Bayesian Machine Learning - Lecture 7

Guido Sanguinetti

Institute for Adaptive and Neural Computation
School of Informatics
University of Edinburgh
gsanguin@inf.ed.ac.uk

March 4, 2015

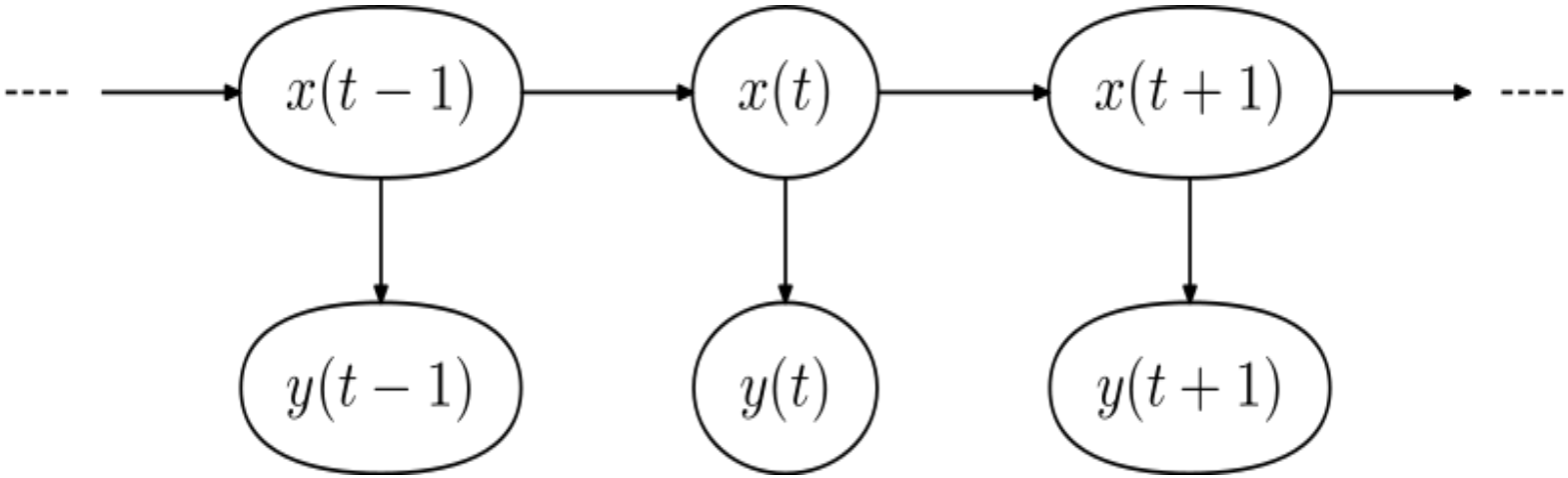
Today's lecture

1. The Forward-Backward algorithm
2. Message passing: Belief Propagation
3. Free energies and the variational principle for inference

Hidden Markov Model definitions

- We have a Markov chain of unobserved states
- Traditionally the latent variables are taken to be discrete and the transition probabilities are given by a transition matrix T
- We have a sequence of observations
- Each observation depends only on the latent state at that time through the *emission probability*

Graphical representation of HMM



- We represent the latent states as a sequence of random variables; each of them depends *only* on the previous one
- The observations depend only on the corresponding state

States and parameters

- We are interested in the posterior distribution of the states $x_{1:T}$ given the observations $y_{1:T}$ (subscript $1:T$ denotes the collection of variables from 1 to T)
- Notice that we only have one observation per time point
- In the independent observations case, this would not be enough
- We also have parameters which we assume known (for now): these are in the known probabilities

$$\pi = p(x(1)) \quad T_{x(t-1),x(t)} = p(x(t)|x(t-1)) \quad O_{x,y} = p(y(t)|x(t)$$

- We assume parameters to be time-independent

The single time marginals

- The joint posterior over the states is, by the rules of probability, proportional to the joint probability of observations and states

$$p\left(x_{1:T}|y_{1:T}\right) \propto p\left(x_{1:T}, y_{1:T}\right)$$

- An object of central importance is the *single time marginal* for the latent variable at time t
- This is obtained by marginalising the latent variables at all other time points; by the proportionality above

$$p\left(x(t)|y_{1:T}\right) \propto p\left(x(t), y_{1:T}\right)$$

Conditional independence and factorisations

- By using the product rule of probability, we can rewrite the joint probability of states and observations as

$$\begin{aligned} & p\left(x_{1:T}, y_{1:T}\right) = \\ & = p\left(y_{t+1:T} | x_{1:T}, y_{1:t}\right) p\left(x_{1:T}, y_{1:t}\right) \end{aligned} \tag{1}$$

- Recall that graphical models encode *conditional independence* relations

Some conditional independencies

- By inspection of the network representation of the model (four slides ago), we see that

$$p\left(y_{t+1:T}|x_{1:T}, y_{1:t}\right) = p\left(y_{t+1:T}|x_{t+1:T}\right) \tag{2}$$

- Also $x_{t+1:T}$ are conditionally independent of $y_{1:t}$ given x_t, so that

$$\begin{aligned} &p\left(x_{1:T}, y_{1:t}\right) = p\left(x_{t+1:T}|x_{1:t}, y_{1:t}\right) p\left(x_{1:t}, y_{1:t}\right) = \\ &p\left(x_{t+1:T}|x_t\right) p\left(x_{1:t}, y_{1:t}\right) \end{aligned} \tag{3}$$

Factorisations and messages

- Putting equations (2,3) into (1), we get

$$p\left(x_{1:T}, y_{1:T}\right) = p\left(y_{t+1:T}, x_{t+1:T} | x_t\right) p\left(x_{1:t}, y_{1:t}\right)$$

- Marginalising $x_{1:t-1}$ and $x_{t+1:T}$ we get the following *fundamental factorisation* of the single time marginal

$$\begin{aligned} p\left(x(t)|y_{1:T}\right) &\propto \alpha(x(t))\beta(x(t)) = \\ &= p\left(x(t)|y_{1:t}\right) p\left(y_{t+1:T}|x(t)\right) \end{aligned} \tag{4}$$

- The single time marginal at time t is the product of the posterior estimate given all the data *up to that point*, times the likelihood of future observations given the state at t

Filtering: computing the forward message

- *Initialisation*:

$$\alpha(1) \propto p\left(y(1), x(1)\right) = \pi O_{x(1),y(1)}$$

- *Recursion*:

$$\begin{aligned}\alpha(t) \propto & p\left(x(t), y_{1:t}\right) = \sum_{x(t-1)} p\left(x(t), x(t-1), y_{1:t}\right) = \\ & = \sum_{x(t-1)} p\left(y(t)|x(t)\right) p\left(x(t)|x(t-1)\right) p\left(x(t-1)|y_{1:t-1}\right) = \\ & = \sum_{x(t-1)} O_{x(t),y(t)} T_{x(t-1),x(t)} \alpha(x(t-1))\end{aligned}$$

  where I used the conditional independences of the network to go from line 1 to 2

- If $x(t)$ is a continuous, replace the sum with an integral

Computing the backward message

- *Initialisation*: $\beta(x(T)) = 1$ (why?)
- *Backward recursion*:

$$\beta(x(t-1)) = p\left(y_{t:T}|x(t-1)\right) = \sum_{x(t)} p\left(y_{t:T}, x(t)|x(t-1)\right) =$$

$$= \sum_{x(t)} p\left(y_{t+1:T}|y(t), x(t), x(t-1)\right) p\left(y(t)x(t)|x(t-1)\right) =$$

$$= \sum_{x(t)} \beta(x(t)) p\left(y(t)|x(t)\right) p\left(x(t)|x(t-1)\right)$$

- Once again, if x is continuous replace sum with integral

Message passing

- The factorisation in equation (4) is an example of *message passing*
- $\alpha(x(t))$ is a message propagated forwards from the previous observations (forward message or filtered process)
- $\beta(x(t))$ is a message propagated backwards from future observations (backward message)
- Message passing algorithms allow exact inference in tree structured graphical models (why?) and approximate inference in more complicated models

General graphical model scenario

- Excellent review by Yedidia et al available here `http://www.merl.com/publications/docs/TR2001-22.pdf`
- We consider the case of *pairwise* Markov random fields, where the joint distribution is given by

$$p(\mathbf{x}) = \frac{1}{Z} \prod_{(ij)} \psi(x_i, x_j) \prod_i \phi(x_i)$$

  where (ij) is the set of edges in the graph and the terms $\phi(x_i)$ are terms coming from (noisy) observations of nodes x_i (*local evidence*)
- The goal is to compute an approximation of the single node marginal, called the *belief* at node i $b(x_i)$

Neighbours and messages

- Given a node i, denote as N_i the set of neighbouring nodes in the graph (recall we work with MRFs, undirected)
- We associate with each edge (ij) a *message* from i to j $m_{ij}(x_j)$
- For graphical models on discrete random variables, the message $m_{ij}(x_j)$ is a vector with dimension the number of states of x_j
- Intuitively, the message $m_{ij}(x_j)$ encodes what node i thinks about the state of node j

Belief propagation

- The *Belief Propagation* (BP) algorithm computes an approximation to the single node belief in terms of messages and local evidence

$$b_i(x_i) \propto \phi(x_i) \prod_{j \in N_i} m_{ji}(x_i) \tag{5}$$

- The messages are computed self-consistently by

$$m_{ji}(x_i) = \sum_{x_j} \phi(x_j)\psi(x_i, x_j) \prod_{l \in N_j \setminus i} m_{lj}(x_j) \tag{6}$$

Why it might make sense

- We proved earlier that BP is exact on trees
- What are the various BP quantities in our Forward-Backward derivation?
- BP essentially posits a *local tree-like* approximation to a general graphical model
- It works well when there are no short loops
- Recent work (by Tony Jebara) showed that exact message passing algorithms (not BP) can also be devised for *perfect graphs* (clique number= colouring number)

Matching distributions through KL divergence

- Given two probability distributions q and p, we have seen that the Kullback-Leibler divergence

$$KL[q\|p] = \sum_x q(x) \log \frac{q(x)}{p(x)}$$

is a convex non-negative functional which vanishes only when the two distributions are the same

- We write the distribution p as

$$p(x) = \frac{1}{Z} \exp\left[-E(x)\right]$$

where E is traditionally known as the *energy*

- The KL divergence then becomes

$$KL[q\|p] = \log Z + \langle E(x) \rangle_q - H\left[q(x)\right] \tag{7}$$

where the functional H is the entropy of a distribution

Free energy

- For fixed p, the KL is a positive functional of q
- If the parametrisation of q is sufficiently rich as to contain the distribution p, minimising the KL divergence would return the *exact* distribution p
- Operationally, this is achieved by minimising the *Free Energy*

$$G(q) = \langle E(x) \rangle_q - H[q(x)]$$

- When p is an (intractable) posterior distribution, this is *the variational principle* of Bayesian inference
- Choosing q in a specific distributional family yields specific variational inference algorithms

Variational formulation of BP

- BP works by positing a local tree-like approximation to the graph
- In terms of distributions, it assumes that an approximating distribution can be defined *only* in terms of 1 and 2 node beliefs $b(x_i)$ and $b(x_i, x_j)$ (subject to $\sum_{x_j} b(x_i, x_j) = b(x_i)$)
- Then, the approximation consists of replacing the entropy term with that of a distribution is defined as

$$q(\mathbf{x}) = \frac{\prod_{(ij)} b(x_i, x_j)}{\prod_i b(x_i)^{(n_i - 1)}} \tag{8}$$

where n_i is the number of neighbours of node i. This is called the *Bethe approximation*

- Exercise: prove that the formula above is the correct joint distribution on a tree (or a singly connected graph in general)

Variational BP again

- Sticking the formulation in equation (8) in the KL divergence (7) yields the so called *Bethe Free Energy*
- Minimising this w.r.t. the beliefs gives an iterative algorithm which is the same as BP
- **PROBLEM**: for general graphs, there may not be a distribution with marginals satisfying the constraints
- In this case, BP will not converge. Theoretical guarantees as to when it will converge are lacking, although the absence of short loops is widely taken to be a good idea