}, \mathbf{y}, s) = \max_{\mathbf{y}' \in \mathcal{Y}^m} \left(\ell(\mathbf{y}, \mathbf{y}') + s(\mathbf{x}, \mathbf{y}') - s(\mathbf{x}, \mathbf{y}) \right),$$

where $\ell : \mathcal{Y}^m \times \mathcal{Y}^m \to \mathbb{R}$ is some target loss function (e.g., Hamming loss or subset 0/1 loss).

Show that if we use scoring functions s of the form

$$s(\mathbf{x}, \mathbf{y}) = \sum_{j=1}^{m} s_j(\mathbf{x}, y_j),$$

where $s_j : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ are scoring functions for the j-th target, then

(a) conditional random fields are very well suited to model the case, where the distributions of the targets $y_1, \ldots, y_m$ are conditionally independent. In other words, show that $\mathbb{P}(\mathbf{y} \mid \mathbf{x}) \approx \prod_{j=1}^{m} \mathbb{P}(y_j \mid \mathbf{x})$. *Hint:* Use the multinomial theorem:

$$(z_1 + z_2 + \ldots + z_g)^m = \sum_{k_1 + k_2 + \ldots k_g = m} \binom{m}{k_1, k_2, \ldots, k_g} \prod_{t=1}^{g} z_t^{k_t}.$$

(b) The structured hinge loss corresponds to the multiclass hinge loss for the targets if we use the (non-averaged) Hamming loss for $\ell(\mathbf{y}, \mathbf{y}') = \sum_{j=1}^{m} \mathbb{1}_{[y_j \neq y_j']}$, i.e.,

$$\ell_{shinge}(\mathbf{x}, \mathbf{y}, s) = \sum_{j=1}^{m} \max_{y_j' \in \mathcal{Y}} \left(\mathbb{1}_{[y_j \neq y_j']} + s_j(\mathbf{x}, y_j') - s_j(\mathbf{x}, y_j) \right).$$