Ch I Markov models on chains and acyclic graphs

1. Markov models on chains

1A Definitions & basic properties

- Sequence $S = (s_1, .., s_n)$ of K-valued random variables $s_i \in K$
- K is a finite set, its elements are called states
- $p(S) = p(s_1, .., s_n)$ is a joint probability distr. on K^n

W.l.o.g. we can write

$$p(s_1, .., s_n) = p(s_n \mid s_1, .., s_{n-1})\, p(s_1, .., s_{n-1})$$
$$= \dots$$
$$= p(s_n \mid s_1, .., s_{n-1})\, p(s_{n-1} \mid s_1, .., s_{n-2}) \cdot \ldots \cdot p(s_2 \mid s_1)\, p(s_1)$$

Definition 1a A p.d. on K^n is a Markov chain if

$$p(s) = p(s_1) \prod_{i=2}^{n} p(s_i \mid s_{i-1})$$

holds $\forall\, s \in K^n$ ◻

Definition 1b A p.d. on K^n is a Markov chain if

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

holds $\forall\, s \in K^n$, where $g_i : K^2 \to \mathbb{R}_+$ are some functions ◻

Equivalence:

a) → b) trivial

b) → a) recursively apply the following step

$$p(s_{n-1}, s_n) = \Big\{ \sum_{s_1, .., s_{n-2}} \prod_{i=2}^{n-1} g_i(s_{i-1}, s_i) \Big\}\, g_n(s_{n-1}, s_n)$$

↳ $g_n(s_{n-1}, s_n) = p(s_n \mid s_{n-1})\, b_{n-1}(s_{n-1})$ with some b_{n-1}

Therefore, we have

$$p(s_1,\dots,s_n) = \underbrace{\left[\prod_{i=2}^{n-1} g_i(s_{i-1},s_i)\right] b_{n-1}(s_{n-1})}_{p(s_1,\dots,s_{n-1})} \cdot p(s_n \mid s_{n-1})$$

Another useful formula

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

Example 1 (Ehrenfest model)

Consider N particles in two containers. At each discrete time $t=1,2,\dots$, independently from the past, a particle is selected at random and moved to the other container. Let S_t denote the number of particles in the first container at time t. Then we have

$$p(S_t=k \mid S_{t-1}=\ell) = \begin{cases} \frac{N-\ell}{N} & \text{if } k=\ell+1 \\ \frac{\ell}{N} & \text{if } k=\ell-1 \\ 0 & \text{otherwise} \end{cases}$$

Q: How does $p(S_t=k)$, $k=1,\dots,N$ behave for $t\to\infty$?

Example 2 (Random walk on a graph)

Consider a random walk on an undirected graph V, E

- $K=V$ states, $S_t\in V$ position of the walker at time t
- $p(s_1)$ some p.d. for the start vertex
- $p(S_t=i \mid S_{t-1}=j) = \begin{cases} W_{ij} & \text{if } \{i,j\}\in E \\ 0 & \text{otherwise}\end{cases}$

where th W_{ij} fulfill $\sum_{i\in\mathcal{N}(j)} W_{ij} = 1 \quad \forall j\in V$

LB. Homogeneous Markov chains, stationary distributions

Definition 2 A Markov chain is homogeneous if its conditional prob's $p(s_i \mid s_{i-1})$ do not depend on the position i, i.e.

$$p(s_i = k \mid s_{i-1} = k') = q(k, k') \quad \forall\, i = 2, \dots, n. \quad \blacksquare$$

We know that

$$p(s_i = k) = \sum_{k' \in K} p(s_i = k \mid s_{i-1} = k')\, p(s_{i-1} = k').$$

Consider $p(s_i = k)$, $k \in K$ as components of a vector $\pi_i \in \mathbb{R}^K_+$ and $p(s_i = k \mid s_{i-1} = k')$, $k, k' \in K$ as elements of a $K \times K$ matrix P. Then the previous eq. reads

$$\pi_i = P\, \pi_{i-1}$$

and more general, we have $\pi_i = P^{i-1} \pi_1$.

It may happen that there $\exists$ a p.d. π^* on K s.t. $P\pi^* = \pi^*$. We call it a stationary p.d. of P.

Definition 3 A homogeneous Markov chain is irreducible if for each pair of states k, k' there is an $m > 0$ s.t. $P^m_{kk'} > 0$. I.e. there is a non-zero probability to reach state k starting from ~~from~~ state k' (after m transitions) $\blacksquare$

A somewhat stronger condition ensures the existence & uniqueness of a stationary distribution and convergence to it.

Theorem 1 (w/o proof) If for some $m > 0$ all elements of the matrix P^m are strictly positive, then the Markov chain has a unique stationary distribution π^*, which is a fixpoint

$$P^n \pi \xrightarrow{n \to \infty} \pi^* \quad \forall\, \pi$$

Moreover

$$P^n = \Pi^* \otimes e + \mathcal{E}(n),$$

where $e = (1,\dots,1)$ and $\mathcal{E}_{kk'}(n) = \mathcal{O}(h^n)$ with some $0 < h < 1$. ■

<u>Definition 4</u> A Markov chain satisfies the <u>detailed balance condition</u> if it has a stationary distribution $\pi \in \mathbb{R}^K_+$ s.t.

$$p(s_i \mid s_{i-1})\, \pi(s_{i-1}) = p(s_{i-1} \mid s_i)\, \pi(s_i).$$ ■

This means that the reverse Markov chain has the same transition probability matrix as the forward chain.

1C. <u>Hidden Markov models on chains</u>

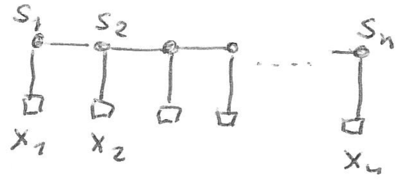
Common models in pattern recognition

$X = (X_1, \dots, X_n)$ sequence of features (observable)

$S = (S_1, \dots, S_n)$ sequence of states (hidden)

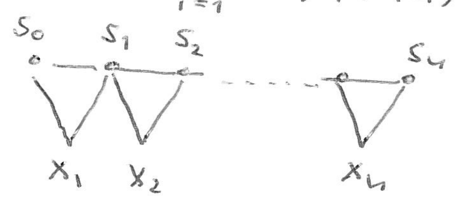
<u>Hidden Markov model</u> (HMM): a p.d. on pairs (X, S) s.t.

a) $$p(x,s) = \underbrace{\prod_{i=1}^{n} p(x_i \mid s_i)}_{p(x\mid s)} \cdot \underbrace{p(s_1) \cdot \prod_{i=2}^{n} p(s_i \mid s_{i-1})}_{p(s)\ -\ \text{Markov model}}$$



b) or slightly more general

$$p(x,s) = p(s_0) \prod_{i=1}^{n} p(x_i, s_i \mid s_{i-1})$$



<u>Remark 1</u> This describes a stochastic regular language. ■

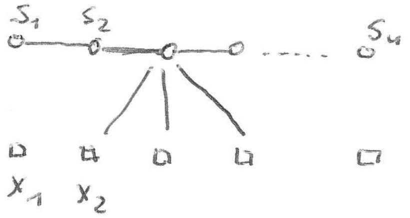
Conditional HMM

As before, $x = (x_1, \dots, x_n)$ - sequence of features and
$s = (s_1, \dots, s_n)$ - sequence of hidden states

Discriminative model → we model only $p(s \mid x)$

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

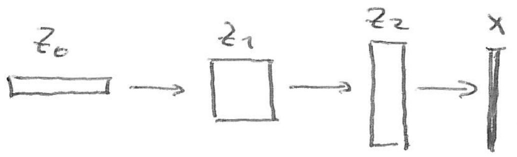
where $Z(x)$ is a normalisation constant



Such models allow to model a direct dependence of s_i on a larger context window of features

Hierarchical variational autoencoders & diffusion models

Generative latent variable models (deep learning)



$z_0, \dots, z_n$ - latent variables (vectors, tensors) x - image

The model is specified by

$p(z_0)$ - simple distribution (uniform, standard Gaussian, etc.)

$p_\theta(z_i \mid z_{i-1})$ - parametrised conditional distributions

$p_\theta(x \mid z_n)$ - conditional distribution on images

If $z_k \in \mathbb{B}^{n_k}$, i.e. z_k is a binary valued vector ⇒

$$\log p(z_k \mid z_{k-1}) = \langle z_k, f(z_{k-1}, \theta) \rangle - C(z_{k-1}),$$

where $f(z_{k-1}, \theta)$ is modelled by a (deep) network.
$C(z_{k-1})$ is the log-partition function (normalising constant)