10. Unsupervised learning, EM algorithm

Given: i.i.d. training data $\mathcal{T}_\ell = \{x^j \in F^n \mid j = 1, \dots, \ell\}$

Task: $$\vec{u}_* \in \arg\max_{\vec{u}} \sum_{x \in F^n}' \beta(x) \log \sum_{s \in K^n}' p_{\vec{u}}(x, s)$$

Substituting $p_{\vec{u}}$ we get

$$\underbrace{\log Z(\vec{u})}_{g(\vec{u})} - \underbrace{\sum_{x \in F^n}' \beta(x) \log \sum_{s \in K^n}' \exp \langle \vec{\varphi}(x,s), \vec{u} \rangle}_{h(\vec{u})} \to \min_{\vec{u}} \quad (1)$$

The objective function is a difference of convex functions, i.e. we have to solve a DC-program

A. DC-programs, DC-duality, DC-algorithm

Definition 1 The Fenchel conjugate of a function $g: \mathbb{R}^n \to [-\infty, +\infty]$ is the function $g^*: \mathbb{R}^n \to [-\infty, +\infty]$ defined by

$$g^*(\vec{v}) = \sup_{\vec{u} \in \mathbb{R}^n} \{ \langle \vec{v}, \vec{u} \rangle - g(\vec{u}) \}$$

- The function g^* is convex
- If g is convex and closed then

$$\vec{v} \in \partial g(\vec{u}) \iff \vec{u} \in \partial g^*(\vec{v}) \quad (2)$$

- Each DC program has a DC-dual program

$$g(\vec{u}) - h(\vec{u}) \to \min_{\vec{u}}$$

$$h^*(\vec{v}) - g^*(\vec{v}) \to \min_{\vec{v}}$$

The optimal values of both tasks coincide

- The <u>DC-algorithm</u> aims at solving both tasks simultaneously by constructing a <u>pair</u> of sequences $\vec{u}^{(t)}, \vec{v}^{(t)}$ in an alternating way

a) $\vec{v}^{(t)} \in \partial h(\vec{u}^{(t)})$

b) $\vec{u}^{(t+1)} \in \partial g^*(\vec{v}^{(t)})$

B. <u>DC-algorithm $\hat{=}$ EM algorithm for (1)</u>

Applying DCA for (1) is a „reincarnation" of the EM-algorithm:

Choose an arbitrary $\vec{u}^{(0)}$. Iterate

a) E-step: compute

$$\vec{v}^{(t)} = \nabla h(u^{(t)})$$

$$= \sum_x \beta(x) \sum_s \exp\langle \vec{\varphi}(x,s), \vec{u}^{(t)} \rangle \vec{\varphi}(x,s) \Big/ \sum_s \exp\langle \varphi(x,s), \vec{u}^{(t)} \rangle$$

$$= \sum_{x,s} \beta(x)\, p_{\vec{u}^{(t)}}(s \mid x)\, \vec{\varphi}(x,s)$$

i.e. $\vec{v}^{(t)}$ is the vector of marginal statistics of the distr. $\beta(x)\, p_{\vec{u}^{(t)}}(s \mid x)$

For this it is necessary to compute pairwise posterior marginals $p_{\vec{u}^{(t)}}(s_{i-1}, s_i \mid x)$ $\forall\, i = 2, .., n$ for each $x \in \mathcal{T}_e$ (see sec. 5).

b) M-step: compute $\vec{u}^{(t+1)} \in \partial g^*(\vec{v}^{(t)})$

From Def. 1 and (2) we have

$$\vec{u}_* \in \underbrace{\operatorname*{argmax}_{\vec{u}} \left[\langle \vec{v}^{(t)}, \vec{u} \rangle - \log Z(\vec{u}) \right]}_{MLE} \Rightarrow \vec{u}_* \in \partial g^*(\vec{v}^{(t)})$$

But this task is easy to solve (see sec. 7)! ∎

Theorem 1 (w/o proof)

- The sequences $g(\vec{u}^{(t)}) - h(\vec{u}^{(t)})$ and $h^*(\vec{v}^{(t)}) - g^*(\vec{v}^{(t)})$ are non-increasing
- The sequence $\vec{v}^{(t)}$ is convergent. Its fixpoint is a local minimum of $h^*(\vec{v}) - g^*(\vec{v})$. ∎