Introduction to Statistical Machine Learning

Christfried Webers

Statistical Machine Learning Group
NICTA
and
College of Engineering and Computer Science
The Australian National University

Canberra
February – June 2013

(Many figures from C. M. Bishop, "Pattern Recognition and Machine Learning")

Outlines

Part XXII

Sequential Data 2

Example of a Hidden Markov Model

Assume Peter and Mary are students in Canberra and Sydney, respectively. Peter is a computer science student and only interested in riding his bicycle, shopping for new computer gadgets, and studying. (Well, he also does other things but because these other activities don't depend on the weather we neglect them here.)

Mary does not know the current weather in Canberra, but knows the general trends of the weather in Canberra. She also knows Peter well enough to know what he does on average every day.

She believes that the weather follows a given discrete Markov chain. She tries to guess the sequence of weather patterns for a number of days after Bob tells her on the phone what he did in the last days.

Example of a Hidden Markov Model

Mary uses the following HMM

		rainy	*sunny*
initial probability		0.2	0.8
transition probability	*rainy*	0.3	0.7
	sunny	0.4	0.6
emission probability	*cycle*	0.1	0.6
	shop	0.4	0.3
	study	0.5	0.1

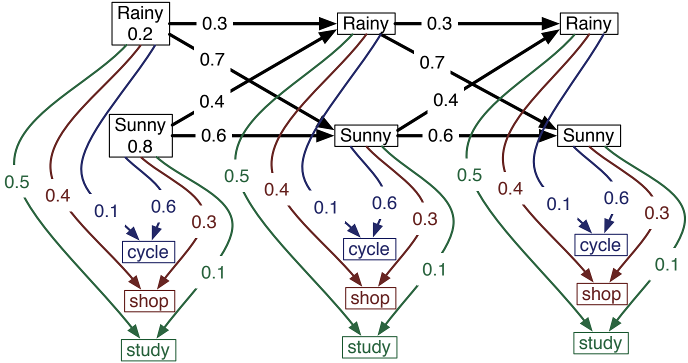
Assume, Peter tells Mary that the list of his activities in the last days was $['cycle', 'shop', 'study']$

(a) Calculate the probability of this observation sequence.

(b) Calculate the most probable sequence of hidden states for these observations.

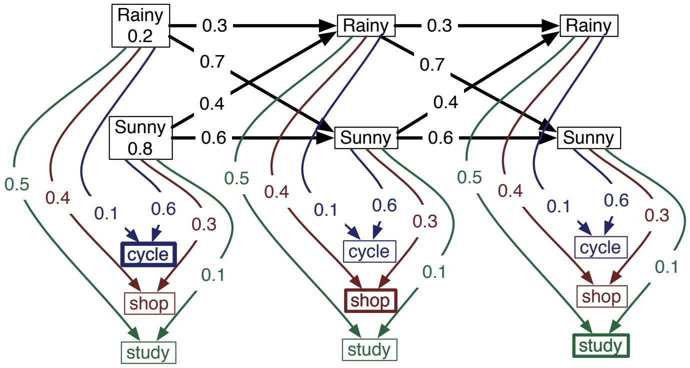
Hidden Markov Model

- Most probable hidden states for 'cycle', 'shop', 'study'



Hidden Markov Model

- Most probable hidden states for 'cycle', 'shop', 'study'



Maximum Likelihood for HMM

- We have observed a data set $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$.
- Assume it came from a HMM with a given structure (number of nodes, form of emission probabilities).
- The likelihood of the data is

$$p(\mathbf{X} \,|\, \boldsymbol{\theta}) = \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z} \,|\, \boldsymbol{\theta})$$

- This joint distribution does not factorise over n (as with the mixture distribution).
- We have N variables, each with K states : K^N terms. Number of terms grows exponentially with the length of the chain.
- But we can use the conditional independence of the latent variables to reorder their calculation later.
- Further obstacle to find a closed loop maximum likelihood solution: calculating the emission probabilities for different states $\mathbf{z}_n$.

Maximum Likelihood for HMM - EM

- Employ the EM algorithm to find the Maximum Likelihood for HMM.

Maximum Likelihood for HMM - EM

- Employ the EM algorithm to find the Maximum Likelihood for HMM.
- Start with some initial parameter settings θ^{old}.

Maximum Likelihood for HMM - EM

- Employ the EM algorithm to find the Maximum Likelihood for HMM.
- Start with some initial parameter settings $\boldsymbol{\theta}^{\text{old}}$.
- E-step: Find the posterior distribution of the latent variables $p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta}^{\text{old}})$.

Maximum Likelihood for HMM - EM

- Employ the EM algorithm to find the Maximum Likelihood for HMM.
- Start with some initial parameter settings $\boldsymbol{\theta}^{\text{old}}$.
- E-step: Find the posterior distribution of the latent variables $p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta}^{\text{old}})$.
- M-step: Maximise

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) = \sum_{\mathbf{Z}} p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta}^{\text{old}}) \ln p(\mathbf{Z}, \mathbf{X} \mid \boldsymbol{\theta})$$

with respect to the parameters $\boldsymbol{\theta} = \{\boldsymbol{\pi}, \mathbf{A}, \phi\}$.

Maximum Likelihood for HMM - EM

- Denote the marginal posterior distribution of $\mathbf{z}_n$ by $\gamma(\mathbf{z}_n)$, and
- the joint posterior distribution of two successive latent variables by $\xi(\mathbf{z}_{n-1}, \mathbf{z}_n)$

$$\gamma(\mathbf{z}_n) = p(\mathbf{z}_n \mid \mathbf{X}, \boldsymbol{\theta}^{\text{old}})$$
$$\xi(\mathbf{z}_{n-1}, \mathbf{z}_n) = p(\mathbf{z}_{n-1}, \mathbf{z}_n \mid \mathbf{X}, \boldsymbol{\theta}^{\text{old}}).$$

- For each step n, $\gamma(\mathbf{z}_n)$ has K nonnegative values which sum to 1.
- For each step n, $\xi(\mathbf{z}_{n-1}, \mathbf{z}_n)$ has $K \times K$ nonnegative values which sum to 1.
- Elements of these vectors are denoted by $\gamma(z_{nk})$ and $\xi(z_{n-1,j}, z_{nk})$ respectively.

Maximum Likelihood for HMM - EM

- Because the expectation of a binary random variable is the probability that it is one, we get with this notation

$$\gamma(z_{nk}) = \mathbb{E}\left[z_{nk}\right] = \sum_{\mathbf{z_n}} \gamma(\mathbf{z_n}) z_{nk}$$

$$\xi(z_{n-1,j}, z_{nk}) = \mathbb{E}\left[z_{n-1,j}, z_{nk}\right] = \sum_{\mathbf{z}_{n-1},\mathbf{z}_n} \gamma(\mathbf{z}) z_{n-1,j}\, z_{nk}.$$

- Putting all together we get

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) = \sum_{k=1}^{K} \gamma(z_{1k}) \ln \pi_k + \sum_{n=2}^{N} \sum_{j=1}^{K} \sum_{k=1}^{K} \xi(z_{n-1,j}, z_{nk}) \ln A_{jk}$$

$$+ \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk}) \ln p(\mathbf{x}_n \mid \phi_k).$$

Maximum Likelihood for HMM - EM

- M-step: Maximising

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) = \sum_{k=1}^{K} \gamma(z_{1k}) \ln \pi_k + \sum_{n=2}^{N} \sum_{j=1}^{K} \sum_{k=1}^{K} \xi(z_{n-1,j}, z_{nk}) \ln A_{jk}$$
$$+ \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk}) \ln p(\mathbf{x}_n \mid \phi_k).$$

- results in

$$\pi_k = \frac{\gamma(z_{1k})}{\sum_{j=1}^{K} \gamma(z_{1j})}$$

$$A_{jk} = \frac{\sum_{n=2}^{N} \xi(z_{n-1,j}, z_{nk})}{\sum_{l=1}^{K} \sum_{n=2}^{N} \xi(z_{n-1,j}, z_{nl})}$$

Maximum Likelihood for HMM - EM

- Still left: Maximising with respect to ϕ.

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) = \sum_{k=1}^{K} \gamma(z_{1k}) \ln \pi_k + \sum_{n=2}^{N} \sum_{j=1}^{K} \sum_{k=1}^{K} \xi(z_{n-1,j}, z_{nk}) \ln A_{jk} + \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk}) \ln p(\mathbf{x}_n \mid \phi_k).$$

- But ϕ only appears in the last term, and under the assumption that all the ϕ_k are independent of each other, this term decouples into a sum.
- Then maximise each contribution $\sum_{n=1}^{N} \gamma(z_{nk}) \ln p(\mathbf{x}_n \mid \phi_k)$ individually.

Maximum Likelihood for HMM - EM

- In the case of Gaussian emission densities

$$p(\mathbf{x} \,|\, \phi_k) = \mathcal{N}(\mathbf{x} \,|\, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k),$$

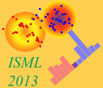
we get for the maximising parameters for the emission densities

$$\boldsymbol{\mu}_k = \frac{\sum_{n=1}^{N} \gamma(z_{nk})\, \mathbf{x}_n}{\sum_{n=1}^{N} \gamma(z_{nk})}$$

$$\boldsymbol{\Sigma}_k = \frac{\sum_{n=1}^{N} \gamma(z_{nk})\, (\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T}{\sum_{n=1}^{N} \gamma(z_{nk})}.$$

Forward-Backward HMM

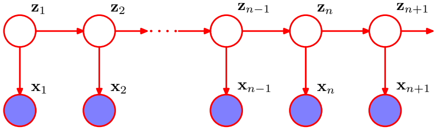
- Need to efficiently evaluate the $\gamma(z_{nk})$ and $\xi(z_{n-1,j}, z_{nk})$.
- The graphical model for the HMM is a tree!
- We know we can use a two-stage message passing algorithm to calculate the posterior distribution of the latent variables.
- For HMM this is called the forward-backward algorithm (Rabiner, 1989), or Baum-Welch algorithm (Baum, 1972).
- Other variants exist, different only in the form of the messages propagated.
- We look at the most widely known, the alpha-beta algorithm.



Conditional Independence for HMM

Given the data $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ the following independence relations hold:

$$
\begin{aligned}
p(\mathbf{X} \mid \mathbf{z}_n) &= p(\mathbf{x}_1, \dots, \mathbf{x}_n \mid \mathbf{z}_n)\, p(\mathbf{x}_{n+1}, \dots, \mathbf{x}_N \mid \mathbf{z}_n) \\
p(\mathbf{x}_1, \dots, \mathbf{x}_{n-1} \mid \mathbf{x}_n, \mathbf{z}_n) &= p(\mathbf{x}_1, \dots, \mathbf{x}_{n-1} \mid \mathbf{z}_n) \\
p(\mathbf{x}_1, \dots, \mathbf{x}_{n-1} \mid \mathbf{z}_{n-1}, \mathbf{z}_n) &= p(\mathbf{x}_1, \dots, \mathbf{x}_{n-1} \mid \mathbf{z}_{n-1}) \\
p(\mathbf{x}_{n+1}, \dots, \mathbf{x}_N \mid \mathbf{z}_n, \mathbf{z}_{n+1}) &= p(\mathbf{x}_{n+1}, \dots, \mathbf{x}_N \mid \mathbf{z}_{n+1}) \\
p(\mathbf{x}_{n+2}, \dots, \mathbf{x}_N \mid \mathbf{x}_{n+1}, \mathbf{z}_{n+1}) &= p(\mathbf{x}_{n+2}, \dots, \mathbf{x}_N \mid \mathbf{z}_{n+1}) \\
p(\mathbf{X} \mid \mathbf{z}_{n-1}, \mathbf{z}_n) &= p(\mathbf{x}_1, \dots, \mathbf{x}_{n-1} \mid \mathbf{z}_{n-1})\, p(\mathbf{x}_n \mid \mathbf{x}_n) \\
&\quad p(\mathbf{x}_{n+1}, \dots, \mathbf{x}_N \mid \mathbf{z}_n) \\
p(\mathbf{x}_{N+1} \mid \mathbf{X}, \mathbf{z}_{N+1}) &= p(\mathbf{x}_{N+1} \mid \mathbf{z}_{N+1}) \\
p(\mathbf{z}_{n+1} \mid \mathbf{X}, \mathbf{z}_N) &= p(\mathbf{z}_{N+1} \mid \mathbf{z}_N)
\end{aligned}
$$

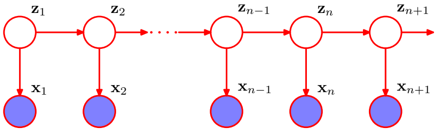


Conditional Independence for HMM - Example

- Let's look at the following independence relation:

$$p(\mathbf{X} \,|\, \mathbf{z}_n) = p(\mathbf{x}_1, \dots, \mathbf{x}_n \,|\, \mathbf{z}_n)\, p(\mathbf{x}_{n+1}, \dots, \mathbf{x}_N \,|\, \mathbf{z}_n)$$

- Any path from the set $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ to the set $\{\mathbf{x}_{n+1}, \dots, \mathbf{x}_N\}$ has to go through $\mathbf{z}_n$.
- In $p(\mathbf{X} \,|\, \mathbf{z}_n)$ the node $\mathbf{z}_n$ is conditioned on ($=$ observed).
- All paths through $\mathbf{z}_n$ are head-tail.
- Therefore all paths through $\mathbf{z}_n$ are blocked.



Alpha-Beta HMM

- Define the joint probability of observing all data up to step n and having $\mathbf{z}_n$ as latent variable to be

$$\alpha(\mathbf{z}_n) = p(\mathbf{x}_1, \dots, \mathbf{x}_n, \mathbf{z}_n).$$

- Define the probability of all future data given $\mathbf{z}_n$ to be

$$\beta(\mathbf{z}_n) = p(\mathbf{x}_{n+1}, \dots, \mathbf{x}_N \mid \mathbf{z}_n).$$

- Then it an be shown the following recursions hold

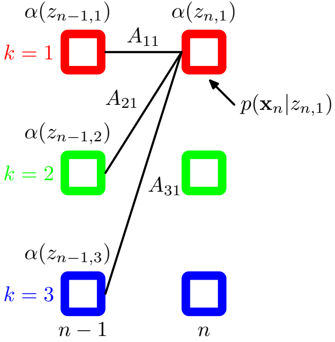
$$\alpha(\mathbf{z}_n) = p(\mathbf{x}_n \mid \mathbf{z}_n) \sum_{\mathbf{z}_{n-1}} \alpha(\mathbf{z}_{n-1})\, p(\mathbf{z}_n \mid \mathbf{z}_{n-1})$$

$$\alpha(\mathbf{z}_1) = \prod_{k=1}^{K} \{\pi_k\, p(\mathbf{x}_1 \mid \phi_k)\}^{z_{1k}}$$

Alpha-Beta HMM

- At step n we can efficiently calculate $\alpha(\mathbf{z}_n)$ given $\alpha(\mathbf{z}_{n-1})$

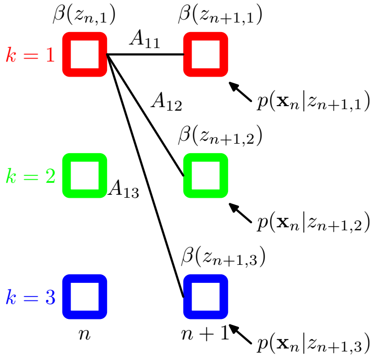
$$\alpha(\mathbf{z}_n) = p(\mathbf{x}_n \mid \mathbf{z}_n) \sum_{\mathbf{z}_{n-1}} \alpha(\mathbf{z}_{n-1})\, p(\mathbf{z}_n \mid \mathbf{z}_{n-1})$$



Alpha-Beta HMM

- And for $\beta(\mathbf{z}_n)$ we get the recursion

$$\beta(\mathbf{z}_n) = \sum_{\mathbf{z}_{n+1}} \beta(\mathbf{z}_{n+1})\, p(\mathbf{x}_{n+1} \mid \mathbf{z}_{n+1})\, p(\mathbf{z}_{n+1} \mid \mathbf{z}_n)$$



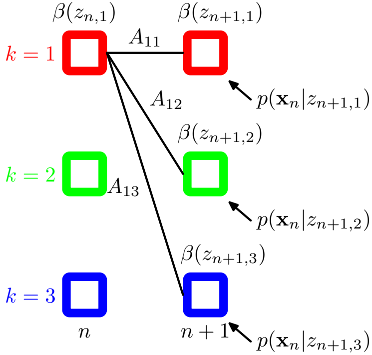
Alpha-Beta HMM

- How do we start the β recursion? What is $\beta(\mathbf{z}_N)$?

$$\beta(\mathbf{z}_N) = p(\mathbf{x}_{N+1}, \ldots, \mathbf{x}_N \mid \mathbf{z}_n).$$

- Can be shown the following is consistent with the approach

$$\beta(\mathbf{z}_N) = 1.$$



Alpha-Beta HMM

- Now we know how to calculate $\alpha(\mathbf{z}_n)$ and $\beta(\mathbf{z}_n)$ for each step.
- What is the probability of the data $p(\mathbf{X})$?
- Use the definition of $\gamma(\mathbf{z}_n)$ and Bayes

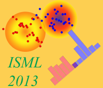
$$\gamma(\mathbf{z}_n) = p(\mathbf{z}_n \mid \mathbf{X}) = \frac{p(\mathbf{X} \mid \mathbf{z}_n)\, p(\mathbf{z}_n)}{p(\mathbf{X})} = \frac{p(\mathbf{X}, \mathbf{z}_n)}{p(\mathbf{X})}$$

- and the following conditional independence statement from the graphical model of the HMM

$$p(\mathbf{X} \mid \mathbf{z}_n) = \underbrace{p(\mathbf{x}_1, \dots, \mathbf{x}_n \mid \mathbf{z}_n)}_{\alpha(\mathbf{z}_n)/p(\mathbf{z}_n)} \underbrace{p(\mathbf{x}_{n+1}, \dots, \mathbf{x}_N \mid \mathbf{z}_n)}_{\beta(\mathbf{z}_n)}$$

- and therefore

$$\gamma(\mathbf{z}_n) = \frac{\alpha(\mathbf{z}_n)\beta(\mathbf{z}_n)}{p(\mathbf{X})}$$



Alpha-Beta HMM

- Marginalising over $\mathbf{z}_n$ results in

$$1 = \sum_{\mathbf{z}_n} \gamma(\mathbf{z}_n) = \frac{\sum_{\mathbf{z}_n} \alpha(\mathbf{z}_n)\beta(\mathbf{z}_n)}{p(\mathbf{X})}$$

  and therefore at each step n

$$p(\mathbf{X}) = \sum_{\mathbf{z}_n} \alpha(\mathbf{z}_n)\beta(\mathbf{z}_n)$$

- Most conviniently evaluated at step N where $\beta(\mathbf{z}_N) = 1$ as

$$p(\mathbf{X}) = \sum_{\mathbf{z}_N} \alpha(\mathbf{z}_n).$$

Alpha-Beta HMM

- Finally, we need to calculate the joint posterior distribution of two successive latent variables by $\xi(\mathbf{z}_{n-1}, \mathbf{z}_n)$ defined by

$$\xi(\mathbf{z}_{n-1}, \mathbf{z}_n) = p(\mathbf{z}_{n-1}, \mathbf{z}_n \mid \mathbf{X}).$$

- This can be done directly from the α and β values in the form

$$\xi(\mathbf{z}_{n-1}, \mathbf{z}_n) = \frac{\alpha(\mathbf{z}_{n-1})\, p(\mathbf{x}_n \mid \mathbf{z}_n)\, p(\mathbf{z}_n \mid \mathbf{z}_{n-1})\, \beta(\mathbf{z}_n)}{p(\mathbf{X})}$$

How to train a HMM using EM?

1. Make an inital selection for the parameters $\boldsymbol{\theta}^{\text{old}}$ where $\boldsymbol{\theta} = \{\boldsymbol{\pi}, \mathbf{A}, \phi)\}$. (Often, $\mathbf{A}, \boldsymbol{\pi}$ initialised uniformly or randomly. The ϕ_k-initialisation depends on the emission distribution; for Gaussians run K-means first and get $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ from there.)
2. (Start of E-step) Run forward recursion to calculate $\alpha(\mathbf{z}_n)$.
3. Run backward recursion to calculate $\beta(\mathbf{z}_n)$.
4. Calculate $\gamma(\mathbf{z})$ and $\xi(\mathbf{z}_{n-1}, \mathbf{z}_n)$ from $\alpha(\mathbf{z}_n)$ and $\beta(\mathbf{z}_n)$.
5. Evaluate the likelihood $p(\mathbf{Z})$.
6. (Start of M-step) Find a $\boldsymbol{\theta}^{\text{new}}$ maximising $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}})$. This results in new settings for the parameters π_k, A_jk and ϕ_k as described before.
7. Iterate until convergence is detected.

Alpha-Beta HMM - Notes

- In order to calculate the likelihood, we need to use the joint probability $p(\mathbf{X}, \mathbf{Z})$ and sum over all possible values of $\mathbf{Z}$.
- That means, every particular choice of $\mathbf{Z}$ corresponds to one path through the lattice diagram. There are exponentially many !
- Using the alpha-beta algorithm, the exponential cost has been reduced to a linear cost in the length of the model.
- How did we do that?
- Swapping the order of multiplication and summation.

HMM - Viterbi Algorithm

- Motivation: The latent states can have some meaningful interpretation, e.g. phonemes in a speech recognition system where the observed variables are the acoustic signals.
- Goal: After the system has been trained, find the most probable states of the latent variables for a given sequence of observations.
- Warning: Finding the set of states which are each indivdually the most probable does NOT solve this problem.

HMM - Viterbi Algorithm

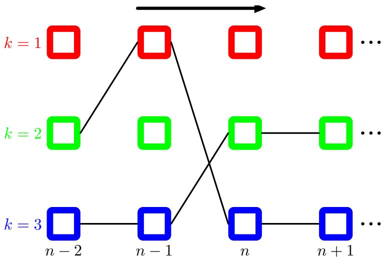
- Define $\omega(\mathbf{z}_n) = \max_{\mathbf{z}_1,\dots,\mathbf{z}_{n-1}} \ln p(\mathbf{x}_1,\dots,\mathbf{x}_n,\mathbf{z}_1,\dots,\mathbf{z}_n)$
- From the joint distribution of the HMM given by

$$p(\mathbf{x}_1,\dots,\mathbf{x}_N,\mathbf{z}_1,\dots,\mathbf{z}_N) = p(\mathbf{z}_1)\left[\prod_{n=2}^{N} p(\mathbf{z}_n \mid \mathbf{z}_{n-1})\right]\prod_{n=1}^{N} p(\mathbf{x}_n \mid \mathbf{z}_n)$$

the following recursion can be derived

$$\omega(\mathbf{z}_n) = \ln p(\mathbf{x}_n \mid \mathbf{z}_n) + \max_{\mathbf{z}_{n-1}}\{\ln p(\mathbf{z}_n \mid \mathbf{z}_{n-1}) + \omega(\mathbf{z}_{n-1})\}$$

$$\omega(\mathbf{z}_1) = \ln p(\mathbf{z}_1) + \ln p(\mathbf{x}_1 \mid \mathbf{z}_1) = \ln p(\mathbf{x}_1, \mathbf{z}_1)$$



HMM - Viterbi Algorithm

- Calculate

$$\omega(\mathbf{z}_n) = \max_{\mathbf{z}_1,\dots,\mathbf{z}_{n-1}} \ln p(\mathbf{x}_1,\dots,\mathbf{x}_n,\mathbf{z}_1,\dots,\mathbf{z}_n)$$

  for $n = 1, \dots, N$.
- For each step n remember which is the best transition to go into each state at the next step.
- At step N : Find the state with the highest probability.
- For $n = 1, \dots, N-1$: Backtrace which transition led to the most probable state and identifiy from which state it came.