Structured learning with latent variables

October 1, 2017

1 Model

In our model, X is a set of input variables, Y is a set of output variables and H is a set of latent variables. $X \cup Y \cup H = V$ is set of all variables.(x_X, x_Y, x_H) follows a conditional model,

$$p(x_Y, x_H|x_X, w) = \frac{1}{Z(x_X; w)} \exp[w^T \phi(x_X, x_Y, x_H)] \tag{1}$$

Writing it as a log-linear model over complete representation,

$$p(x_Y, x_H|x_X, \theta) = \frac{1}{Z(x_X; \theta)} \prod_{\alpha\beta\gamma} \exp[\theta_{\alpha\beta\gamma}(x_\alpha, x_\beta, x_\gamma)] \tag{2}$$

Here $\alpha \subseteq X, \beta \subseteq Y, \gamma \subseteq H$ and $\alpha\beta\gamma \in F$. α, β and γ form a clique $\alpha\beta\gamma$ in the graph and is associated with a factor $\theta_{\alpha\beta\gamma}$.
We use power sum operator, which is defined as,

$$\sum_{x_i}^{\tau_i} f(x_i) = \left[\sum_{x_i} f(x_i)^{\frac{1}{\tau_i}}\right]^{\tau_i}$$

The power sum reduces to standard sum when τ_i=1 and approaches to $\max_x f(x)$ when $\tau_i \to 0^+$.
Define $\phi_A(\theta)$ for some subset A of variables V as following,

$$\phi_A(\theta) = \log \sum_{x_A}^{\tau_A} \exp\left[\sum_{\alpha\beta\gamma} \theta_{\alpha\beta\gamma}(x_\alpha, x_\beta, x_\gamma)\right]$$

τ_A is set of τ values associated with each variable in A. By setting these variables τ_A to 0 or 1, we can convert the equation above into max or sum problem.

2 Perceptron learning

To classify all data points correctly, we want, for each data point m

$$\sum_{x_H} p(x_Y^m, x_H|x_X^m; \theta) \geq \max_{x_Y} \sum_{x_H} p(x_Y, x_H|x_X^m; \theta)$$

Equivalently,

$$\sum_{x_H} p(x_X^m, x_Y^m, x_H|\theta) \geq \max_{x_Y} \sum_{x_H} p(x_X^m, x_Y, x_H|\theta)$$

Rewriting it using power sum operator,

$$\log \sum_{x_H}^{\tau_H} \exp\left[\sum_{\alpha\beta\gamma} \theta_{\alpha\beta\gamma}(x_\alpha^m, x_\beta^m, x_\gamma)\right] \geq \log \sum_{x_Y, x_H}^{\tau_Y, \tau_H} \exp\left[\sum_{\alpha\beta\gamma} \theta_{\alpha\beta\gamma}(x_\alpha^m, x_\beta, x_\gamma)\right]$$

Here τ_H is set of 1s and τ_y is set of 0s. Power sum operations are applied in order, first on H variables then Y variables, along a fixed order and are not commutative.

The equation above can be written as,

$$\phi_H(\theta|x_X^m, x_Y^m) \geq \phi_{Y\cup H}(\theta|x_X^m)$$

Let's define L as,

$$L(\theta) = \phi_H(\theta|x_X^m, x_Y^m) - \phi_{Y\cup H}(\theta|x_X^m) \geq 0$$

As described in [1], including cost-shifting variables to both $\phi_H(\theta|x_X^m, x_Y^m)$ and $\phi_{Y\cup H}(\theta|x_X^m)$ in the equation above,

$$\begin{aligned} L(\theta, \delta, \zeta) = \log \sum_{x_H}^{\tau_H} \exp\left[\sum_{i\in H} \sum_{\alpha\beta\gamma\in N_i} \delta_i^{\alpha\beta\gamma}(x_i) + \sum_{\alpha\beta\gamma\in F} (\theta_{\alpha\beta\gamma}(x_\alpha^m, x_\beta^m, x_\gamma) - \sum_{i\in\gamma} \delta_i^{\alpha\beta\gamma}(x_i))\right] \\ - \log \sum_{x_Y, x_H}^{\tau_Y, \tau_H} \exp\left[\sum_{i\in Y\cup H} \sum_{\alpha\beta\gamma\in N_i} \zeta_i^{\alpha\beta\gamma}(x_i) + \sum_{\alpha\beta\gamma\in F} (\theta_{\alpha\beta\gamma}(x_\alpha^m, x_\beta, x_\gamma) - \sum_{i\in\beta\gamma} \zeta_i^{\alpha\beta\gamma}(x_i))\right] \end{aligned}$$

where $N_i = \{\alpha\beta\gamma|\alpha\beta\gamma \ni i\}$ is set of cliques incident to i. $\delta_i^{\alpha\beta\gamma}$ and $\zeta_i^{\alpha\beta\gamma}$ are set of cost-shifting variables defined on each variable-clique pair, which can be optimized to provide tighter upper bound later.

Using split-weights according to Theorem4.1 in [1],

$$\begin{aligned} L(\theta, \delta, \zeta, w, \omega) \approx \sum_{i\in H} \log \sum_{x_i}^{w_i} \exp\left[\sum_{\alpha\beta\gamma\in N_i} \delta_i^{\alpha\beta\gamma}(x_i)\right] + \sum_{\alpha\beta\gamma\in F} \log \sum_{x_\gamma}^{w^\gamma} \exp\left[\theta_{\alpha\beta\gamma}(x_\alpha^m, x_\beta^m, x_\gamma) - \sum_{i\in\gamma} \delta_i^{\alpha\beta\gamma}(x_i)\right] \\ - \sum_{i\in Y\cup H} \log \sum_{x_i}^{\omega_i} \exp\left[\sum_{\alpha\beta\gamma\in N_i} \zeta_i^{\alpha\beta\gamma}(x_i)\right] - \sum_{\alpha\beta\gamma\in F} \log \sum_{x_{\beta\gamma}}^{\omega^{\beta\gamma}} \exp\left[\theta_{\alpha\beta\gamma}(x_\alpha^m, x_\beta, x_\gamma) - \sum_{i\in\beta\cup\gamma} \zeta_i^{\alpha\beta\gamma}(x_i)\right] \end{aligned} \tag{3}$$

Here, in first part of the equation, the new weights $w = \{w_i, w_i^\gamma | \forall(i,\gamma), i \in \gamma, w_i^\gamma \geq 0\}$ should satisfy,

$$w_i + \sum_{\alpha\beta\gamma\in N(i)} w_i^\gamma = \tau_i$$

where $\tau_i \in \tau_H$ is 1 as we set it earlier. Similarly for the second part of the equation,

$$\omega_i + \sum_{\alpha\beta\gamma\in N(i)} \omega_i^{\beta\gamma} = \tau_i$$

where $i \in Y \cup H$ and τ_i is either 0 or 1.
All power sum operations in 3 are applied in order, first on H variables and then on Y variables, along a fixed order and are not commutative.

Converting L in 3 to dual representations as described in Theorem-4.2 in [1],

$$L(\theta, b, b', w, \omega) = \max_{b \in L(G_1)} \left\{ \langle \theta, b \rangle + \sum_{i \in H} w_i H(x_i; b_i) + \sum_{\alpha\beta\gamma \in F} \sum_{i \in \gamma} w_i^{\gamma} H(x_i | x_{pa_i^{\alpha\beta\gamma}}; b_{\alpha\beta\gamma}) \right\}$$
$$- \max_{b' \in L(G_2)} \left\{ \langle \theta, b' \rangle + \sum_{i \in Y \cup H} \omega_i H(x_i; b'_i) + \sum_{\alpha\beta\gamma \in F} \sum_{i \in \beta\gamma} \omega_i^{\beta\gamma} H(x_i | x_{pa_i^{\alpha\beta\gamma}}; b'_{\alpha\beta\gamma}) \right\} \quad (4)$$

$pa_i^{\alpha\beta\gamma}$ are variables that are summed out later than i in clique $\alpha\beta\gamma$. We can expand and rearrange conditional entropy terms in equation 4 and rewrite it as,

$$L(\theta, b, b', w, \omega) = \max_{b \in L(G_1)} \left\{ \langle \theta, b \rangle + \sum_{i \in H} w_i H(x_i; b_i) + \sum_{\alpha\beta\gamma \in F} \left\{ w_1^{\gamma} H(x_{\alpha\beta\gamma}; b_{\alpha\beta\gamma}) + \sum_{[i,j] \sqsubseteq \gamma} (w_j^{\gamma} - w_i^{\gamma}) H(x_{pa_i^{\alpha\beta\gamma}}; b_{pa_i^{\alpha\beta\gamma}}) \right\} \right\}$$
$$- \max_{b' \in L(G_2)} \left\{ \langle \theta, b' \rangle + \sum_{i \in Y \cup H} \omega_i H(x_i; b'_i) + \sum_{\alpha\beta\gamma \in F} \left\{ \omega_1^{\beta\gamma} H(x_{\alpha\beta\gamma}; b'_{\alpha\beta\gamma}) + \sum_{[i,j] \sqsubseteq \beta\gamma} (\omega_j^{\beta\gamma} - \omega_i^{\beta\gamma}) H(x_{pa_i^{\alpha\beta\gamma}}; b'_{pa_i^{\alpha\beta\gamma}}) \right\} \right\}$$

where $x_{\alpha\beta\gamma} = \{x_1, x_2, ..., x_i, x_j, ..., x_n\}$ such that i and j are adjacent in summation order.

2.1 Frank-Wolfe optimization

We need to optimize following function (with L2 regularizer),

$$L(\theta, b, b', w, \omega) = \max_{\theta} \left\{ \sum_{m=1}^{M} \max_{b^m \in L(G_1)} \min_{b'^m \in L(G_2)} \left\{ \langle \theta, b^m \rangle + \sum_{i \in H} w_i H(x_i; b_i^m) + \sum_{\alpha\beta\gamma \in F} \left\{ w_1^{\gamma} H(x_{\alpha\beta\gamma}; b^m_{\alpha\beta\gamma}) \right.\right.\right.$$
$$\left. + \sum_{[i,j] \sqsubseteq \gamma} (w_j^{\gamma} - w_i^{\gamma}) H(x_{pa_i^{\alpha\beta\gamma}}; b^m_{pa_i^{\alpha\beta\gamma}}) \right\} - \langle \theta, b'^m \rangle - \sum_{i \in Y \cup H} \omega_i H(x_i; b_i'^m) - \sum_{\alpha\beta\gamma \in F} \left\{ \omega_1^{\beta\gamma} H(x_{\alpha\beta\gamma}; b'^m_{\alpha\beta\gamma}) \right.$$
$$\left.\left.\left. + \sum_{[i,j] \sqsubseteq \beta\gamma} (\omega_j^{\beta\gamma} - \omega_i^{\beta\gamma}) H(x_{pa_i^{\alpha\beta\gamma}}; b'^m_{pa_i^{\alpha\beta\gamma}}) \right\} \right\} - \frac{\lambda}{2} ||\theta||^2 \right\} \quad (5)$$

Below are second order partial derivatives with respect to θ and b, which are diagonal elements of Hessian matrix:

$$\frac{\partial^2 L}{\partial \theta_i^2} = -\lambda$$
$$\frac{\partial^2 L}{\partial \theta_{\alpha\beta\gamma}^2} = -\lambda$$
$$\frac{\partial^2 L}{\partial {b_i^m}^2} = -\frac{w_i}{b_i^m(x_i)}$$
$$\frac{\partial^2 L}{\partial {b^m_{\alpha\beta\gamma}}^2} = -\frac{w_1^{\gamma}}{b^m_{\alpha\beta\gamma}}$$
$$\frac{\partial^2 L}{\partial {b^m_{pa_i^{\alpha\beta\gamma}}}^2} = -\frac{(w_j^{\gamma} - w_i^{\gamma})}{b^m_{pa_i^{\alpha\beta\gamma}}}$$

We can see that these terms are negative given $w_j^\gamma > w_i^\gamma$. Also, off diagonal terms of Hessian matrix are:

$$\frac{\partial^2 L}{\partial \theta_i b_i^m} = 1$$
$$\frac{\partial^2 L}{\partial b_i^m \theta_i} = 1$$
$$\frac{\partial^2 L}{\partial \theta_{\alpha\beta\gamma} b_{\alpha\beta\gamma}^m} = 1$$
$$\frac{\partial^2 L}{\partial b_{\alpha\beta\gamma}^m \theta_{\alpha\beta\gamma}} = 1$$

All other off-diagonal terms are zero. For the Hessian matrix to be diagonally dominant and negative semi-definite, following conditions need to hold for each clique $\alpha\beta\gamma$ and each assignment to x_i and $x_{\alpha\beta\gamma}$:

$$w_j^\gamma \geq w_i^\gamma$$
$$\omega_j^{\beta\gamma} \geq \omega_i^{\beta\gamma}$$
$$\lambda > \frac{b_i^m(x_i)}{w_i}$$
$$\lambda > \frac{b_{\alpha\beta\gamma}^m(x_{\alpha\beta\gamma})}{w_1^\gamma}$$

Extra constraints on $b^m_{pa_i^{\alpha\beta\gamma}}$ are following:

$$b^m_{pa_i^{\alpha\beta\gamma}}(x_{pa_i^{\alpha\beta\gamma}}) = \sum_{j \preceq i} \sum_{x_j} b^m_{\alpha\beta\gamma}(x_{\alpha\beta\gamma})$$
$$b'^m_{pa_i^{\alpha\beta\gamma}}(x_{pa_i^{\alpha\beta\gamma}}) = \sum_{j \preceq i} \sum_{x_j} b'^m_{\alpha\beta\gamma}(x_{\alpha\beta\gamma})$$

If these constraints are satisfied, then L is concave in $\{\theta, b\}$ and convex in b'. Partial derivatives with respect to b' are similar to those of b, but with inverted sign. Simplifying this equation and expanding entropy terms to compute ΔL.

$$\begin{aligned}
L(\theta, b, b', w, \omega) = \max_\theta \max_b \min_{b'} \sum_{m=1}^{M} \Bigg\{ & \langle \theta, b^m - b'^m \rangle - \sum_{i \in H} \sum_{x_i} w_i b_i^m(x_i) \log b_i^m(x_i) \\
& - \sum_{\alpha\beta\gamma \in F} \sum_{x_\gamma} \Bigg\{ w_1^\gamma b^m_{\alpha\beta\gamma}(x_{\alpha\beta\gamma}) \log b^m_{\alpha\beta\gamma}(x_{\alpha\beta\gamma}) + \sum_{[i,j] \sqsubseteq \gamma} (w_j^\gamma - w_i^\gamma) b^m_{pa_i^{\alpha\beta\gamma}}(x_{pa_i^{\alpha\beta\gamma}}) \log b^m_{pa_i^{\alpha\beta\gamma}}(x_{pa_i^{\alpha\beta\gamma}}) \Bigg\} \\
& + \sum_{i \in Y \cup H} \sum_{x_i} \omega_i b'^m_i(x_i) \log b'^m_i(x_i) + \sum_{\alpha\beta\gamma \in F} \sum_{x_{\beta\gamma}} \Bigg\{ \omega_1^{\beta\gamma} b'^m_{\alpha\beta\gamma}(x_{\alpha\beta\gamma}) \log b'^m_{\alpha\beta\gamma}(x_{\alpha\beta\gamma}) \\
& + \sum_{[i,j] \sqsubseteq \beta\gamma} (\omega_j^{\beta\gamma} - \omega_i^{\beta\gamma}) b'^m_{pa_i^{\alpha\beta\gamma}}(x_{pa_i^{\alpha\beta\gamma}}) \log b'^m_{pa_i^{\alpha\beta\gamma}}(x_{pa_i^{\alpha\beta\gamma}}) \Bigg\} \Bigg\} - \frac{\lambda}{2} ||\theta||^2
\end{aligned}$$

We will use Frank-Wolfe algorithm to maximize with respect to b, θ and to minimize with respect to b' one step at a time as described in [2]. In Frank-Wolfe implementation, we need first order derivatives of L with

b, θ and b', which are given below:

$$\frac{\partial L}{\partial b_i^m(x_i)} = \theta_i(x_i) - w_i - w_i \log\left(b_i^m(x_i)\right)$$
$$\frac{\partial L}{\partial b_{\alpha\beta\gamma}^m(x_{\alpha\beta\gamma})} = \theta_{\alpha\beta\gamma}(x_{\alpha\beta\gamma}) - w_1^\gamma - w_1^\gamma \log b_{\alpha\beta\gamma}^m(x_{\alpha\beta\gamma})$$
$$\frac{\partial L}{\partial b_{pa_i^{\alpha\beta\gamma}}^m(x_{pa_i^{\alpha\beta\gamma}})} = w_i^\gamma - w_j^\gamma + (w_i^\gamma - w_j^\gamma) \log b_{pa_i^{\alpha\beta\gamma}}^m(x_{pa_i^{\alpha\beta\gamma}})$$
$$\text{where, } [i,j] \sqsubseteq \alpha\beta\gamma.$$
$$\frac{\partial L}{\partial \theta_i(x_i)} = \sum_{m=1}^{M} \{b_i^m(x_i) - b_i'^m(x_i)\} - \lambda\theta_i(x_i)$$
$$\frac{\partial L}{\partial \theta_{\alpha\beta\gamma}(x_{\alpha\beta\gamma})} = \sum_{m=1}^{M} \{b_{\alpha\beta\gamma}^m(x_{\alpha\beta\gamma}) - b_{\alpha\beta\gamma}'^m(x_{\alpha\beta\gamma})\} - \lambda\theta_{\alpha\beta\gamma}(x_{\alpha\beta\gamma})$$

Derivatives with respect to b' are same as above, but with inverted signs. They are given below.

$$\frac{\partial L}{\partial b_i'^m(x_i)} = -\theta_i(x_i) + \omega_i + \omega_i \log\left(b_i'^m(x_i)\right)$$
$$\frac{\partial L}{\partial b_{\alpha\beta\gamma}'^m(x_{\alpha\beta\gamma})} = -\theta_{\alpha\beta\gamma}(x_{\alpha\beta\gamma}) + \omega_1^{\beta\gamma} + \omega_1^{\beta\gamma} \log b_{\alpha\beta\gamma}'^m(x_{\alpha\beta\gamma})$$
$$\frac{\partial L}{\partial b_{pa_i^{\alpha\beta\gamma}}'^m(x_{pa_i^{\alpha\beta\gamma}})} = \omega_i^{\beta\gamma} - \omega_j^{\beta\gamma} + (\omega_i^{\beta\gamma} - \omega_j^{\beta\gamma}) \log b_{pa_i^{\alpha\beta\gamma}}'^m(x_{pa_i^{\alpha\beta\gamma}})$$
$$\text{where, } [i,j] \sqsubseteq \alpha\beta\gamma.$$

2.1.1 Exact inference for trees

We can use exact inference algorithms such as Belief Propagation on tree structures instead of doing approximate inference. Our exact objective that we need to optimize is following:

$$L = \max_\theta \left\{ \log \sum_{x_H}^{\tau_H} \exp\left[\sum_{\alpha\beta\gamma} \theta_{\alpha\beta\gamma}(x_\alpha^m, x_\beta^m, x_\gamma)\right] - \log \sum_{x_Y, x_H}^{\tau_Y, \tau_H} \exp\left[\sum_{\alpha\beta\gamma} \theta_{\alpha\beta\gamma}(x_\alpha^m, x_\beta, x_\gamma)\right] \right\}$$

Where, $\tau_Y \to 0^+$ and $\tau_H = 1$. Writing it in variational forms,

$$L(\theta, b, b', w, \omega) = \max_\theta \left\{ \max_{b \in L(G_1)} \left\{ \langle \theta, b \rangle + \sum_{i \in H} \tau_i H(x_i; b_i) + \sum_{\alpha\beta\gamma \in F} \sum_{i \in \gamma} \tau_i H(x_i | x_{pa_i^{\alpha\beta\gamma}}; b_{\alpha\beta\gamma}) \right\} \right.$$
$$\left. - \max_{b' \in L(G_2)} \left\{ \langle \theta, b' \rangle + \sum_{i \in Y \cup H} \tau_i H(x_i; b_i') + \sum_{\alpha\beta\gamma \in F} \sum_{i \in \beta\gamma} \tau_i H(x_i | x_{pa_i^{\alpha\beta\gamma}}; b_{\alpha\beta\gamma}') \right\} \right\}$$

To find b and b' that maximize these two Bethe free energy formulations, we can carry out two-stage message passing with following messages:

$$m_{i \to j}(x_j) = \sum_{x_i} \phi_i(x_i) \psi_{ij}(x_i, x_j) \prod_{k \in N(i) \setminus j} m_{k \to i}(x_i)$$

where, $\phi_i(x_i) = \exp(\frac{\theta_i(x_i)}{\tau_i})$ and $\psi_{ij}(x_i, x_j) = exp(\frac{\theta_i j(x_i, x_j)}{\tau_i})$

References

[1] Wei Ping, Qiang Liu, and Alexander Ihler. Decomposition Bounds for Marginal MAP. *Advances in Neural Information Processing Systems 28 (NIPS 2015)*, pages 1–9, 2015.

[2] Kui Tang, Nicholas Ruozzi, David Belanger, and Tony Jebara. Bethe Learning of Graphical Models via MAP Decoding. *Artificial Intelligence and Statistics (AISTATS)*, 51, 2016.