8. Supervised learning of HMMs: Empirical risk minimisation

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

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

Recall: optimal predictor $h: F^n \to K^n$ for 0-1 loss is

$$h_u(x) \in \operatorname*{argmax}_{s \in K^n} p_u(x, s)$$

But, the model is not known $\Rightarrow$ minimise empirical risk

$$\frac{1}{m} \sum_{j=1}^{m} \mathbb{1}\{s^j \neq h_u(x^j)\} \to \min_u$$

This task is not tractable in the general case of an HMM because the objective function is neither convex nor differentiable.

Special case: Suppose there $\exists$ a u^* for which the empirical risk is zero. Can we find it?

Conditions for u^*:

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

or, equivalently

$$\langle \Phi(x^j, s^j), u^* \rangle > \langle \Phi(x^j, s), u^* \rangle \quad \forall s \neq s^j, \forall j = 1, \dots, m$$

This is a system of linear inequalities for $u^* \Rightarrow$ it can be solved by perceptron algorithm:

Start with arbitrary u and iterate

- find $\tilde{s}^j = \operatorname*{argmax}_{s \in K^n} \langle \Phi(x^j, s), u \rangle \quad \forall j = 1, \dots, m$

This can be done by the algorithm given in Sec. 4

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

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

General case:

Idea: overcome intractability by approximating the loss (as a function of u) by a convex upper bound

E.g. „margin rescaling" loss surrogate

$$\mathbb{1}\{s \neq h_u(x)\} \leq \max_{s' \in K^n} \{ \mathbb{1}\{s \neq s'\} + \langle \varphi(x,s') - \varphi(x,s), u \rangle \}$$

The approximation task reads

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

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