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.

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

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

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

Consider the maximum likelihood estimator

$$L(u) = \frac{1}{\ell} \sum_{s \in T} \log p_u(s) \to \max_u$$

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

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

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

The task has the structure $\langle \Psi, u \rangle - g(u) \to \max_u$ with convex $g(u)$. Can we solve it by gradient ascent?

$$\nabla g(u) = \nabla \log Z(u) = \mathbb{E}_u(\varphi)$$

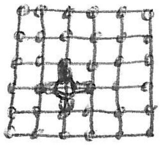
Computing the gradient of g requires to compute statistics of φ, i.e. to compute unary and pairwise marginal prob's.

Remark 1 The learning task is easy to solve for acyclic graphs (V, E) because there $\exists$ a closed form expression for the joint p.d. in terms of marginal statistics. ∎

Pseudo-likelihood estimator (Besag, 1975)

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

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



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

Idea: use the pseudo-likelihood estimator

$$L_p(u) = \frac{1}{\ell} \sum_{s \in T}{}' \sum_{i \in V}{}' \log p_u(s_i \mid s_{\mathcal{N}_i}) \to \max_u$$

where

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

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

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

Theorem 1 (w/o proof)

The pseudo-likelihood estimator is consistent for GRFs, but has a higher variance than the MLE. ∎

3. Discriminative learning

- X, S a pair of F-valued and K-valued random fields on a graph (V, E) with 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 $\ell(s,s') = \sum_{i\in V} \mathbb{1}\{s_i \neq s_i'\}$

- i.i.d. training data $\mathcal{T} = \{(x^j, s^j) \mid x^j \in F^V, s^j \in K^V, j=1,\dots,m\}$

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

$$R(u,\mathcal{T}) = \frac{1}{m}\sum_{j=1}^{m} \ell(s^j, \operatorname*{argmax}_{s\in K^V} p_u(x^j, s))$$

$$= \frac{1}{m}\sum_{j=1}^{m} \ell(s^j, \operatorname*{argmax}_{s\in K^V} \langle \varphi(x^j, s), u\rangle) \to \min_u$$

The objective function is neither continuous nor convex $\Rightarrow$ replace the true loss by some surrogate loss, e.g. margin rescaling loss

$$\tilde{R}(u,\mathcal{T}) = \frac{1}{m}\sum_{j=1}^{m} \max_{s\in K^V}\Big[\ell(s^j, s) - \langle \varphi(x^j, s^j), u\rangle + \langle \varphi(x^j, s), u\rangle\Big]$$

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

Remark 2 The same approach can be applied for Conditional Random Fields, where $p(s|x)$ is modelled as a GRF.