6. Formulation of learning tasks for HMMs

⋮

E. Representing an HMM as an exponential family

According to Def 1b in sec. 1, the joint p.d. for a Markov model on a chain can be written as

$$p(s) = \prod_{i=2}^{n} g_i(s_{i-1}, s_i)$$

To allow arbitrary non-negative g-s, we introduce a normalising factor Z. If, in addition, all g-s are strictly positive, we may write

$$p(s) = \frac{1}{Z} \exp \sum_{i=2}^{n} u_i(s_{i-1}, s_i) = \frac{1}{Z} \exp \langle \vec{\varphi}(s), \vec{u} \rangle$$

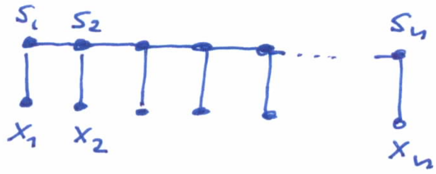
The vector $\vec{\varphi}(s)$ is a binary valued indicator vector. The components of $\vec{\varphi}(s)$ are defined as follows

$$\varphi_{ikk'}(s_{i-1}, s_i) = \delta(s_{i-1}, k)\, \delta(s_i, k'), \quad i = 2, \dots, n, \quad k, k' \in K$$

where δ is the Kronecker delta. The components of the parameter vector $\vec{u}$ are given by $u_{ikk'} = u_i(k, k')$

Remark 1 Notice that the normalising factor Z depends on the model parameters $\vec{u}$, i.e. Z is a function of $\vec{u}$.

Since an Hidden Markov model is defined on the graph



its joint p.d. can be written as

$$p(x,s) = \prod_{i=2}^{n} g_i(s_{i-1}, s_i) \prod_{i=1}^{n} \tilde{g}_i(x_i, s_i) .$$

Repeating the same steps, we get the representation

$$p_{\vec{u}}(x,s) = \frac{1}{Z(\vec{u})} \exp\langle \vec{\varphi}(x,s), \vec{u} \rangle$$

7. Supervised learning, ML-estimator

Training data: i.i.d. sample of pairs (x, s)

$$\mathcal{T}_\ell = \{(x^j, s^j) \mid x^j \in F^n,\ s^j \in K^n,\ j = 1, \dots, \ell\}$$

$\hookrightarrow$ empirical probability $\beta(x, s)$

Learning task:

$$\vec{u}_* \in \arg\max_{\vec{u}} \sum_{x \in F^n} \sum_{s \in K^n} \beta(x, s) \log p_{\vec{u}}(x, s) \qquad (1)$$

Intuitive answer: $\vec{u}_*$ is given by

$$p_{\vec{u}_*}(x_i \mid s_i) = \beta(x_i \mid s_i)$$

$$p_{\vec{u}_*}(s_{i-1}, s_i) = \beta(s_{i-1}, s_i)$$

Remark 1 The formula

$$p(s_1, \dots, s_n) = \frac{p(s_1, s_2) \cdot p(s_2, s_3) \cdot \ldots \cdot p(s_{n-1}, s_n)}{p(s_2) \cdot p(s_3) \cdot \ldots \cdot p(s_{n-1})}$$

for a Markov chain (see sec. 1) provides an easy way to compute the components of $\vec{u}$ given the pairwise marginal prob's – simply put, the components of the former are the logarithms of the latter $\blacksquare$

Let us prove, that the intuitive answer given above is indeed true. The objective function of the learning task is

$$L(\vec{u}) = \sum_{x, s} \beta(x, s) \left[\langle \vec{\varphi}(x, s), \vec{u} \rangle - \log Z(\vec{u}) \right]$$

$$= \langle \overline{\vec{\varphi}}, \vec{u} \rangle - \log Z(\vec{u})$$

where $\vec{\Phi} = \sum'_{x,s} \beta(x,s)\, \vec{\varphi}(x,s)$ denotes the empirical mean of the random vector $\vec{\Phi}$.

The first term in L is linear and thus concave.
Let us prove that $\log Z(\vec{u})$ is a convex function of $\vec{u}$

- $$\log Z(\vec{u}) = \log \sum'_{x,s} \exp\langle \vec{\varphi}(x,s), \vec{u} \rangle$$

- $$\nabla \log Z(\vec{u}) = \frac{1}{Z(\vec{u})} \sum'_{x,s} \exp\langle \vec{\varphi}(x,s), \vec{u}\rangle\, \vec{\varphi}(x,s) \overset{!}{=} \mathbb{E}_{\vec{u}}(\vec{\Phi})$$

  i.e. the gradient of $\log Z$ is the expectation of the random vector $\vec{\Phi}$

- $$\nabla^2 \log Z(\vec{u}) = \mathbb{E}_{\vec{u}}(\vec{\Phi} \otimes \vec{\Phi}) - \mathbb{E}_{\vec{u}}(\vec{\Phi}) \otimes \mathbb{E}_{\vec{u}}(\vec{\Phi})$$
  $$= \mathbb{E}_{\vec{u}}\left[\left(\vec{\Phi} - \mathbb{E}_{\vec{u}}(\vec{\Phi})\right) \otimes \left(\vec{\Phi} - \mathbb{E}_{\vec{u}}(\vec{\Phi})\right)\right]$$

  i.e. the second derivative of $\log Z$ is the covariance matrix of the random vector $\vec{\Phi}$.
  It is symmetric and positive semidefinite.

<u>Lemma 1</u> The partition function $\log Z(\vec{u})$ of an HMM (with strictly positive p.d.) is convex in $\vec{u}$.

We conclude, that the objective function $L(\vec{u})$ of the learning task (1) is concave. Hence, it has global maxima only. They are given by

$$\nabla L(\vec{u}) = \sum_{x,s}' \beta(x,s)\vec{\varphi}(x,s) - \mathbb{E}_{\vec{u}}(\vec{\varphi}) = 0$$

But, the components of $\mathbb{E}_{\vec{u}}(\vec{\varphi})$ are the pairwise marginals of the model $p_{\vec{u}}(x,s)$! This proves that the intuitive answer given above is indeed correct: The optimiser $\vec{u}_*$ defines the model which has precisely the same pairwise marginals as the empirical prob. distr. $\beta(x,s)$

Theorem 1 (w/o proof) The maximum likelihood estimator for HMMs is consistent, i.e.

$$P_{\vec{u}}\left(\| \vec{u}_*(\mathcal{T}_\ell) - \vec{u} \| > \varepsilon \right) \xrightarrow{\ell \to \infty} 0$$

for every $\varepsilon > 0$.