

Hidden markov models

Course of Machine Learning
Master Degree in Computer Science
University of Rome "Tor Vergata"

Giorgio Gambosi

a.a. 2017-2018

Sequence of observations (usually in time) $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$.

Joint probability

$$p(\mathbf{x}_1, \dots, \mathbf{x}_n) = \prod_{i=1}^n p(\mathbf{x}_i | \mathbf{x}_{i-1}, \dots, \mathbf{x}_1)$$

From past observations we wish to predict the next value \mathbf{x}_n .

The conditional probability distribution $p(\mathbf{x}_n | \mathbf{x}_{n-1}, \dots, \mathbf{x}_1)$ is considered.

This would result into an exponential number of cases to be considered. The set of observations is sparse, making it impossible to infer all probabilities.

Weather predictions

Assume there exist only three possible weather conditions: sunny (☀️), rainy (🌧️), foggy (🌫️).

We wish to predict tomorrow weather from the knowledge of the past weather (time series).

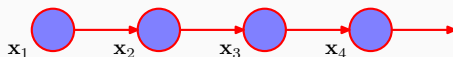
Let q_n be the weather at day n : in general, we may assume that q_n depends on $q_{n-1}, q_{n-2}, \dots, q_1$.

Markov models

The next value is assumed dependent only on the k latest observations (k -th order Markov model).

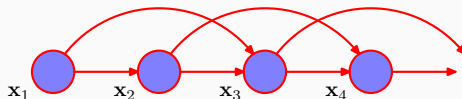
First order Markov models

$$p(\mathbf{x}_n | \mathbf{x}_{n-1}, \dots, \mathbf{x}_1) = p(\mathbf{x}_n | \mathbf{x}_{n-1})$$



Second order Markov models

$$p(\mathbf{x}_n | \mathbf{x}_{n-1}, \dots, \mathbf{x}_1) = p(\mathbf{x}_n | \mathbf{x}_{n-1}, \mathbf{x}_{n-2})$$



We wish to derive the conditional probability distribution $P(q_n | q_{n-1}, \dots, q_1)$
(e.g. $P(\text{☁️} | \text{🌧️}, \text{☀️}, \text{☀️})$).

Problem: As n increases, the number of possible time series increases exponentially
(i.e. 3 observations for 6 days = $3^{6-1} = 243$ possible time series)







Weather predictions

Assume a **First order Markov Model**.

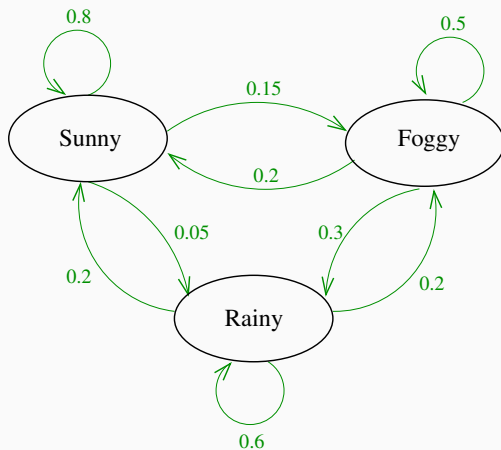
$$P(q_n | q_{n-1}, \dots, q_1) = P(q_n | q_{n-1})$$

This allows to consider 9 time series out of the original 243.

The output sequence (q_i) is a (homogeneous) **First order Markov Chain**.

Today	Tomorrow		
			
	0.8	0.05	0.15
	0.2	0.6	0.2
	0.2	0.3	0.5

Weather predictions



Since today it is $q_1 = \text{☀}$, what is the probability that tomorrow it will be $q_2 = \text{☀}$ and the day after tomorrow it will be $q_3 = \text{☁}$?

$$\begin{aligned} P(q_2 = \text{☀}, q_3 = \text{☁} | q_1 = \text{☀}) &= \\ P(q_3 = \text{☁} | q_2 = \text{☀}, q_1 = \text{☀}) \cdot P(q_2 = \text{☀} | q_1 = \text{☀}) &= \\ P(q_3 = \text{☁} | q_2 = \text{☀}) \cdot P(q_2 = \text{☀} | q_1 = \text{☀}) &= 0.05 \cdot 0.8 = 0.04 \end{aligned}$$

Since yesterday it was $q_1 = \text{☁}$ and today it is $q_2 = \text{☁}$, what is the probability that tomorrow it will be $q_3 = \text{☀}$?

$$P(q_3 = \text{☀} | q_2 = \text{☁}, q_1 = \text{☁}) =$$

$$P(q_3 = \text{☀} | q_2 = \text{☁}) = 0.2$$

$$p(\mathbf{x}_n) = q(\mathbf{x}|\boldsymbol{\theta}_n)$$

where q is some predefined distribution with parameters $\boldsymbol{\theta}$.

$\boldsymbol{\theta}_i$ is a function of $\mathbf{x}_{n-1}, \dots, \mathbf{x}_1$

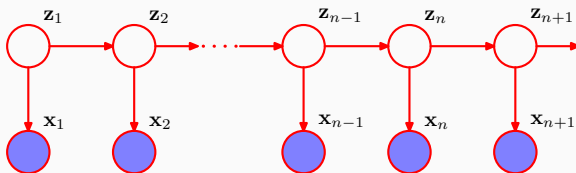
Autoregressive (AR) model: q is the gaussian distribution, $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is a linear function of $\mathbf{x}_{n-1}, \dots, \mathbf{x}_1$.

Latent variable models

Assume we wish to predict the next value by not strictly considering the latest values observed, but by using a limited number of parameters.

Introduce a latent variable:

- the observations depends on the corresponding values of the latent variable
- latent variables values are related through a Markov model



$$p(\mathbf{x}_1, \dots, \mathbf{x}_n, \mathbf{z}_1, \dots, \mathbf{z}_n) = \prod_{i=1}^n p(\mathbf{x}_i | \mathbf{z}_i) \prod_{i=2}^n p(\mathbf{z}_i | \mathbf{z}_{i-1}) p(\mathbf{z}_1)$$

Discrete latent variables: Hidden Markov Models

Gaussian latent and observed variables: linear dynamical systems

Undirect weather prediction

We are not able to directly observe weather conditions. However, we may observe whether our mother gets out carrying an umbrella(☂️) or not (☂️🚫).

Weather is **hidden**.

The only evidence we get about the weather is the umbrella.

- Weather is the latent variable
- Umbrella is the observed variable

Undirect weather prediction

The relation between weather and umbrella can be expressed through the conditional probability distribution $p(x_i|q_i)$, where $q_i \in \{\text{☀️}, \text{☁️}, \text{☔️}\}$ and $x_i \in \{\text{☔️}, \text{☔️}\}$.

By Bayes rule, it is possible to derive the probability distribution of the latent variable given the observation






$$p(q_i|x_i) = \frac{p(x_i|q_i)p(q_i)}{p(x_i)}$$

If a sequence of observations is considered, by the model assumptions,:







$$\begin{aligned} p(q_1, \dots, q_n | x_1, \dots, x_n) &= \frac{p(x_1, \dots, x_n | q_1, \dots, q_n) p(q_1, \dots, q_n)}{p(x_1, \dots, x_n)} \\ &\propto \prod_{i=1}^n p(x_i | q_i) \prod_{i=2}^n p(q_i | q_{i-1}) p(q_1) \end{aligned}$$

Undirect weather prediction

Conditional probability of umbrella x_i , given the weather condition q_i

Tempo		
	0.1	0.9
	0.8	0.2
	0.3	0.7

Conditional probability of weather condition on a day q_i given the weather condition on the previous day q_{i-1} .

Today	Tomorrow		
			
	0.8	0.05	0.15
	0.2	0.6	0.2
	0.2	0.3	0.5

Prior probability of weather condition

Hidden Markov Model

- $S = \{s_1, \dots, s_{N_s}\}$ set of states
- $\Theta = \{\pi, \mathbf{A}, \mathbf{B}\}$:
 - π : prior probabilities $\pi_i = q(q_1 = s_i)$
 - \mathbf{A} transition probabilities matrix $a_{ij} = p(q_{n+1} = s_j | q_n = s_i)$ (independent of n by the homogeneity assumption)
 - \mathbf{B} emission probabilities:
 - discrete observations, e.g. $x_n \in \{v_1, \dots, v_k\}$: \mathbf{B} is a matrix $b_{ik} = p(x_n = v_k | q_n = s_i)$.
 - continuous observations, e.g. $x_n \in \mathbb{R}^d$, $\mathbf{B}(x)$ is a vector of probability densities $b_i(x_n) = p(x_n | q_n = s_i)$ (actually, parameters).

Random variables in a Hidden Markov Model

- A hidden state sequence $Q = \{q_1, q_2, \dots, q_n\}$, $q_i \in S$.
- An observation sequence $X = \{x_1, x_2, \dots, x_n\}$.

Assume the last weather observation was performed the day before yesterday (it was ☀️). Yesterday your mother got out carrying an umbrella. Which was the weather?

$$\begin{aligned} L(q_2 = \text{☀️} | q_1 = \text{☀️}, x_2 = \text{☂️}) &= P(x_2 = \text{☂️} | q_2 = \text{☀️}) \cdot P(q_2 = \text{☀️} | q_1 = \text{☀️}) \\ &= 0.1 \cdot 0.8 = 0.08 \end{aligned}$$

$$\begin{aligned} L(q_2 = \text{☁️} | q_1 = \text{☀️}, x_2 = \text{☂️}) &= P(x_2 = \text{☂️} | q_2 = \text{☁️}) \cdot P(q_2 = \text{☁️} | q_1 = \text{☀️}) \\ &= 0.8 \cdot 0.05 = 0.04 \end{aligned}$$

$$\begin{aligned} L(q_2 = \text{🌧️} | q_1 = \text{☀️}, x_2 = \text{☂️}) &= P(x_2 = \text{☂️} | q_2 = \text{🌧️}) \cdot P(q_2 = \text{🌧️} | q_1 = \text{☀️}) \\ &= 0.3 \cdot 0.15 = 0.045 \end{aligned}$$

Assume you do not remember the last weather observed, three days ago. In the following three days your mother always carried no umbrella. What is the probability that in those days the weather was

$(q_1 = \text{☀}, q_2 = \text{☁}, q_3 = \text{☀})$?

$$\begin{aligned} L(q_1 = \text{☀}, q_2 = \text{☁}, q_3 = \text{☀} | x_1 = \text{☂}, x_2 = \text{☂}, x_3 = \text{☂}) = \\ P(x_1 = \text{☂} | q_1 = \text{☀}) \cdot P(x_2 = \text{☂} | q_2 = \text{☁}) \cdot P(x_3 = \text{☂} | q_3 = \text{☀}) \cdot \\ P(q_1 = \text{☀}) \cdot P(q_2 = \text{☁} | q_1 = \text{☀}) \cdot P(q_3 = \text{☀} | q_2 = \text{☁}) = \\ 0.9 \cdot 0.7 \cdot 0.9 \cdot 1/3 \cdot 0.15 \cdot 0.2 = 0.0057 \end{aligned}$$

State sequence distribution

$$p(Z|\Theta) = \pi_{z_1} \prod_{i=1}^{n-1} a_{z_i, z_{i+1}}$$

Conditional observation sequence distribution

$$p(X|Z, \Theta) = \prod_{i=1}^{n-1} p(x_i|z_i, \Theta) = \prod_{i=1}^{n-1} b_{z_i, x_i}$$

Joint distribution

$$p(X, Z|\Theta) = p(X|Z, \Theta)p(Z|\Theta) = \pi_{z_1} \prod_{i=1}^{n-1} a_{z_i, z_{i+1}} \prod_{i=1}^n b_{z_i, x_i}$$

Likelihood of an observation sequence

$$p(X|\Theta) = \sum_{\text{all } Z} p(X, Z|\Theta)$$

Given an *observation sequence* $X = (\mathbf{x}_1, \dots, \mathbf{x}_n)$, it is possible to consider three different objectives:

- computing the probability $p(\mathbf{X}|\Theta)$ of the sequence, given the parameters
- estimating the set of parameter values: this can be performed by a suitable application of the EM algorithm (Baum-Welch algorithm)
- finding the most likely sequence of hidden states $\mathbf{z}_1, \dots, \mathbf{z}_n$. It can be shown that this is not equivalent to maximizing $p(\mathbf{z}_i|\mathbf{X}, \Theta)$ for each latent variable, which could be performed after the previous problem is solved. The Viterbi algorithm can be applied here.

Forward probabilities

Given an observed sequence $X = (x_1, \dots, x_n)$, let us define the following probabilities.

$$\alpha_t(z) = p(x_1, x_2, \dots, x_t, z_t = z | \Theta)$$

the probability of observing the first t values of \mathbf{X} and that the hidden value, at the end of such sequence, is z . The α values enable us to solve the first problem since, marginalizing z , we obtain

$$p(X | \Theta) = \sum_{\text{all } z} p(x_1, x_2, \dots, x_n, z_n = z | \Theta) = \sum_{\text{all } z} \alpha_n(z)$$

The α values can be derived by the following recursive definition

$$\begin{aligned}\alpha_1(z_i) &= \pi_i b_{z_i, x_1} \\ \alpha_{t+1}(z_i) &= \sum_{\text{all } z_j} \alpha_t(z_j) a_{z_j, z_i} b_{z_i, x_{t+1}}\end{aligned}$$

The following (backward) probabilities can also be defined

$$\beta_t(z) = p(x_{t+1}, \dots, x_n | z_t = z, \Theta)$$

that, is, the probability that, if the current hidden value is z , the following $n - t$ observed values correspond to the last values of \mathbf{X} .

The β values can also be derived by the following recursive definition

$$\begin{aligned}\beta_n(z_i) &= 1 \\ \beta_t(z_i) &= \sum_{\text{all } z_j} a_{z_i, z_j} b_{z_j, x_{t+1}} \beta_{t+1}(z_j)\end{aligned}$$

Observe that

$$p(X, z_t = z | \Theta) = p(x_1, x_2, \dots, x_t, z_t = z | \Theta) p(x_{t+1}, \dots, x_n | z_t = z, \Theta) = \alpha_t(z) \beta_t(z)$$

Useful probabilities

The probability of being in state z at time t given the observation sequence and model.

$$\gamma_t(z) = p(z_t = z | X, \Theta) = \frac{p(X, z_t = z | \Theta)}{p(X | \Theta)} = \frac{\alpha_t(z)\beta_t(z)}{p(X | \Theta)}$$

The probability of being in state z at time t and making the transition from z to z' given the observation sequence and model.

$$\xi_t(z, z') = p(z_t = z, z_{t+1} = z' | X, \Theta)$$

Since

$$p(z_t = z, z_{t+1} = z', X | \Theta) = \alpha_t(z) a_{z, z'} b_{z', x_{t+1}} \beta_{t+1}(z')$$

it derives that

$$\xi_t(z, z') = \frac{p(z_t = z, z_{t+1} = z', X | \Theta)}{p(X | \Theta)} = \frac{\alpha_t(z) a_{z, z'} b_{z', x_{t+1}} \beta_{t+1}(z')}{p(X | \Theta)}$$

Moreover,

$$\gamma_t(z) = p(z_t = z | X, \Theta) = \sum_{\text{all } z'} p(z_t = z, z_{t+1} = z' | X, \Theta) = \sum_{\text{all } z'} \xi_t(z, z')$$

It is an application of EM to HMM.

- $\sum_{t=1}^n \xi_t(z, z')$ is the expected number of transitions from z_i to z_j
- $\sum_{t=1}^n \gamma_t(z)$ is the expected number of transitions from z_i

They both depend on $\Theta = \{\mathbf{A}, \mathbf{B}, \pi\}$ (E-step).

However, the parameter values can be derived from $\gamma_t(z)$ and $\xi_t(z, z')$ as follows:

$$a_{z_i, z_j} = \frac{\sum_{k=1}^{n-1} \xi_t(z_i, z_j)}{\sum_{k=1}^{n-1} \gamma_t(z_j)}$$
$$b_{z_i, x_j} = \frac{\sum_{k=1}^{n-1} \gamma_t(z_i) \phi_t(x_j)}{\sum_{k=1}^{n-1} \gamma_t(z_j)}$$
$$\pi_{z_i} = \gamma_1(z_i)$$

where $\phi_t(x) = 1$ if $x_t = x$, 0 otherwise.

Viterbi Algorithm

The problem of deriving the most likely sequence of hidden states, given the sequence of observed states, can be solved by means of the **Viterbi** algorithm.

The algorithm uses two variables

- $\delta_i(k)$, the *maximum* joint likelihood of a sequence of latent states (and the observed values) among all sequences q_1, \dots, q_i with $q_i = s_k$

$$\delta_i(k) = \max_{q_1, q_2, \dots, q_{i-1}} p(q_1, q_2, \dots, q_{i-1}, q_i = s_k, x_1, x_2, \dots, x_i | \Theta)$$

- the corresponding sequence

$$(q_1^{(i)}, q_2^{(i)}, \dots, q_{i-1}^{(i)}) = \operatorname{argmax}_{q_1, q_2, \dots, q_{i-1}} p(q_1, q_2, \dots, q_{i-1}, q_i = s_k, x_1, x_2, \dots, x_i | \Theta)$$

is represented by setting

$$\psi_1(k) = \perp$$

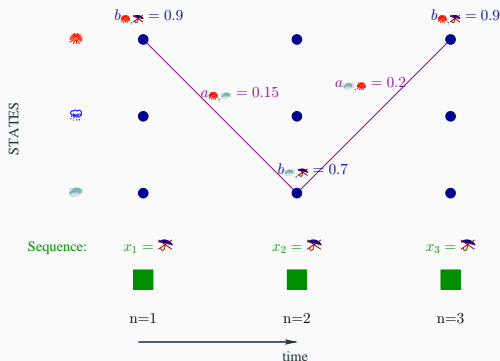
$$\psi_i(k) = q_{i-1}^{(i)} \quad i > 1$$

The idea of the algorithm is to apply dynamic programming to find the most likely sequence for each intermediate state of the Trellis diagram, up to the final state. For each, intermediate state, only the sequence providing the

Trellis Diagram

A Trellis Diagram is used to visualize sequences of states. Rows correspond to states, columns to steps.

The observed sequence $X = (x_1, \dots, x_n)$ is also visualized.



$$\begin{aligned}
 L &= \pi_{\text{Sun}} \cdot b_{\text{Sun}, \text{Sun}} \cdot a_{\text{Sun}, \text{Bug}} \cdot b_{\text{Bug}, \text{Sun}} \cdot a_{\text{Bug}, \text{Sun}} \cdot b_{\text{Sun}, \text{Sun}} \\
 &= 1/3 \cdot 0.9 \cdot 0.15 \cdot 0.7 \cdot 0.2 \cdot 0.9 = 0.0057
 \end{aligned}$$

This is the case $i = 1$:

$$\begin{aligned}\delta_1(k) &= \pi_k b_{k,x_1} \quad k = 1, \dots, n_s \\ \psi_1(k) &= 0\end{aligned}$$

where π_k is the initial probability of state s_k .

This is for all $2 \leq i \leq n$ and $1 \leq k \leq n_s$:

$$\delta_i(k) = \max_{1 \leq j \leq n_s} (\delta_{i-1}(j) a_{kj}) b_{k,x_i}$$

$$\psi_i(k) = \operatorname{argmax}_{1 \leq j \leq n_s} (\delta_{i-1}(j) a_{kj})$$

$$p^*(X, Q | \Theta) = \max_{1 \leq i \leq n_s} \delta_n(i)$$
$$q_n^* = \operatorname{argmax}_{1 \leq i \leq n_s} \delta_n(i)$$

The sequence $Q^* = \{q_1^*, \dots, q_n^*\}$ of hidden states can be derived as

$$q_n^* = q_n^{(n)}$$

$$q_i^* = \psi_{i+1}(q_{i+1}^*) \quad i < n$$

Let $n = 3$ and let $Q = \{\text{☔}, \text{☀}, \text{☔}\}$: we wish to find the most likely sequence of weather conditions in the three days.

Initialization: $i = 1$:

$$\delta_1(\text{☀}) = \pi_{\text{☀}} \cdot b_{\text{☀}, \text{☔}} = 1/3 \cdot 0.9 = 0.3$$

$$\psi_1(\text{☀}) = 0$$

$$\delta_1(\text{☔}) = \pi_{\text{☔}} \cdot b_{\text{☔}, \text{☔}} = 1/3 \cdot 0.2 = 0.0667$$

$$\psi_1(\text{☔}) = 0$$

$$\delta_1(\text{☁}) = \pi_{\text{☁}} \cdot b_{\text{☁}, \text{☔}} = 1/3 \cdot 0.7 = 0.233$$

$$\psi_1(\text{☁}) = 0$$

Recursion, 1st step:

$$\begin{aligned}\delta_2(\text{☀}) &= \max(\delta_1(\text{☀})a_{\text{☀},\text{☀}}, \delta_1(\text{☁})a_{\text{☁},\text{☀}}, \delta_1(\text{🌧})a_{\text{🌧},\text{☀}}) \cdot b_{\text{☀},\text{☔}} \\ &= \max(0.3 \cdot 0.8, 0.0667 \cdot 0.2, 0.233 \cdot 0.2) \cdot 0.1 = 0.024\end{aligned}$$

$$\psi_2(\text{☀}) = \text{☀}$$

$$\begin{aligned}\delta_2(\text{☁}) &= \max(\delta_1(\text{☀}) \cdot a_{\text{☀},\text{☁}}, \delta_1(\text{☁}) \cdot a_{\text{☁},\text{☁}}, \delta_1(\text{🌧}) \cdot a_{\text{🌧},\text{☁}}) \cdot b_{\text{☁},\text{☔}} \\ &= \max(0.3 \cdot 0.05, 0.0667 \cdot 0.6, 0.233 \cdot 0.3) \cdot 0.8 = 0.056\end{aligned}$$

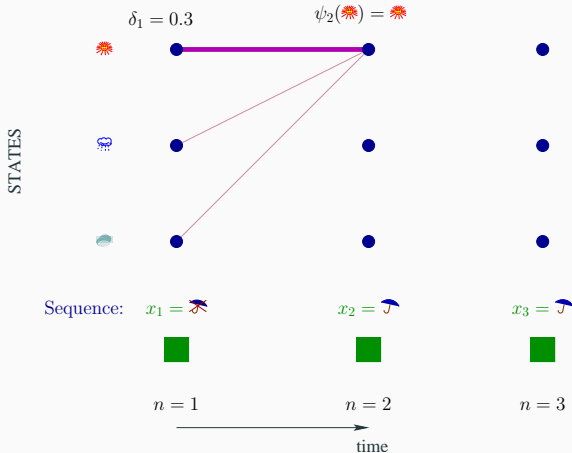
$$\psi_2(\text{☁}) = \text{🌧}$$

$$\begin{aligned}\delta_2(\text{🌧}) &= \max(\delta_1(\text{☀}) \cdot a_{\text{☀},\text{🌧}}, \delta_1(\text{☁}) \cdot a_{\text{☁},\text{🌧}}, \delta_1(\text{🌧}) \cdot a_{\text{🌧},\text{🌧}}) \cdot b_{\text{🌧},\text{☔}} \\ &= \max(0.3 \cdot 0.15, 0.0667 \cdot 0.2, 0.233 \cdot 0.5) \cdot 0.3 = 0.0350\end{aligned}$$

$$\psi_2(\text{🌧}) = \text{🌧}$$

Viterbi and weather prediction

$$\delta_2(\text{☀}) = \max(\delta_1(\text{☀}) \cdot a_{\text{☀}, \text{☀}}, \delta_1(\text{☁}) \cdot a_{\text{☁}, \text{☀}}, \delta_1(\text{🌧}) \cdot a_{\text{🌧}, \text{☀}} \cdot b_{\text{☀}, \text{☔}})$$



Recursion, 2nd step:

$$\begin{aligned}\delta_3(\text{☀}) &= \max(\delta_2(\text{☀}) \cdot a_{\text{☀},\text{☀}}, \delta_2(\text{☁}) \cdot a_{\text{☁},\text{☀}}, \delta_2(\text{☂}) \cdot a_{\text{☂},\text{☀}}) \cdot b_{\text{☀},\text{☂}} \\ &= \max(0.024 \cdot 0.8, 0.056 \cdot 0.2, 0.035 \cdot 0.2) \cdot 0.1 = 0.0019\end{aligned}$$

$$\psi_3(\text{☀}) = \text{☀}$$

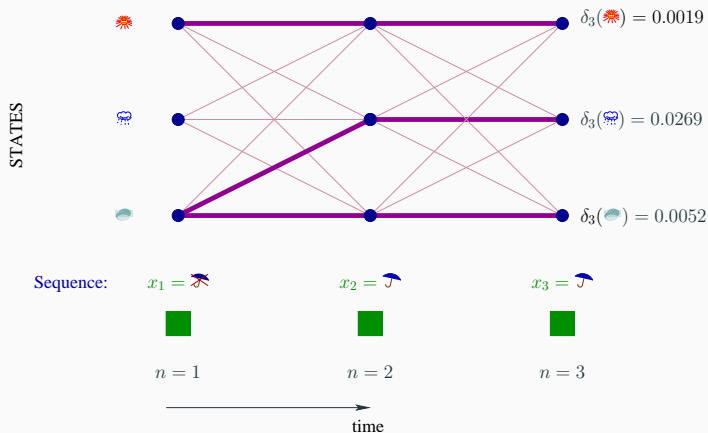
$$\begin{aligned}\delta_3(\text{☁}) &= \max(\delta_2(\text{☀}) \cdot a_{\text{☀},\text{☁}}, \delta_2(\text{☁}) \cdot a_{\text{☁},\text{☁}}, \delta_2(\text{☂}) \cdot a_{\text{☂},\text{☁}}) \cdot b_{\text{☁},\text{☂}} \\ &= \max(0.024 \cdot 0.05, 0.056 \cdot 0.6, 0.035 \cdot 0.3) \cdot 0.8 = 0.0269\end{aligned}$$

$$\psi_3(\text{☁}) = \text{☁}$$

$$\begin{aligned}\delta_3(\text{☂}) &= \max(\delta_2(\text{☀}) \cdot a_{\text{☀},\text{☂}}, \delta_2(\text{☁}) \cdot a_{\text{☁},\text{☂}}, \delta_2(\text{☂}) \cdot a_{\text{☂},\text{☂}}) \cdot b_{\text{☂},\text{☂}} \\ &= \max(0.0024 \cdot 0.15, 0.056 \cdot 0.2, 0.035 \cdot 0.5) \cdot 0.3 = 0.0052\end{aligned}$$

$$\psi_3(\text{☂}) = \text{☂}$$

Viterbi and weather prediction



Termination

$$p^*(X, Q|\Theta) = \max(\delta_3(i)) = \delta_3(\text{☁}) = 0.0269$$

$$q_3^* = \operatorname{argmax}(\delta_3(i)) = \text{☁}$$

Backtracking $i = n - 1 = 2$:

$$q_2^* = \psi_3(q_3^*) = \psi_3(\text{☁}) = \text{☁}$$

$i = n - 2 = 1$:

$$q_1^* = \psi_2(q_2^*) = \psi_2(\text{☁}) = \text{☀}$$

The most likely sequence of latent states is $Q^* = \{q_1^*, q_2^*, q_3^*\} = \{\text{☀}, \text{☁}, \text{☁}\}$.

Viterbi and weather prediction

Backtracking

