8. Empirical risk minimisation

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

loss function $C(s, s') = \mathbb{1}\{s \neq s'\}$

Recall: risk (average loss) for an inference strategy $q: F^n \to K^n$

$$\sum_{x,s} p(x,s)\, \mathbb{1}\{s \neq q(x)\} = \sum_{x,s} p(x,s)\left[1 - \delta(s, q(x))\right]$$

$$= 1 - \sum_{x} p(x, q(x)) \to \min_{q}.$$

I.e., if the true model $p_{\vec{u}}(x,s)$ is known $\to$ optimal inference strategy is

$$q_{\vec{u}}(x) \in \operatorname*{argmax}_{s \in K^n} p_{\vec{u}}(x,s)$$

But, the true model is not known $\to$ replace risk by empirical risk

$$\sum_{x,s} p(x,s)\, \mathbb{1}\{s \neq q_{\vec{u}}(x)\} \approx \frac{1}{\ell} \sum_{(x,s) \in T} \mathbb{1}\{s \neq q_{\vec{u}}(x)\} \to \min_{\vec{u}}$$

This task is not tractable in general, the objective function is neither convex nor differentiable.

However, it is simple to solve for HMMs if there exists a $\vec{u}^*$ such that

$$s^j = \operatorname*{argmax}_{s \in K^n} p_{\vec{u}^*}(x^j, s) \quad \forall\, j = 1, \dots, \ell$$

i.e. there $\exists$ a $\vec{u}^*$ for which the empirical risk is zero.

The above equations are equivalent to the following system of inequalities

$$\langle \vec{\varphi}(x^j, s^j), \vec{u}^* \rangle > \langle \vec{\varphi}(x^j, s), \vec{u}^* \rangle \quad \forall\, s \neq s^j$$
$$\forall\, j = 1, 2, \dots, \ell$$

Such a $\vec{u}^*$ can be found by the perceptron algorithm (if it exists):

Start with arbitrary $\vec{u}$ and iterate

- solve $\tilde{s}^j = \underset{s \in K^n}{\operatorname{argmax}} \langle \vec{\varphi}(x^j, s), \vec{u} \rangle \quad \forall j = 1, \dots, \ell$

  for an HMM this can be done by the algorithm given in sec. 4 (dynamic programming)

- if for some j $\tilde{s}^j \neq s^j$, update $\vec{u}$ by

$$\vec{u} \to \vec{u} + \vec{\varphi}(x^j, s^j) - \vec{\varphi}(x^j, \tilde{s}^j)$$

□

Let us reconsider the general task of empirical risk minimisation

$$q_{\vec{u}}(x) \in \underset{s \in K^n}{\operatorname{argmax}} \langle \vec{\varphi}(x, s), \vec{u} \rangle$$

$$\frac{1}{\ell} \sum_{(x,s) \in T} \mathbb{1}\{s \neq q_{\vec{u}}(x)\} \to \min_{\vec{u}}$$

Approximate the loss (as a function of $\vec{u}$) by a convex upper bound. E.g. „margin rescaling" loss surrogate

$$\mathbb{1}\{s \neq q_{\vec{u}}(x)\} \leq \max_{s' \in K^n} \left[\mathbb{1}\{s \neq s'\} + \langle \vec{\varphi}(x, s') - \vec{\varphi}(x, s), u \rangle \right]$$

The approximation task reads

$$\frac{1}{\ell} \sum_{j=1}^{\ell} \max_{s \in K^n} \left[\mathbb{1}\{s \neq s^j\} + \langle \vec{\varphi}(x^j, s) - \vec{\varphi}(x^j, s^j), \vec{u} \rangle \right] \to \min_{\vec{u}}$$

- Solve it by subgradient descent, cutting plane algorithm, ...
- The inner optimisation tasks $\max_{s \in K^n} [\dots]$ are solved by the algorithm given in sec. 4

9. A sidestep: non-convex optimisation, DC-algorithm

Let $\mathbb{E}$ denote a finite dimensional Euclidian space

- A function $f: \mathbb{E} \to (-\infty, +\infty]$ is convex if its epigraph

$$epi(f) = \{(x, r) \in \mathbb{E} \times \mathbb{R} \mid f(x) \leq r\}$$

is a convex set.

- The domain of a function $f: \mathbb{E} \to (-\infty, +\infty]$ is the set

$$dom\, f = \{x \in \mathbb{E} \mid f(x) < +\infty\}$$

- An element φ of $\mathbb{E}$ is a subgradient of f in $\bar{x}$ if it satisfies

$$\langle \varphi, x - \bar{x} \rangle \leq f(x) - f(\bar{x}) \quad \forall\, x \in \mathbb{E}$$

The set of subgradients of f in $\bar{x}$ (called subdifferential) is denoted by $\partial f(\bar{x})$. ($\partial f(\bar{x}) = \emptyset$ if $\bar{x} \notin dom\, f$)

- The Fenchel conjugate of a function $f: \mathbb{E} \to [-\infty, +\infty]$ is the function $f^*: \mathbb{E} \to [-\infty, +\infty]$ defined by

$$f^*(\varphi) = \sup_{x \in \mathbb{E}} \{\langle \varphi, x \rangle - f(x)\}$$

The function f^* is convex. If f is convex and closed then

$$\varphi \in \partial f(x) \iff x \in \partial f^*(\varphi)$$

Let $g, h : \mathbb{E} \to (-\infty, +\infty]$ be convex functions and consider the problem

$$g(x) - h(x) \to \min_{x \in \mathbb{E}}$$

We call such a problem DC-problem (difference of convex functions).

- Each DC-program has a DC-dual program

$$g(x) - h(x) \to \min_{x \in \mathbb{E}}$$

$$h^*(y) - g^*(y) \to \min_{y \in \mathbb{E}}$$

The optimal values of both tasks coincide

- The <u>DC-algorithm</u> aims at solving both task simultaneously by constructing a <u>pair</u> of decreasing sequences $x^{(t)}, y^{(t)}$:

a) $y^{(t)} \in \partial h(x^{(t)})$

b) $x^{(t+1)} \in \partial g^*(y^{(t)})$