An introduction to hidden Markov models (HMMs)

25/03/2025

Hidden Markov Models (HMMs) are used to compress, infer & predicts behaviours of stochastic processes.

Stochastic processes

A Stochastic process is a bi-infinite sequence of random variables in time.

$$\underbrace{\ldots X_{-3}\, X_{-2}\, X_{-1}}_{\text{past } X_{-\infty:0}}\ \underbrace{X_0}_{\text{present } X_0}\ \underbrace{X_1\, X_2\, X_3 \ldots}_{\text{future } X_{1:\infty}}$$ where subscripts denote time.

In the most general case the random variable X_t is correlated with the entire past, i.e.,

$$Pr(X_t) = \sum_{x_{-\infty:t}} Pr(X_t | X_{-\infty:t} = x_{-\infty:t})\, Pr(X_{-\infty:t} = x_{-\infty:t}).$$ (the sum becomes an integral in the continuous case)

Some *special* kinds of stochastic processes

1. Independently & identically distributed (IID) processes.
   $Pr(X_t | X_{-\infty:t}) = Pr(X_t)$; $Pr(X_t, X_{t'}) = Pr(X_t)Pr(X_{t'})$; $Pr(X_t) = Pr(X_{t'})\ \forall$
   An example of an IID process is a coin flip.
   Every coinflip has the same statistics over outputs and is independent from the past coin flips.
2. Processes with finite Markov order.
   $Pr(X_t) = f(X_{t-k:t})$ or equivalently $Pr(X_t | X_{-\infty:t}) = Pr(X_t | X_{t-k:t})$

This means that we can discard the history of the past prior to k time steps prior

$$\underbrace{\ldots X_{-k-2}\, X_{-k-3} \ldots}_{\text{we can discard this part of the process.}} X_{-k}\, X_{-k+1} \ldots X_0$$

3. Markov order-1 process when the process at time t depends only one one time step prior.
   $$Pr(X_t | X_{-\infty:t}) = \underbrace{Pr(X_t | X_{t-1:t})}_{\text{We normally write } Pr(X_{t+1}|X_t).}$$

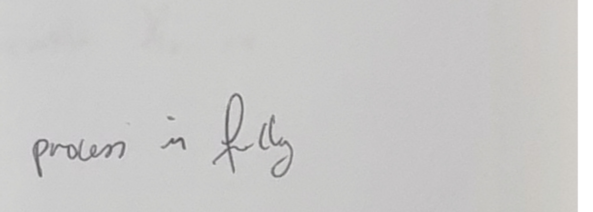
Markov models

If a process has Markov order 1 then the process is fully captured by the transition matrix

$M(X_{t+1} | X_t)$

e.g. if X takes values in $\mathcal{A} := \{x^{(0)}, x^{(1)}\}$, then

$$\underset{\uparrow\ \text{stochastic matrix}}{M(X_{t+1}|X_t)} = \begin{pmatrix} pr(x^{(0)}_{t+1} | x^{(0)}_t) & pr(x^{(0)}_{t+1} | x^{(1)}_t) \\ \underbrace{pr(x^{(1)}_{t+1} | x^{(0)}_t)}_{\text{adds to 1}} & \underbrace{pr(x^{(1)}_{t+1} | x^{(1)}_t)}_{\text{adds to 1.}} \end{pmatrix} \Longleftrightarrow$$



(3)

Then the process at any time-step $t = L$ is given by,

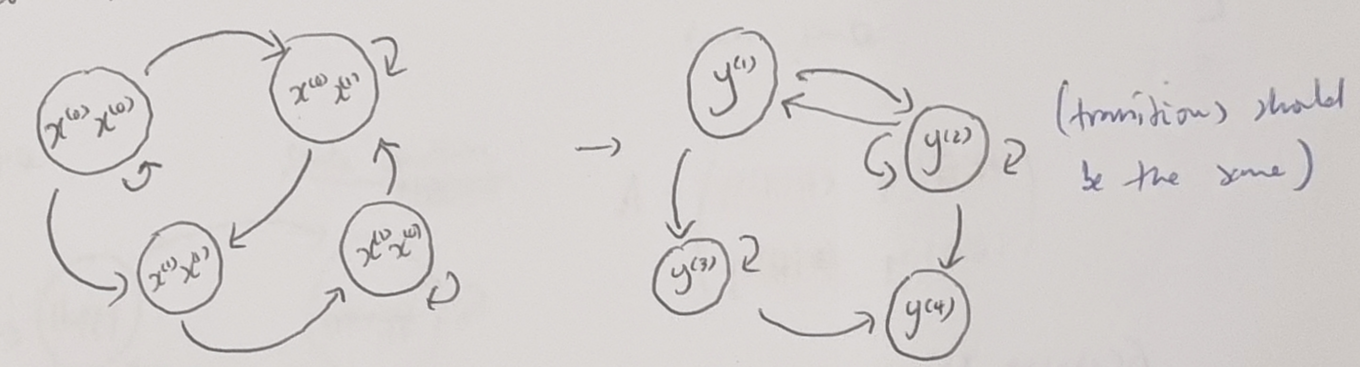
$$\underbrace{Pr(X_L)}_{\text{final distribution}} = \underbrace{Pr(X_L | X_{0:L-1})}_{M} Pr(X_{0:L-1})$$
$$= M\, Pr(X_{0:L-1})$$
$$= M M\, Pr(X_{0:L-1})$$
$$\vdots$$
$$= M^{(L-1)} Pr(X_{0:1})$$
$$= M^{(L-1)} Pr(X_1 | X_0) \underbrace{Pr(X_0)}_{\text{initial distribution}}$$

If a process has Markov order k, then it can still be expressed as a Markov order 1 process by grouping k variables into a new random variable.

$$\ldots X_{-3} X_{-2} X_{-1} X_0 X_1 X_2 X_3 \ldots \rightarrow \ldots Y_{-3} Y_{-2} Y_{-1} Y_0 Y_1 Y_2 Y_3$$

where $Y_t = X_{t:k}$ and the new random variable takes values from an alphabet with $|X|^k$ symbols.

e.g. for $k = 2$.



But say that we have a process that outputs symbols from an alphabet with 2 symbols, and has Markov order 10, then the markov model will require $2^{10} = 1024$ states to express this.

We can do much better with *Hidden* Markov models (HMMs)

(6)

Hidden Markov Models

Markov models fail in cases where there is infinite Markov order or the internal dynamics of the process are unknown (we instead only have a witness of the Markov model).

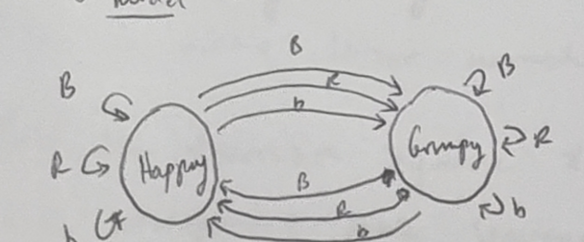
Def: An HMM is composed of

1. $\mathcal{X} := \{x_0, x_1, \ldots, x_N\}$ a set of N observables
2. $\mathcal{S} := \{s_0, s_1, \ldots, s_M\}$ a set of M internal states
3. An $M \times M$ matrix $\underline{A}$, $A_{jk} = \underbrace{Pr(s^{(j)}_{i+1} | s^{(k)}_i)}$ the probability of transitioning between states of the model
4. An $N \times N$ matrix $\underline{B}$, $B_{lm} = \underbrace{Pr(x^{(l)}_{i+1} | s^{(m)}_i)}$ the emission probabilities for each state.

or $\underset{M\times M}{\overset{N}{\{T^{(x^{(i)})}\}}}$

Example

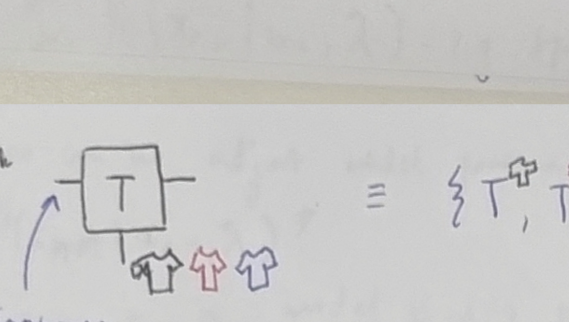
state emitting model



$$A = \begin{pmatrix} pr(H|H) & pr(G|H) \\ pr(H|G) & pr(G|G) \end{pmatrix}$$

$$B = \begin{pmatrix} \cancel{pr(B|B)} & \cancel{pr(R|R)} & \cancel{pr(H|b)} \\ \cancel{pr(G|B)} & \cancel{pr(G|R)} & \cancel{pr(G|b)} \end{pmatrix}$$
$$= \begin{pmatrix} \underbrace{\begin{matrix} p(B|H) \\ p(B|G) \end{matrix}}_{B_B} & \underbrace{\begin{matrix} p(R|H) \\ p(R|G) \end{matrix}}_{B_R} & \underbrace{\begin{matrix} p(b|H) \\ p(b|G) \end{matrix}}_{B_b} \end{pmatrix}$$

edge emitting model



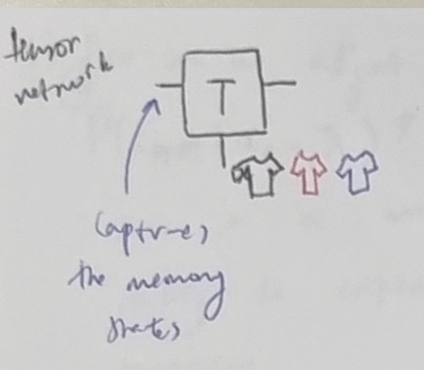
$\{T, T, T\}$

e.g., $pr(R, G|H) = pr(R|G)\, p(G|H)$

with $T^x = \begin{pmatrix} pr(x, H|H) & pr(x, G|H) \\ pr(x, H|G) & pr(x, G|G) \end{pmatrix}$

$\sum T^x = T$ (stochastic)

tensor network



$\equiv \{T, T, T\}$.

!! Generally we are not guaranteed that the model representation is unique. Many presentations can lead to the same output process !!

We can extract probabilities of sequences by once again considering an initial state η_0 which is a distribution over hidden states.

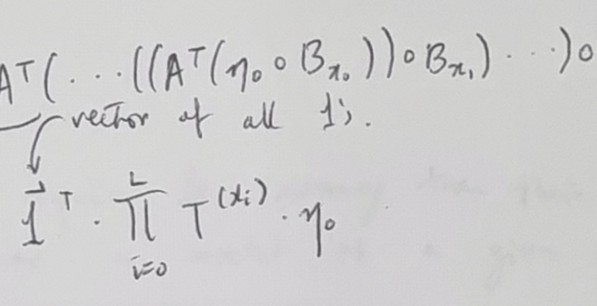
an initial distribution over internal states, expressed as a vector

~~edge~~ state emitting : $Pr(X_{0:L})_{\eta_0} = 1^T((A^T(\ldots((A^T(\eta_0 \circ B_{x_0})) \circ B_{x_1}) \ldots) \circ B_{x_L}))$

vector of all 1's.

edge emitting : $Pr(X_{0:L} = x_{0:L}) = \vec{1}^T \cdot \prod_{i=0}^{L} T^{(x_i)} \cdot \eta_0$

tensor framework : $Pr(X_{0:L}) =$



We can define
→ Stationary states $\sum_x T^{(x)} \eta_s = \eta_s$
→ entropy rates $\sum_{r_s} p(r_s) \log(p(r_s))$ where $\eta_s = (r_s^1, r_s^2, \ldots)$
→ complexity measures.

The three basic problems of HMMs

- Rabiner tutorial

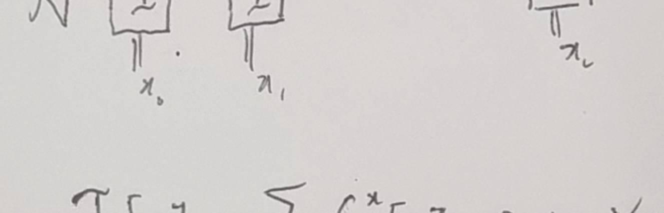
1) Given an observation sequence $x_{0:L}$ & model $\lambda = (A, B, \eta_0) \equiv (\{T^x\}, \eta_0)$ how do we efficiently compute $P(x_{0:L} | \lambda)$, the probability of the sequence given by the model
   (this is where tensor network formulations are particularly useful).
2) Given the observation sequence $x_{0:L}$ & model λ, how can we find the corresponding state sequence $s_{0:L}$ which best explains the data.
   $\max_{s_{0:L}} Pr(x_{0:L} | s_{0:L}, \lambda)$ - e.g. speech recognition, finding patterns in data ML
3) How can we adjust model parameters to maximise $P(x_{0:L} | \lambda)$ or even $P(x_{0:M} | x_{0:m}, \lambda)$?
   This is a model finding problem to maximise the model's ability to explain / fit the data or to predict future sequences.
   (forward/backward method, scissor method, Baum-Welch algorithm).

Quantum Hidden Markov models

Motivation: Quantum models often require less memory than their ~~quantum~~ classical counterparts to generate a model of a given process

arXiv: 2412.12812 for our recent work which partly answers why this is

Def: A quantum hidden Markov model (qHMM) is given by a quantum instrument $\mathcal{I} := \{\mathcal{E}^x\}$ with $\sum_x \mathcal{E}^x = \mathcal{E}$ CPTP.
Together with an initial state ρ_0, the instrument gives rise to a classical stochastic process



$$\mathcal{I}[\rho] = \sum_x \underbrace{\mathcal{E}^x[\rho]}_{\text{quantum update}} \otimes \underbrace{|x\rangle\langle x|}_{\text{classical register}}.$$ } A quantum instrument is a quantum classical tool that can give rise to classical stochastic processes.

$$Pr(X_{0:L} = x_{0:L}) = tr(\mathcal{E}^{x_L} \circ \mathcal{E}^{x_{L-1}} \circ \ldots \circ \mathcal{E}^{x_0}[\rho_0])$$