8. Supervised learning of HMMs: Empirical risk minimisation

Given: i.i.d. training data $T = \{(x^j, s^j) \mid x^j \in F^n, s^j \in K^n, j = 1, \dots, m\}$
and a loss function $\ell(s, s') = \mathbb{1}\{s \neq s'\}$

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

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

Empirical risk minimisation:

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

This task is not tractable because the objective function is piecewise constant.

Special case Suppose, there $\exists$ a u^* for which the empirical risk is zero. How to find it?

Conditions for u^*:

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

or, equivalently

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

This is a system of linear inequalities $\Rightarrow$ perceptron algorithm

Start with arbitrary u and iterate

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

  This can be done by the algorithm 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 replacing 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

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

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

9. Unsupervised learning: EM algorithm for HMMs

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

Task: $$u^* \in \arg\max_u \frac{1}{|T|} \sum_{x \in T} \log \sum_{s \in K^n} p_u(x, s)$$

Recall EM algorithm

$$L(u) = \frac{1}{|T|} \sum_{x \in T} \log \sum_{s \in K^n} \frac{\alpha(s \mid x)}{\alpha(s \mid x)} p_u(x, s),$$

where $\alpha(s \mid x) \geq 0$, $\sum_{s \in K^n} \alpha(s \mid x) = 1 \quad \forall x \in T$

Using concavity of log, we get a lower bound

$$L(u) \geq L_B(u) = \frac{1}{|T|} \sum_{x \in T} \sum_{s \in K^n} \alpha(s \mid x) \log p_u(x, s) -$$

$$- \frac{1}{|T|} \sum_{x \in T} \sum_{s \in K^n} \alpha(s \mid x) \log \alpha(s \mid x)$$

EM algorithm: maximise $L_B(u, \alpha)$ by block-coordinate ascent w.r.t. to α-s and u-s. Start with some $u^{(0)}$

E-step: set $\alpha^{(t)}(s \mid x) = p_{u^{(t)}}(s \mid x) \quad \forall s \in K^n, \forall x \in T$

M-step: set

$$u^{(t+1)} \in \arg\max_u \frac{1}{|T|} \sum_{x \in T} \sum_{s \in K^n} \alpha^{(t)}(s \mid x) \log p_u(x, s)$$

Let us analyse the M-step for HMMs. The objective is

$$\frac{1}{|T|} \sum_{x \in T} \sum_{s \in K^n} \alpha^{(t)}(s \mid x) \langle \Phi(x, s), u \rangle \quad - \log Z(u) \to \max_u$$

Denoting $\Psi = \frac{1}{|T|} \sum_{x \in T} \sum_{s \in K^n} \alpha^{(t)}(s|x) \Phi(x,s)$, we get

$$\langle \Psi, u \rangle - \log Z(u) \to \max_u$$

This is equivalent to supervised learning task in Sec. 7. We know how to solve it, provided we can compute Ψ.

Computing Ψ:

- For each $x \in T$ compute

$$\Psi(x) = \sum_{s \in K^n} \alpha^{(t)}(s|x) \Phi(x,s) = \sum_{s \in K^n} p_{u^{(t)}}(s|x) \Phi(x,s),$$

  i.e. we have to compute posterior pairwise marginals $p(s_{i-1}, s_i \mid x)$ $\forall\, i = 2, \dots, n$ and $\forall s_{i-1}, s_i \in K$. This can be done by an algorithm similar to the one discussed in Sec. 5

- The components of Ψ are then obtained by averaging the components of $\Psi(x)$ over all $x \in T$, i.e.

$$\Psi = \frac{1}{|T|} \sum_{x \in T} \Psi(x)$$

Theorem 1 (w/o) proof

The sequence $L(u^{(t)})$ is monotonously increasing and the sequence $\alpha^{(t)}$ is convergent.