A Simple Markov Model For Weather Prediction

What is a first-order Markov chain?

$$P[q_t = j | (q_{t-1} = i, q_{t-2} = k, \ldots)] = P[q_t = j | q_{t-1} = i]$$

We consider only those processes for which the right-hand side is independent of time:

$$a_{ij} = P[q_t = j | q_{t-1} = i] \qquad 1 \le i, j \le N$$

with the following properties:

$$a_{ij} \ge 0 \qquad \forall j, i$$

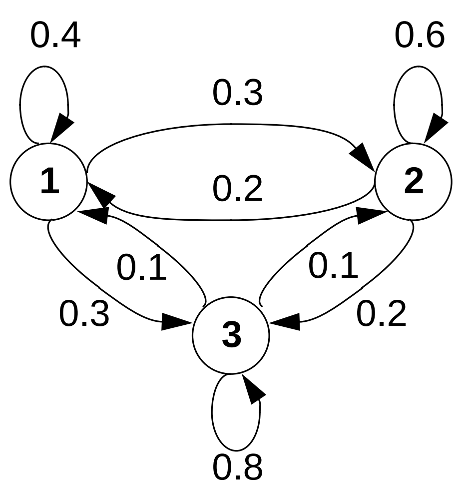
$$\sum_{j=1}^{N} a_{ij} = 1 \qquad \forall i$$

The above process can be considered observable because the output process is a set of states at each instant of time, where each state corresponds to an observable event.

Later, we will relax this constraint, and make the output related to the states by a second random process.

Example: A three-state model of the weather

State 1: precipitation (rain, snow, hail, etc.)
State 2: cloudy
State 3: sunny



Basic Calculations

Example: What is the probability that the weather for eight consecutive days is “sun-sun-sun-rain-rain-sun-cloudy-sun”?

Solution:

O =	sun	sun	sun	rain	rain	sun	cloudy	sun
	3	3	3	1	1	3	2	3

$$\begin{aligned} P(\bar{O}|Model) &= P[3]P[3|3]P[3|3]P[1|3]P[1|1]P[3|1]P[2|3]P[3|2] \\ &= \pi_3 a_{33} a_{31} a_{11} a_{13} a_{32} a_{23} \\ &= 1.536 \times 10^{-4} \end{aligned}$$

Example: Given that the system is in a known state, what is the probability that it stays in that state for *d* days?

O = i i i ... i j

$$\begin{aligned} P(\bar{O}|Model, q_1 = i) &= P(\bar{O}, q_1 = i|Model)/P(q_1 = i) \\ &= \pi_i a_{ii}^{d-1}(1 - a_{ii})/\pi_i \\ &= a_{ii}^{d-1}(1 - a_{ii}) \\ &= p_i(d) \end{aligned}$$

Note the exponential character of this distribution.

We can compute the expected number of observations in a state given that we started in that state:

$$\bar{d}_i = \sum_{d=1}^{\infty} d p_i(d) = \sum_{d=1}^{\infty} d a_{ii}^{d-1}(1 - a_{ii}) = \frac{1}{1 - a_{ii}}$$

Thus, the expected number of consecutive sunny days is (1/(1-0.8)) = 5; the expected number of cloudy days is 2.5, etc.

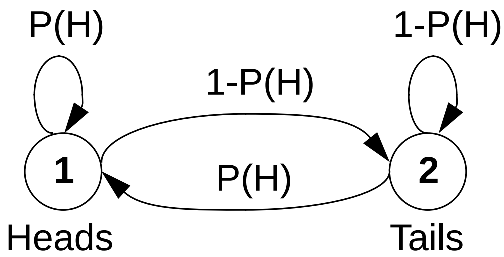
What have we learned from this example?

Why Are They Called "Hidden" Markov Models?

Consider the problem of predicting the outcome of a coin toss experiment. You observe the following sequence:

$$\bar{O} = (HHTTTHTTH \ldots H)$$

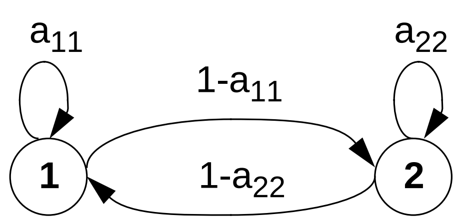
What is a reasonable model of the system?



1-Coin Model
(Observable Markov Model)

O = H H T T H T H H T T H ...

S = 1 1 2 2 1 2 1 1 2 2 1 ...



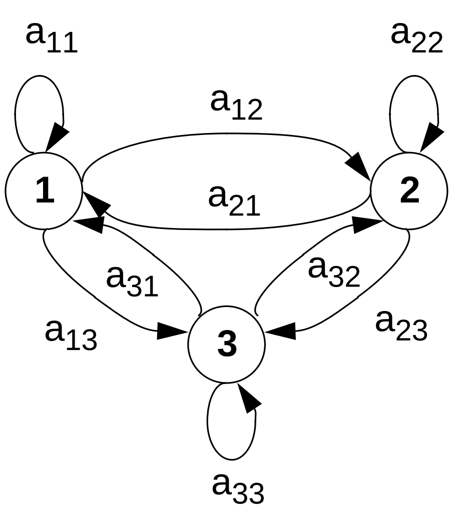
2-Coins Model
(Hidden Markov Model)

O = H H T T H T H H T T H ...

S = 1 1 2 2 1 2 1 1 2 2 1 ...

$P(H) = P_1$ $P(H) = P_2$

$P(T) = 1-P_1$ $P(T) = 1-P_2$



3-Coins Model
(Hidden Markov Model)

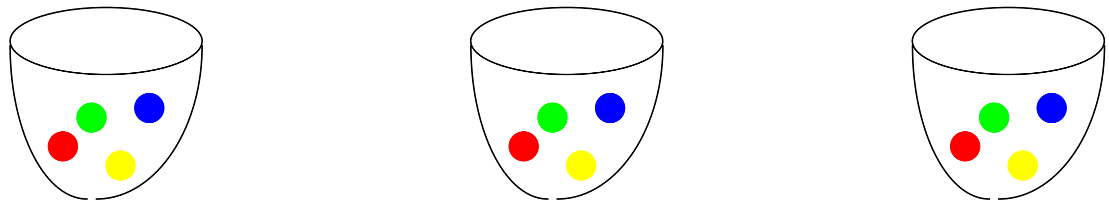
O = H H T T H T H H T T H ...

S = 3 1 2 3 3 1 1 2 3 1 3 ...

P(H):	P_1	P_2	P_3
P(T):	$1-P_1$	$1-P_2$	$1-P_3$

Why Are They Called Doubly Stochastic Systems?

The Urn-and-Ball Model



P(red)	$= b_1(1)$	P(red)	$= b_2(1)$	P(red)	$= b_3(1)$
P(green)	$= b_1(2)$	P(green)	$= b_2(2)$	P(green)	$= b_3(2)$
P(blue)	$= b_1(3)$	P(blue)	$= b_2(3)$	P(blue)	$= b_3(3)$
P(yellow)	$= b_1(4)$	P(yellow)	$= b_2(4)$	P(yellow)	$= b_3(4)$
...		...		...	

$$\overline{O} = \{\text{green, blue, green, yellow, red, ..., blue}\}$$

How can we determine the appropriate model for the observation sequence given the system above?

Elements of a Hidden Markov Model (HMM)

- N — the number of states
- M — the number of distinct observations per state
- The state-transition probability distribution $\underline{A} = \{a_{ij}\}$
- The output probability distribution $\underline{B} = \{b_j(k)\}$
- The initial state distribution $\pi = \{\pi_i\}$

We can write this succinctly as: $\lambda = (\underline{A}, \underline{B}, \pi)$

Note that the probability of being in any state at any time is completely determined by knowing the initial state and the transition probabilities:

$$\pi(t) = \underline{A}^{t-1}\pi$$

Two basic problems:

(1) how do we train the system?

(2) how do we estimate the probability of a given sequence (recognition)?

This gives rise to a third problem:

If the states are hidden, how do we know what states were used to generate a given output?

How do we represent continuous distributions (such as feature vectors)?

Formalities

The *discrete observation* HMM is restricted to the production of a finite set of discrete observations (or sequences). The output distribution at any state is given by:

$$b(k, i) \equiv P(\underline{y}(t) = k | \underline{x}(t) = i)$$

The observation probabilities are assumed to be independent of time. We can write the probability of observing a particular observation, $\underline{y}(t)$, as:

$$b(y(t)|i) \equiv P(\underline{y}(t) = y(t) | \underline{x}(t) = i)$$

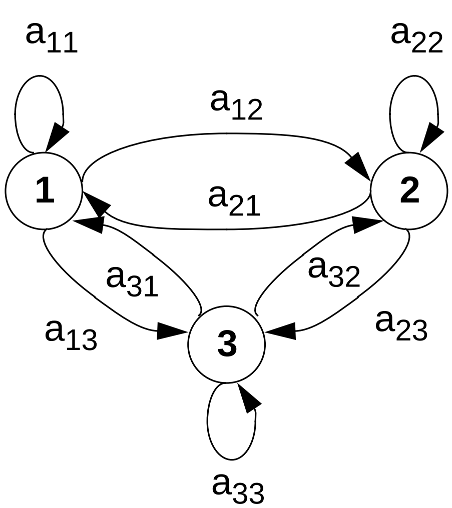
The observation probability distribution can be represented as a matrix whose dimension is K rows x S states.
We can define the observation probability vector as:

$$\boldsymbol{p}(t) = \begin{bmatrix} P(\underline{y}(t) = 1) \\ P(\underline{y}(t) = 2) \\ \dots \\ P(\underline{y}(t) = K) \end{bmatrix}, \quad \text{or,} \quad \boldsymbol{p}(t) = \boldsymbol{B}\pi(t) = \boldsymbol{B}\boldsymbol{A}^{t-1}\pi(1)$$

The mathematical specification of an HMM can be summarized as:

$$\boldsymbol{M} = \{S, \pi(1), \boldsymbol{A}, \boldsymbol{B}, \{\boldsymbol{y}_k, 1 \le k \le K\}\}$$

For example, reviewing our coin-toss model:



P(H): P_1 P_2 P_3
P(T): $1-P_1$ $1-P_2$ $1-P_3$

$$S = 3$$

$$\pi(1) = \begin{Bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{Bmatrix}$$

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

$$\boldsymbol{B} = \begin{bmatrix} P_1 & P_2 & P_3 \\ 1-P_1 & 1-P_2 & 1-P_3 \end{bmatrix}$$

Recognition Using Discrete HMMs

Denote any partial sequence of observations in time by:

$$y_{t_1}^{t_2} \equiv \{y(t_1), y(t_1+1), y(t_1+2), \ldots, y(t_2)\}$$

The forward partial sequence of observations at time t is

$$y_1^t \equiv \{y(1), y(2), \ldots, y(t)\}$$

The backward partial sequence of observations at time t is

$$y_{t+1}^T \equiv \{y(t+1), y(t+2), \ldots, y(T)\}$$

A complete set of observations of length T is denoted as $y \equiv y_1^T$.

What is the likelihood of an HMM?

We would like to calculate $P(M | \underline{y} = y)$ — however, we can't. We can (see the introductory notes) calculate $P(\underline{y} = y | M)$. Consider the brute force method of computing this. Let $\vartheta = \{i_1, i_2, \ldots, i_T\}$ denote a specific state sequence. The probability of a given observation sequence being produced by this state sequence is:

$$P(y | \vartheta, M) = b(y(1) | i_1) b(y(2) | i_2) \ldots b(y(T) | i_T)$$

The probability of the state sequence is

$$P(\vartheta | M) = P(\underline{x}(1) = i_1) a(i_2 | i_1) a(i_3 | i_2) \ldots a(i_T | i_{T-1})$$

Therefore,

$$P(y, (\vartheta | M)) = P(\underline{x}(1) = i_1) a(i_2 | i_1) a(i_3 | i_2) \ldots a(i_T | i_{T-1})$$
$$x \;\; b(y(1) | i_1) b(y(2) | i_2) \ldots b(y(T) | i_T)$$

To find $P(y | M)$, we must sum over all possible paths:

$$P(y | M) = \sum_{\forall \vartheta} P(y, (\vartheta | M))$$

This requires $O(2TS^T)$ flops. For $S = 5$ and $T = 100$, this gives about 1.6×10^{72} computations per HMM!