Engineering Tripos Part IIA THIRD YEAR

Paper 3F8: Inference

Solutions to Example Sheet 3: Sequence Modelling and Monte Carlo Methods

Straightforward questions are marked †
Tripos standard (but not necessarily Tripos length) questions are marked *

Markov Models

1. Markov Models: fitting bi-gram models

   A data scientist observes part of a long sequence that contains $K = 3$ characters: $ABAAABBABCCCBC$. She would like to use a bi-gram model to fit the data with parameters $p(y_1 = k|\theta) = \pi^0_k$ and $p(y_t = k|y_{t-1} = l, \theta) = T_{k,l}$.

   (a) Write down the log-likelihood for the model and optimise it with respect to π^0 and T to find the maximum likelihood parameter estimates.

   (b) Is the maximum-likelihood estimate sensible? How might you improve the estimate?

a) $p(y_{1:T} \mid \underline{\pi}^0, \underline{\underline{T}}) = \left(\prod_k \pi_k^{0\, \mathbb{1}(y_1 = k)} \right) \left(\prod_{t=2}^{T} \prod_{k=1}^{K} \prod_{\ell=1}^{K} T_{k\ell}^{\mathbb{1}(y_t = k, y_{t-1} = \ell)} \right)$

where

$$\mathbb{1}(x) = \begin{cases} 1 & \text{if } x \text{ is true} \\ 0 & \text{if } x \text{ is false} \end{cases}$$

called the indicator function

$$\therefore \log p(y_{1:T} \mid \underline{\pi}^0, \underline{\underline{T}}) = \sum_k \mathbb{1}(y_1 = k) \log \pi_k^0 + \sum_{t=2}^{T} \sum_{k=1}^{K} \sum_{\ell=1}^{K} \mathbb{1}(y_t = k, y_{t-1} = \ell) \log T_{k\ell}$$

$$= \sum_k \mathbb{1}(y_1 = k) \log \pi_k^0 + \sum_{k\ell} N_{k\ell} \log T_{k\ell} = \mathcal{L}(\theta)$$

($N_{k\ell}$: number of transitions from state ℓ to state k)

Maximum likelihood:

$$\frac{d}{d\pi_a^0} \left[\mathcal{L}(\theta) + \lambda \left(\sum_k \pi_k^0 - 1 \right) \right] = \frac{\mathbb{1}(y_1 = a)}{\pi_a^0} + \lambda = 0$$

$$\Rightarrow \pi_a^0 = \mathbb{1}(y_1 = a)$$

$$\frac{d}{dT_{ab}} \left[\mathcal{L}(\theta) + \sum_{\ell=1}^{K} \lambda_\ell \left(\sum_{k=1}^{K} T_{k\ell} - 1 \right) \right] = \frac{N_{ab}}{T_{ab}} + \lambda_b \Rightarrow T_{ab} = \frac{N_{ab}}{\sum_a N_{ab}}$$

(MLE = fraction of transitions from b that end up at a)

In the specific case given:

$$\underline{\pi}^0 = [1, 0, 0] \qquad \underline{\underline{T}} = \begin{bmatrix} 2/5 & 2/5 & 0 \\ 3/5 & 1/5 & 1/3 \\ 0 & 2/5 & 2/3 \end{bmatrix}$$

$$\underline{\underline{N}} = \begin{array}{c} \\ A \\ B \\ C \end{array} \overset{\downarrow \text{from}}{\begin{array}{ccc} A & B & C \\ \hline 2 & 2 & 0 \\ 3 & 1 & 1 \\ 0 & 2 & 2 \end{array}} \quad \text{to} \quad \begin{array}{ccc} 5 & 5 & 4 \end{array}$$

$\sum_a N_{ab}$

$$\mathcal{L}(\theta) = \log \pi_1^{(0)} + 2\log T_{11} + 2\log T_{12} + 3\log T_{21} + \log T_{22} + \log T_{23} + 2\log T_{32} + 2\log T_{33}$$

$$= \log \pi_1^{(0)} + 2\left[\log T_{11} + \log T_{12} + \log T_{32} + \log T_{33}\right] + 3\log T_{21} + \log T_{22} + \log T_{23}$$

b) The MLE is prone to over fitting (eg the estimate for $\underline{\pi}^0$ is deterministic even though we have only seen one data point from $\underline{\pi}^0$)

To avoid such pathologies in estimating $\underline{\pi}^0$ & $\underline{\underline{T}}$ we can use priors to get a MAP or Bayesian estimate.

2. Markov Models: Gaussian AR(1) models

A data scientist observes a sequence of scalar variables $y_{1:T} = \{y_t\}_{t=1}^T$ generated from a Gaussian AR(1) process $y_t = \lambda y_{t-1} + \epsilon_t$ where $\epsilon_t \sim \mathcal{N}(\mu, \sigma^2)$.

She knows that the invariant distribution of the process has the following properties

$$\lim_{t\to\infty} \mathbb{E}(y_t) = \mu_\infty, \quad \lim_{t\to\infty} \mathbb{E}(y_t^2) = \sigma_\infty^2 + \mu_\infty^2, \quad \lim_{t\to\infty} \mathbb{E}(y_t y_{t-1}) = \alpha_\infty.$$

(a) Derive the parameters of the Gaussian AR(1) process $\{\lambda, \mu, \sigma^2\}$ in terms of the properties of the invariant distribution $\{\mu_\infty, \sigma_\infty^2, \alpha_\infty\}$.

(b) The data scientist reinterprets the original Markov model in terms of a new random variable z_t such that $y_t = z_t + \mu_\infty$. State the form of the distribution $p(z_{1:T})$ required for this model to be equivalent to the original one.

Q2 a) First the long way!

$$y_t = \lambda y_{t-1} + \varepsilon_t \qquad \varepsilon_t \sim N(\mu, \sigma^2)$$

① Taking expectations of both sides wrt to all sources of randomness

$$\langle y_t \rangle = \langle \lambda y_{t-1} \rangle + \langle \varepsilon_t \rangle = \lambda \langle y_{t-1} \rangle + \mu$$

If invariant distribution exists then

$$\mu_\infty = \lambda \mu_\infty + \mu \qquad \Rightarrow \mu_\infty = \frac{\mu}{1-\lambda}$$

② Similarly

$$\langle y_t^2 \rangle = \langle (\lambda y_{t-1} + \varepsilon_t)^2 \rangle = \lambda^2 \langle y_{t-1}^2 \rangle + \langle \varepsilon_t^2 \rangle + 2\lambda \langle \varepsilon_t y_{t-1} \rangle$$

$$= \lambda^2 \langle y_{t-1}^2 \rangle + \mu^2 + \sigma^2 + 2\mu\lambda \langle y_{t-1} \rangle$$

if invariant distribution exists then

$$\sigma_\infty^2 + \mu_\infty^2 = \lambda^2 \left[\sigma_\infty^2 + \mu_\infty^2 \right] + \mu^2 + \sigma^2 + 2\mu\lambda\mu_\infty$$

$$\sigma_\infty^2 + \mu_\infty^2 = \frac{\mu^2 + \sigma^2 + 2\mu\lambda\mu_\infty}{1-\lambda^2}$$

③ $\langle y_t y_{t-1} \rangle = \langle (\lambda y_{t-1} + \varepsilon_t) y_{t-1} \rangle = \lambda \langle y_{t-1}^2 \rangle + \mu\mu_\infty$

If invariant distribution exists then;

$$\alpha_\infty = \lambda\left[\sigma_\infty^2 + \mu_\infty^2\right] + \mu\mu_\infty$$

rearrange: $\lambda = \dfrac{\alpha_\infty - \mu_\infty \mu}{\sigma_\infty^2 + \mu_\infty^2}$ & $\mu = \mu_\infty(1-\lambda)$

eliminate λ

So, a procedure for finding λ, μ & σ^2 is

① $\mu = \left(\sigma_\infty^2 + \mu_\infty^2 - \alpha_\infty\right)\dfrac{\mu_\infty}{\sigma_\infty^2}$

② $\lambda = \dfrac{\alpha_\infty - \mu_\infty^2}{\sigma_\infty^2}$

sub ① into $\lambda = \dfrac{\alpha_\infty - \mu_\infty \mu}{\sigma_\infty^2 + \mu_\infty^2}$

③ $\sigma^2 = \left(\sigma_\infty^2 + \mu_\infty^2\right)(1-\lambda^2) - \mu^2 - 2\mu\lambda\mu_\infty$

$= \sigma_\infty^2(1-\lambda^2) + \mu_\infty^2\left[(1-\lambda^2) - (1-\lambda)^2 - 2\lambda(1-\lambda)\right]$

$= \sigma_\infty^2(1-\lambda^2) + \mu_\infty^2\left[1 - \lambda^2 - 1 + 2\lambda - \lambda^2 - 2\lambda + 2\lambda^2\right]$

$= \sigma_\infty^2(1-\lambda^2)$

b) $y_t = z_t + \mu_\infty = \lambda y_{t-1} + \varepsilon_t \qquad \varepsilon_t \sim N(\mu, \sigma^2)$

$= \lambda(z_{t-1} + \mu_\infty) + \varepsilon_t$

$\Rightarrow \quad z_t = \lambda(z_{t-1} + \mu_\infty) - \mu_\infty + \varepsilon_t$

$= \lambda z_{t-1} + \underbrace{\lambda\mu_\infty - \mu_\infty}_{-\mu \text{ (see expression for } \mu_\infty)} + \mu + n_t \qquad n_t \sim N(0, \sigma^2)$

$z_t = \lambda z_t + n_t$

$\Rightarrow p(z_{1:T}) = \prod_t N(z_t; \lambda z_{t-1}, \sigma^2)$ [standard AR(1) model]

This leads to a much faster way of doing part (a) & a useful way of checking this result:

$\sigma_\infty^2 = \langle z_t^2 \rangle$ [standard AR(1) process]

For standard AR(1):

$\sigma_\infty^2 = \dfrac{\sigma^2}{1-\lambda^2}$ (see lectures)

$\therefore \quad \sigma^2 = \sigma_\infty^2(1-\lambda^2)$ much faster!

Hidden Markov Models

3. Discrete Valued Hidden Markov Models*

   (a) Provide the probabilistic equations that define a Hidden Markov Model (HMM) for observed data that takes discrete values. Indicate what aspects of the model the following terms refer to: *initial state probabilities*, *transition matrix* and *emission matrix*.

   (b) Consider a dataset consisting of the following string of 160 symbols from the alphabet $\{A, B, C\}$:

   $$AABBBACABBBACAAAAAAAAAABBBACAAAAABACAAAAAA$$
   $$BBBBACAAAAAAAAAAAAABACABACAABBACAAABBBBACA$$
   $$AABACAAAABACAABACAAABBACAAAABBBBACABBACAA$$
   $$AAAABACABACAAABACAABBBACAAAABACABBACA$$

   Carefully analyse the string. Describe an HMM model for the string. Your description should include the number of states in the HMM, the transition matrix including the values of the elements of the matrix, the emission matrix including the values of its elements, and the initial state probabilities. Explain your reasoning.

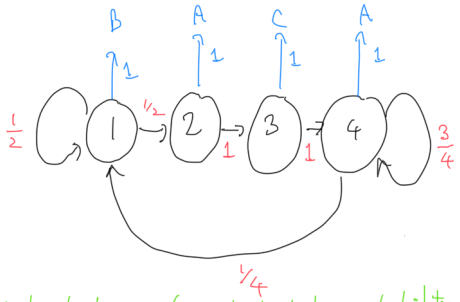
a) bookwork

$p(x_1 = k) = \pi_k$ initial state probabilities

$p(x_t = k \mid x_{t-1} = l) = T_{kl}$ transition probabilities

$p(y_t = m \mid x_t = k) = S_{mk}$ emission probabilities

b) BACA common motif & only way 'C' is emitted & only way A follows B



- only one symbol deterministically emitted from each latent state
- calculate state transition probabilities as follows

Lots of choices for initial state probability

eg $p(4) = 1$ $p(k) \propto n_k$ ↑ number of times in that state

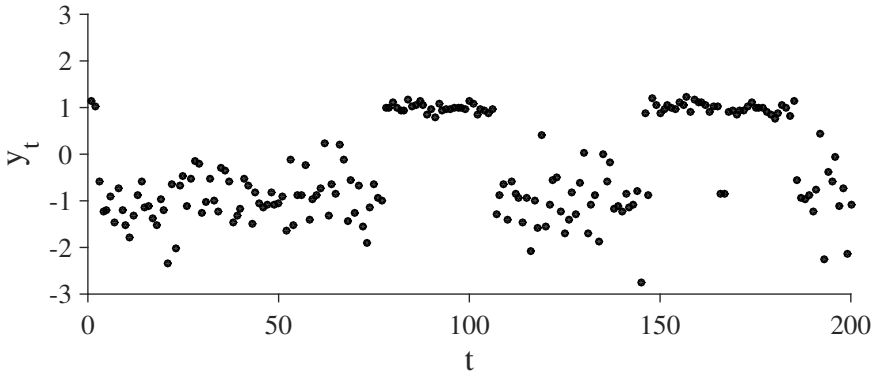
[There are a number of other possible solutions]

AA BB[BACA] BB[BACA] BB[BACA] AAAA AAAA BB [BACA] AAAA [BACA] AAAAA BBB [BACA] AAAAAAA AAAA [BACA] [BACA] A B [BACA] AA BBB [BACA] AA [BACA] AAA [BACA] A [BACA] AAB [BACA] AAA BBB [BACA] B [BACA] AAAAA [BACA] [BACA] AA [BACA] ABB [BACA] AAA [BACA] B [BACA]

from	to	[rough counts - rounded to fractions a human might choose]	
			total
(1)	(1)	(2)	45
	23	22	
(4)	(1)	(4)	
	22	54	76

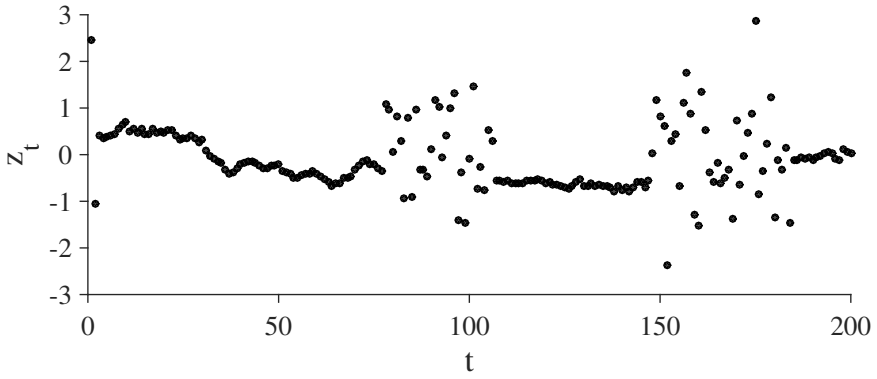
4. Probabilistic Modelling using HMMs for continuous valued observations

   (a) A machine learner observes the time-series, y_t, shown below:



   Suggest a suitable Hidden Markov Model (HMM) for this sequence and state the model's probabilistic equations. Indicate plausible numerical values for the parameters where possible.

   (b) The machine learner is provided with a second set of observations z_t that were measured simultaneously with y_t, shown below:



   Extend the HMM you proposed for part (a) so that it can jointly model the first and second set of observations.

a) Binary hidden state $S \in \{0, 1\}$

$$p(s_1 = 1) = \frac{1}{2} \qquad p(s_t \mid s_{t-1}) = \begin{bmatrix} 0.99 & 0.01 \\ 0.01 & 0.99 \end{bmatrix}$$

$$p(y_t \mid s_t = 1) = \mathcal{N}(y_t;\ \underbrace{1}_{\text{mean}},\ \underbrace{0.1^2}_{\text{variance}})$$

$$p(y_t \mid s_t = 0) = \mathcal{N}(y_t;\ -1,\ (1/2)^2)$$

b) $$p(z_t^{(1)}) = \mathcal{N}(z_t^{(1)};\ 0,\ 1)$$

{ This is known as a switching state-space model

$$p(z_t^{(2)} \mid z_{t-1}^{(2)}) = \mathcal{N}\left(z_t^{(2)};\ \lambda z_{t-1}^{(2)},\ \left(\frac{1}{2}\right)^2 (1 - \lambda^2)\right) \qquad \lambda = 0.99$$

$z_t = s_t z_t^{(1)} + (1 - s_t) z_t^{(2)}$ i.e. $z_t = z_t^{(1)}$ if $s_t = 1$ & $z_t = z_t^{(2)}$ if $s_t = 0$

Again rough estimates for parameter values is fine, the general structure is the main thing to convey

5. Inference in HMMs with Discrete Hidden States†

A Hidden Markov Model contains a discrete hidden state variable x_t that takes one of two values and a discrete observed state y_t that also takes one of two values. The hidden state has a transition probability,

$$\begin{bmatrix} P(x_t = 1|x_{t-1} = 1) & P(x_t = 1|x_{t-1} = 2) \\ P(x_t = 2|x_{t-1} = 1) & P(x_t = 2|x_{t-1} = 2) \end{bmatrix} = \begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix} = \begin{bmatrix} 2/3 & 1/3 \\ 1/3 & 2/3 \end{bmatrix}.$$

The filtering distribution at time $t-1$ is

$$P(x_{t-1}|y_{1:t-1}) = \begin{bmatrix} P(x_{t-1} = 1|y_{1:t-1}) \\ P(x_{t-1} = 2|y_{1:t-1}) \end{bmatrix} = \begin{bmatrix} 1/4 \\ 3/4 \end{bmatrix}.$$

(a) Compute the predictive distribution for the next hidden state variable, $P(x_t|y_{1:t-1})$.

(b) Explain how your solution to part (a) can be used to compute the filtering distribution $P(x_t|y_{1:t})$. What additional piece of information would you require to carry out this computation?

Q5 a)

$$\begin{bmatrix} p(x_t=1 \mid x_{t-1}=1) & p(x_t=1 \mid x_{t-1}=2) \\ p(x_t=2 \mid x_{t-1}=1) & p(x_t=2 \mid x_{t-1}=2) \end{bmatrix} \begin{bmatrix} p(x_{t-1}=1 \mid y_{1:t-1}) \\ p(x_{t-1}=2 \mid y_{1:t-1}) \end{bmatrix}$$

(T ≡ the transition matrix; $p(x_{t-1} \mid y_{1:t-1})$)

$$= \begin{bmatrix} p(x_t=1 \mid x_{t-1}=1)\, p(x_{t-1}=1 \mid y_{1:t-1}) + p(x_t=1 \mid x_{t-1}=2)\, p(x_{t-1}=2 \mid y_{1:t-1}) \\ p(x_t=2 \mid x_{t-1}=1)\, p(x_{t-1}=1 \mid y_{1:t-1}) + p(x_t=2 \mid x_{t-1}=2)\, p(x_{t-1}=2 \mid y_{1:t-1}) \end{bmatrix}$$

$$= \begin{bmatrix} p(x_t=1 \mid y_{1:t-1}) \\ p(x_t=2 \mid y_{1:t-1}) \end{bmatrix}$$

∴ Predictive is

$$\begin{bmatrix} 2/3 & 1/3 \\ 1/3 & 2/3 \end{bmatrix} \begin{bmatrix} 1/4 \\ 3/4 \end{bmatrix} = \begin{bmatrix} 2/12 + 3/12 \\ 1/12 + 6/12 \end{bmatrix} = \frac{1}{12} \begin{bmatrix} 5 \\ 7 \end{bmatrix}$$

b) $p(x_t \mid y_{1:t}) \propto p(y_t \mid x_t)\, p(x_t \mid y_{1:t-1})$

↑ need the likelihood to get the posterior

6. Forecasting in Linear Gaussian State Space Models*

   A simple linear Gaussian state space model with scalar hidden state variables x_t has been used to model scalar observations y_t,

   $$p(x_t|x_{t-1}, \lambda, \sigma^2) = \mathcal{N}(x_t; \lambda x_{t-1}, \sigma^2), \quad p(y_t|x_t, \sigma_y^2) = \mathcal{N}(y_t; x_t, \sigma_y^2).$$

   The Kalman filter recursions have been used to process T observations, $y_{1:T}$, in order to return the posterior distribution over the Tth latent state, $p(x_T|y_{1:T}) = \mathcal{N}(x_T; \mu_T, \sigma_T^2)$.

   (a) Explain how to transform the posterior distribution over the Tth latent state into a forecast for the observations one time step into the future, i.e. express $p(y_{T+1}|y_{1:T})$ in terms of μ_T and σ_T^2.

   (b) Now provide a forecast for the observations τ time steps into the future by expressing $p(y_{T+\tau}|y_{1:T})$ in terms of μ_T and σ_T^2.

   (c) What happens to $p(y_{T+\tau}|y_{1:T})$ as $\tau \to \infty$?

a)

$$p(y_{T+1} \mid y_{1:T}) = N(y_{T+1};\ \lambda\mu_T,\ \lambda^2\sigma_T^2 + \sigma^2 + \sigma_y^2)$$

Calculated by passing $N(x_T; \mu_T, \sigma_T^2)$ through $x_{T+1} = \lambda x_T + \sigma\varepsilon_T$ & then noting $y_{T+1} = x_{T+1} + \sigma_y n_T$ where $\varepsilon_T, n_T \sim N(0,1)$

b) $$x_{T+\tau} = \lambda x_{T+\tau-1} + \sigma\varepsilon_{T+\tau}$$

$$= \lambda(\lambda x_{T+\tau-2} + \sigma\varepsilon_{T+\tau-1}) + \sigma\varepsilon_{T+\tau}$$

$$= \lambda(\lambda(\lambda x_{T+\tau-3} + \sigma\varepsilon_{T+\tau-2}) + \sigma\varepsilon_{T+\tau-1}) + \sigma\varepsilon_{T+\tau}$$

$$\vdots$$

$$x_{T+\tau} = \lambda^\tau x_T + \sigma\sum_{t'=0}^{\tau-1}\lambda^{t'}\varepsilon_{T+\tau-t'}$$

$$\therefore p(y_{T+\tau} \mid y_{1:T}) = N\left(y_{T+\tau};\ \lambda^\tau\mu_T,\ \sigma_y^2 + \sigma^2\sum_{t'=0}^{\tau-1}\lambda^{2t'} + \lambda^{2\tau}\sigma_T^2\right)$$

geometric series: $S_\tau = \sum_{t'=0}^{\tau-1}\lambda^{2t'}$ $\quad S_{\tau+1} = \lambda^2 S_\tau + 1$

c) As $\tau \to \infty$ the forecast will tend to the stationary distribution of the chain:

$S_\infty = \lambda^2 S_\infty + 1$

$$p(y_\infty \mid y_{1:T}) = N\left(y_\infty;\ 0,\ \frac{\sigma^2}{1-\lambda^2} + \sigma_y^2\right)$$

Selected solutions and hints

1. a) $\log p(y_{1:T}|\theta) = \log \pi_1^0 + 2\log(T_{11}T_{12}T_{32}T_{33}) + 3\log T_{21} + \log(T_{22}T_{23})$

$$T = \begin{bmatrix} 2/5 & 2/5 & 0 \\ 3/5 & 1/5 & 1/3 \\ 0 & 2/5 & 2/3 \end{bmatrix}, \quad \pi^0 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

2. a) $\lambda = (\alpha_\infty - \mu_\infty^2)/\sigma_\infty^2$, $\mu = \mu_\infty(1+(\mu_\infty^2 - \alpha_\infty)/\sigma_\infty^2)$, $\sigma^2 = \sigma_\infty^2\left(1 - \left(\frac{\alpha_\infty - \mu_\infty^2}{\sigma_\infty^2}\right)^2\right)$

3. b) pay close attention to repeated patterns and remember that some parts of a HMM can be deterministic

4. b) consider whether the low variance z_t regions are correlated through time and whether a standard HMM could model this

5. a) $p(x_t|y_{1:t-1}) = [5, 7]^\top/12$

6. b) $p(y_{T+\tau}|y_{1:T}) = \mathcal{N}(y_{T+\tau}; \lambda^\tau \mu_T, \sigma_y^2 + \lambda^{2\tau}\sigma_T^2 + \sigma^2 \sum_{t'=0}^{\tau-1} \lambda^{2t'})$

Richard E. Turner