15. Supervised parameter estimation for GRFs

A. Generative learning

- $S = \{S_i \mid i \in V\}$ is a K-valued GRF on a graph (V,E) with joint p.d.

$$p_u(s) = \frac{1}{Z(u)} \exp\Big[\sum_{i\in V} u_i(s_i) + \sum_{ij\in E} u_{ij}(s_i, s_j)\Big]$$

- $T = \{s^j \in K^V \mid j = 1, \dots, m\}$ is an i.i.d. training sample

<u>Task</u>: Estimate unary and pairwise potentials (i.e. model parameters) u_i, u_{ij} from training data

Maximum likelihood estimator

$$L(u) = \frac{1}{m} \sum_{s\in T} \log \frac{1}{Z(u)} \exp U(s) \to \max_u$$

Using the exponential family representation (Sec. 6), we get

$$L(u) = \frac{1}{m} \sum_{s\in T} \log \frac{1}{Z(u)} e^{\langle \varphi(s), u\rangle}$$

$$= \frac{1}{m} \sum_{s\in T} \langle \varphi(s), u\rangle - \log \sum_{s\in K^V} e^{\langle \varphi(s), u\rangle} \to \max_u$$

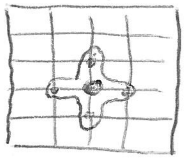
The task has the structure $\langle \psi, u\rangle - g(u) \to \max_u$ with a convex function $g(u)$. Can we solve it by gradient ascent? Computing $\nabla g(u)$ requires to compute statistics of $\varphi(u)$, i.e. computing unary and pairwise marginal probabilities.

<u>Remark 1</u> The learning task is easy to solve for acyclic graphs (V,E). See Sec. 7.

B. Pseudo-likelihood estimator (Besag, 1975)

Recall the Gibbs sampler (Sec. 14), which is defined by the conditional distributions

$$p(s_i \mid s_{\mathcal{N}_i}), \quad i \in V, \; s_i \in K$$



and in turn defines the joint p.d. $p(s)$

Idea: Use the pseudo-likelihood estimator defined by

$$L_p(u) = \frac{1}{m} \sum_{s \in T} \sum_{i \in V} \log p_u(s_i \mid s_{\mathcal{N}_i}) \to \max_u$$

where

$$\log p_u(s_i \mid s_{\mathcal{N}_i}) = \log \frac{\exp\left[u_i(s_i) + \sum_{j \in \mathcal{N}_i} u_{ij}(s_i, s_j)\right]}{\sum_{s_i \in K} \exp\left[\text{—— " ——}\right]}$$

$$= u_i(s_i) + \sum_{j \in \mathcal{N}_i} u_{ij}(s_i, s_j) - \log \sum_{k \in K} \exp\left[u_i(k) + \sum_{j \in \mathcal{N}_i} u_{ij}(k, s_j)\right]$$

$\Rightarrow$ $L_p(u)$ is a concave function of u and its gradient is easy to compute.

Theorem 1 (w/o proof)

The pseudo-likelihood estimator is asymptotically consistent for GRFs. However, it has a higher variance than MLE. ■

Remark 2

The pseudo-likelihood estimator can be easily generalised for GRF models as in C. ■

C. Discriminative learning

- X, S is a pair of $\mathcal{X}$-valued and K-valued random fields on a graph (V, E) with joint p.d.

$$p_u(x,s) = \frac{1}{Z(u)} \exp\Big[\sum_{i\in V} u_i(x_i,s_i) + \sum_{ij\in E} u_{ij}(s_i,s_j)\Big]$$

- loss function $\ell(s,s') = \sum_{i\in V} [\![s_i \neq s_i']\!]$
- i.i.d. training data $\mathcal{T} = \{(x^j, s^j) \mid x^j \in \mathcal{X}^V, s^j \in K^V, j = 1,\dots,m\}$

Task Estimate unary and pairwise potentials by minimising the empirical risk on training data.

$$R(u, \mathcal{T}) = \frac{1}{m} \sum_{(x,s)\in\mathcal{T}} \ell\big(s, \operatorname*{argmax}_{s'\in K^V} p_u(x,s')\big)$$

$$= \frac{1}{m} \sum_{(x,s)\in\mathcal{T}} \ell\big(s, \operatorname*{argmax}_{s'\in K^V} \langle \varphi(x,s'), u\rangle\big) \to \min_u$$

The objective function is piecewise constant $\Rightarrow$ replace the true loss by a surrogate loss, e.g. margin rescaling loss

$$\tilde{R}(u, \mathcal{T}) = \frac{1}{m} \sum_{(x,s)\in\mathcal{T}} \max_{s'\in K^V} \Big[\ell(s,s') - \langle \varphi(x,s), u\rangle + \langle \varphi(x,s'), u\rangle\Big]$$

The objective upper bounds $R(u, \mathcal{T})$ and is convex in u. Computing its subgradient amounts to solve a $(\max,+)$ problem for each training example.

Remark 3 The same approach can be applied for conditional random fields, where $p(s \mid x)$ is modelled as a GRF. ■