Solution 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

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

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:

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

Solution:

The idea of conditional random fields is to model the joint conditional distribution $\mathbb{P}(\mathbf{y} \mid \mathbf{x})$ by means of

$\pi(\mathbf{x}, \mathbf{y})$. Thus, it should hold $\mathbb{P}(\mathbf{y} \mid \mathbf{x}) \approx \pi(\mathbf{x}, \mathbf{y})$ and with this,

$$\begin{aligned}
\mathbb{P}(\mathbf{y} \mid \mathbf{x}) &\approx \pi(\mathbf{x}, \mathbf{y}) \\
&= \frac{\exp(s(\mathbf{x}, \mathbf{y}))}{\sum_{\mathbf{y}' \in \mathcal{Y}^m} \exp(s(\mathbf{x}, \mathbf{y}'))} \\
&= \frac{\exp\left(\sum_{j=1}^m s_j(\mathbf{x}, y_j)\right)}{\sum_{\mathbf{y}' \in \mathcal{Y}^m} \exp\left(\sum_{j=1}^m s_j(\mathbf{x}, y'_j)\right)} \\
&= \frac{\prod_{j=1}^m \exp\left(s_j(\mathbf{x}, y_j)\right)}{\sum_{\mathbf{y}' \in \mathcal{Y}^m} \prod_{j=1}^m \exp\left(s_j(\mathbf{x}, y'_j)\right)} \\
&= \frac{\prod_{j=1}^m \exp\left(s_j(\mathbf{x}, y_j)\right)}{\prod_{j=1}^m \sum_{y'_j \in \mathcal{Y}} \exp\left(s_j(\mathbf{x}, y'_j)\right)} \\
&= \prod_{j=1}^m \underbrace{\frac{\exp\left(s_j(\mathbf{x}, y_j)\right)}{\sum_{y'_j \in \mathcal{Y}} \exp\left(s_j(\mathbf{x}, y'_j)\right)}}_{=:\pi_j(\mathbf{x}, y_j)} .
\end{aligned}$$

Note that we used $\sum_{\mathbf{y} \in \mathcal{Y}^m} \prod_{j=1}^m \exp\left(s_j(\mathbf{x}, y_j)\right) = \prod_{j=1}^m \sum_{y_j \in \mathcal{Y}} \exp\left(s_j(\mathbf{x}, y_j)\right)$. We prove this as follows. For brevity let's assume $|\mathcal{Y}| = g$ and define $S_j = s_j(\mathbf{x}, y_j)$. The left hand side can be written as

$$\sum_{\mathbf{y} \in \mathcal{Y}^m} \prod_{j=1}^m \exp\left(s_j(\mathbf{x}, y_j)\right) = \sum_{k_1 + k_2 + \ldots + k_g = m} \binom{m}{k_1, k_2, \ldots, k_g} \prod_{t=1}^g S_t^{k_t}.$$

So we enumerate all the possible $\mathbf{y}$. By using the binomial theorem, this boils down to

$$\begin{aligned}
\sum_{k_1 + k_2 + \ldots + k_g = m} \binom{m}{k_1, k_2, \ldots, k_g} \prod_{t=1}^g S_t^{k_t} &= (S_1 + S_2 + \ldots + S_g)^m \\
&= \prod_{j=1}^m \left(\sum_{t=1}^g S_t \right) \\
&= \prod_{j=1}^m \left(\sum_{y_j \in \mathcal{Y}} \exp(s_j(\mathbf{x}, y_j)) \right) .
\end{aligned}$$

(If you find a problem understanding this part of proof, try a simple example with $|\mathcal{Y}| = 2$ and $m = 3$ and compute the both sides of the equation by hand.)

So, if $\pi_j(\mathbf{x}, y_j)$ is interpreted as a model for the marginal conditional distribution $\mathbb{P}(y_j \mid \mathbf{x})$, we see from above

$$\mathbb{P}(\mathbf{y} \mid \mathbf{x}) \approx \prod_{j=1}^m \mathbb{P}(y_j \mid \mathbf{x}),$$

i.e., the targets are conditionally independent.

(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.,

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

Solution:

This can be seen immediately from the definition:

$$
\begin{aligned}
\ell_{shinge}(\mathbf{x}, \mathbf{y}, s) &= \max_{\mathbf{y}' \in \mathcal{Y}^m} \left(\ell(\mathbf{y}, \mathbf{y}') + s(\mathbf{x}, \mathbf{y}') - s(\mathbf{x}, \mathbf{y}) \right) \\
&= \max_{\mathbf{y}' \in \mathcal{Y}^m} \left(\sum_{j=1}^{m} \mathbb{1}_{[y_j \neq y'_j]} + s(\mathbf{x}, \mathbf{y}') - s(\mathbf{x}, \mathbf{y}) \right) \\
&= \max_{\mathbf{y}' \in \mathcal{Y}^m} \left(\sum_{j=1}^{m} \mathbb{1}_{[y_j \neq y'_j]} + \sum_{j=1}^{m} s_j(\mathbf{x}, y'_j) - \sum_{j=1}^{m} s_j(\mathbf{x}, y_j) \right) \\
&= \max_{\mathbf{y}' \in \mathcal{Y}^m} \left(\sum_{j=1}^{m} \mathbb{1}_{[y_j \neq y'_j]} + s_j(\mathbf{x}, y'_j) - s_j(\mathbf{x}, y_j) \right) \\
&= \sum_{j=1}^{m} \max_{y'_j \in \mathcal{Y}} \left(\mathbb{1}_{[y_j \neq y'_j]} + s_j(\mathbf{x}, y'_j) - s_j(\mathbf{x}, y_j) \right). \qquad \text{(Summands are independent.)}
\end{aligned}
$$