Stochastic Processes, Markov Chains, and Markov Models

L645 / B659
Dept. of Linguistics, Indiana University
Fall 2015
Where we’re going

Big picture: developing models for sequence labeling (e.g., POS tagging)

▶ We can accomplish a lot by making some local independence assumptions

Today:

▶ Review of finite-state automata
▶ Stochastic processes
▶ Markov chains
▶ Markov models

Next time (and beyond): POS tagging
Finite-State Automata

Definition FSA:
A finite-state automaton $A$ is a 5-tuple $(\Sigma, Q, i, F, E)$ where

- $\Sigma$ is a finite set called the alphabet
- $Q$ is a finite set of states
- $i \in Q$ is the initial state
- $F \subseteq Q$ is the set of final states, and
- $E \subseteq Q \times (\Sigma \cup \epsilon) \times Q$ is the set of transitions
Finite-State Automata (2)

S \rightarrow aA
S \rightarrow aB
A \rightarrow bC
A \rightarrow bB
B \rightarrow cA
B \rightarrow cC
C \rightarrow aD
Finite-State Automata (2)

S → aA
S → aB
A → bC
A → bB
B → cA
B → cC
C → aD
Finite-State Automata (2)

- $S \rightarrow aA$
- $S \rightarrow aB$
- $A \rightarrow bC$
- $A \rightarrow bB$
- $B \rightarrow cA$
- $B \rightarrow cC$
- $C \rightarrow aD$

Diagram:

- States: $S$, $A$, $B$, $C$, $D$
- Transitions:
  - $S \xrightarrow{a} A$
  - $A \xrightarrow{b} C$
  - $C \xrightarrow{a} D$
  - $C \xrightarrow{c} B$
  - $B \xrightarrow{c} C$
  - $B \xrightarrow{b} A$

- Initial state: $S$
- Final state: $D$
Definition DFSA:
A deterministic finite-state automaton is a 5-tuple \((\Sigma, Q, i, F, d)\) where

\(\Sigma\) is a finite set called the alphabet
\(Q\) is a finite set of states
\(i \in Q\) is the initial state
\(F \subseteq Q\) is the set of final states, and
\(d\) is the transition function that maps \(Q \times \Sigma\) to \(Q\).
Important Properties of FSAs

- **Determinization**: for every non-deterministic finite-state automaton there exists an equivalent deterministic FSA.

- **Minimization**: for every non-deterministic finite-state automaton, there exists an equivalent deterministic automaton with a minimal number of states.
What is in a State

Definition (state):
Given a deterministic FSA $M = (\Sigma, Q, i, F, d)$,

a state of $M$ is a triple $(x, q, y)$

where $q \in Q$ and $x, y \in \Sigma^*$
What is in a State

**Definition (state):**
Given a deterministic FSA $M = (\Sigma, Q, i, F, d)$,

a state of $M$ is a triple $(x, q, y)$

where $q \in Q$ and $x, y \in \Sigma^*$

example: $x = aaaaa$, $q = S$, $y = bbbbbbbbbbb$ for language $a^*b^*$
The “directly derives” Relation

Definition (directly derives):
Given a deterministic FSA $M = (\Sigma, Q, i, F, d)$,

a state $(x, q, y)$ directly derives state $(x', q', y')$: $(x, q, y) \vdash (x', q', y')$ iff

1. there is $\sigma \in \Sigma$ such that $y = \sigma y'$ and $x' = x\sigma$ (i.e., the reading head moves right one symbol $\sigma$)
2. $d(q, \sigma) = q'$
The “derives” Relation

Definition (derives): Given a deterministic FSA $M = (\Sigma, Q, i, F, d)$,

a state $A$ derives state $B$: $(x, q, y) \vdash^* (x’, q’, y’)$ iff

there is a sequence $S_0 \vdash S_1 \ldots S_k$

such that $A = S_0$ and $B = S_k$
Acceptance

**Definition (acceptance):**
Given a deterministic FSA $M = (\Sigma, Q, i, F, d)$ and a string $x \in \Sigma^*$

$M$ accepts $x$ iff there is a $q \in F$ such that $(0, i, x) \vdash^* (x, q, 0)$. 
Definition (language accepted by M):
Given a deterministic FSA $M = (\Sigma, Q, i, F, d)$,

the language $L(M)$ accepted by $M$ is the set of all strings accepted by $M$. 
Important Properties of FSAs

Given the FSAs $A$, $A_1$, and $A_2$ and the string $w$, the following properties are decidable:

- Membership: $w \in (A)$?
- Emptiness: $L(A) = \emptyset$?
- Totality: $L(A) = \Sigma^*$?
- Subset: $L(A_1) \subseteq L(A_2)$?
- Equality: $L(A_1) = L(A_2)$?
Encoding FSAs as Matrices

- Basic idea: encode alphabet symbols that appear on transitions in a given FSA by a state transition matrix.
- The transition matrix will have a 1 in a given cell in case its row & column match the from- and to-states of a transition for the alphabet symbol in the FSA.
  - All other cells are filled with 0.
- By matrix multiplication, one can determine the number of successful paths through an automaton.
FSA Example

A non-deterministic automaton:

What is its language?
FSA Matrix

\[ a = \begin{bmatrix} s_0 & s_1 & s_2 \\ s_0 & 1 & 1 & 0 \\ s_1 & 1 & 0 & 0 \\ s_2 & 0 & 0 & 0 \end{bmatrix} \]

\[ b = \begin{bmatrix} s_0 & s_1 & s_2 \\ s_0 & 0 & 1 & 0 \\ s_1 & 0 & 1 & 0 \\ s_2 & 0 & 0 & 0 \end{bmatrix} \]
### FSA Matrix (2)

$$c = \begin{bmatrix}
 s_0 & s_1 & s_2 \\
 s_0 & 0 & 0 & 0 \\
 s_1 & 0 & 0 & 1 \\
 s_2 & 0 & 0 & 0 \\
\end{bmatrix}$$

$$\text{init} = \begin{bmatrix}
 s_0 & s_1 & s_2 \\
 1 & 0 & 0 \\
\end{bmatrix}$$

$$\text{final} = \begin{bmatrix}
 s_0 \\
 s_1 \\
 s_2 \\
\end{bmatrix}$$
A stochastic or random process is a sequence $\xi_1, \xi_2, \ldots$ of random variables based on the same sample space $\Omega$.

The possible outcomes of the random variables are called the set of possible states of the process.

The process will be said to be in state $\xi_t$ at time $t$.

Note that the random variables are in general not independent.

In fact, the interesting thing about stochastic process is the dependence between the random variables.
The sequence of outcomes when repeatedly casting the die is a stochastic process with discrete random variables and a discrete time parameter.
There are three telephone lines, and at any given moment 0, 1, 2 or 3 of them can be busy. Once every minute we will observe how many of them are busy. This will be a random variable with $\Omega_\xi = \{0, 1, 2, 3\}$. Let $\xi_1$ be the number of busy lines at the first observation time, $\xi_2$ the number of busy lines at the second observation time, etc.

The sequence of the number of the busy lines then forms a random process with discrete random variables and a discrete time parameter.
Characterizing a random process

To fully characterize a random process with time parameter $t$, we specify:

1. the probability $P(\xi_1 = x_j)$ of each outcome $x_j$ for the first observation, i.e., the initial state $\xi_1$
2. for each subsequent observation/state $\xi_{t+1} : t = 1, 2, \ldots$ the conditional probabilities $P(\xi_{t+1} = x_{i_{t+1}} | \xi_1 = x_i, \ldots, \xi_t = x_{i_t})$

Terminate after some finite number of steps $T$
A Markov chain is a special type of stochastic process where the probability of the next state conditional on the entire sequence of previous states up to the current state is in fact only dependent on the current state.

This is called the Markov property and can be stated as:

\[
P(\xi_{t+1} = x_{i_{t+1}} | \xi_1 = x_i, \ldots, \xi_t = x_i) = P(\xi_{t+1} = x_{i_{t+1}} | \xi_t = x_i)
\]
Markov Chains and the Markov Property (2)

The probability of a Markov chain \( \xi_1, \xi_2, \ldots \) can be calculated as:

\[
P(\xi_1 = x_{i_1}, \ldots, \xi_t = x_{i_t}) = \\
= P(\xi_1 = x_{i_1}) \cdot P(\xi_2 = x_{i_2} | \xi_1 = x_{i_1}) \cdot \\
\ldots \cdot P(\xi_t = x_{i_t} | \xi_{t-1} = x_{i_{t-1}})
\]

The conditional probabilities \( P(\xi_{t+1} = x_{i_{t+1}} | \xi_t = x_{i_t}) \) are called the *transition probabilities* of the Markov chain.

A *finite Markov chain* must at each time be in one of a finite number of states.
Markov Chains: Example

We can turn the telephone example with 0, 1, 2 or 3 busy lines into a (finite) Markov chain by assuming that the number of busy lines will depend only on the number of lines that were busy the last time we observed them, and not on the previous history.
Transition Matrix for a Markov Process

Consider a Markov chain with $n$ states $s_1, \ldots, s_n$. Let $p_{ij}$ denote the transition probability from state $s_i$ to state $s_j$, i.e., $P(\xi_{t+1} = s_j | \xi_t = s_i)$.

The transition matrix for this Markov process is then defined as

$$
P = \begin{bmatrix}
p_{11} & \cdots & p_{1n} \\
\vdots & \ddots & \vdots \\
p_{n1} & \cdots & p_{nn}
\end{bmatrix}, \quad p_{ij} \geq 0, \quad \sum_{j=1}^{n} p_{ij} = 1, \quad i = 1, \ldots, n
$$

In general, a matrix with these properties is called a stochastic matrix.
Transition Matrix: Example

In the telephone example, a possible transition matrix could for example be:

\[
P = \begin{bmatrix}
0.2 & 0.5 & 0.2 & 0.1 \\
0.3 & 0.4 & 0.2 & 0.1 \\
0.1 & 0.3 & 0.4 & 0.2 \\
0.1 & 0.1 & 0.3 & 0.5
\end{bmatrix}
\]
Transition Matrix: Example (2)

Assume that currently, all three lines are busy.

What is then the probability that at the next point in time exactly one line is busy?

The element in Row 3, Column 1 ($p_{31}$) is 0.1, and thus $p(1|3) = 0.1$. (Note that we have numbered the rows and columns 0 through 3.)
Transition Matrix after Two Steps

\[ p_{ij}^{(2)} = P(\xi_{t+2} = s_j | \xi_t = s_i) \]

\[ = \sum_{r=1}^{n} P(\xi_{t+1} = s_r | \xi_t = s_i) \cdot P(\xi_{t+2} = s_j | \xi_{t+1} = s_r) \]

\[ = \sum_{r=1}^{n} p_{ir} \cdot p_{rj} \]

The element \( p_{ij}^{(2)} \) can be determined by matrix multiplication as the value in row \( i \) and column \( j \) of the transition matrix \( P^2 = PP \).

More generally, the transition matrix for \( t \) steps is \( P^t \).
Matrix Multiplication

Practice at:

http://www.zweigmedia.com/RealWorld/tutorialsf1/frames3_2.html
Matrix after Two Steps: Example

For the telephone example: Assuming that currently all three lines are busy, what is the probability of exactly one line being busy after two steps in time?

\[
P^2 = PP = \begin{bmatrix}
  s_0 & s_1 & s_2 & s_3 \\
  s_0 & 0.22 & 0.37 & 0.25 & 0.16 \\
  s_1 & 0.21 & 0.38 & 0.25 & 0.16 \\
  s_2 & 0.17 & 0.31 & 0.30 & 0.22 \\
  s_3 & 0.13 & 0.23 & 0.31 & 0.33 \\
\end{bmatrix}
\]

\[\Rightarrow p_{31}^{(2)} = 0.23\]
Initial Probability Vector

A vector $\mathbf{v} = [v_1, \ldots, v_n]$ with $v_i \geq 0$ and $\sum_{i=1}^{n} v_i = 1$ is called a probability vector.

The probability vector that determines the state probabilities of the observations of the first element (state) in a Markov chain, i.e., where $v_i = P(\xi_1 = s_i)$, is called an initial probability vector.

The initial probability vector and the transition matrix together determine the probability of the chain in a particular state at a particular point in time.
Initial Probability Vector

If $p^t(s_i)$ is the probability of a Markov chain being in state $s_i$ at time $t$, i.e., after $t - 1$ steps, then

$$[p^t(s_1), \ldots, p^t(s_n)] = vP^{t-1}$$
Initial Probability Vector: Example

Let \( v = [0.5 \ 0.3 \ 0.2 \ 0.0] \) be the initial probability vector for the telephone example of a Markov chain.

What is then the probability that after two steps exactly two lines are busy?

\[
vP^2 = vPP = \begin{bmatrix}
0.207 & 0.361 & 0.260 & 0.172
\end{bmatrix}
\]

\[\Rightarrow p(\xi_3 = s_2) = 0.26\]
Markov Models

Markov models add the following to a Markov chain:

- A sequence of random variables $\eta_t$ for $t = 1, \ldots, T$
- These represent the signal emitted at time $t$
  - We will use $\sigma_j$ to refer to a particular signal
A Markov Model consists of:

- a finite set of states $\Omega = \{s_1, \ldots, s_n\}$;
- an signal alphabet $\Sigma = \{\sigma_1, \ldots, \sigma_m\}$;
- an $n \times n$ state transition matrix $P = [p_{ij}]$ where $p_{ij} = P(\xi_{t+1} = s_j | \xi_t = s_i)$;
- an $n \times m$ signal matrix $A = [a_{ij}]$, which for each state-signal pair determines the probability $a_{ij} = P(\eta_t = \sigma_j | \xi_t = s_i)$ that signal $\sigma_j$ will be emitted given that the current state is $s_i$;
- and an initial vector $v = [v_1, \ldots, v_n]$ where $v_i = P(\xi_1 = s_i)$. 

Markov Models (1)
HMM Example 1: Crazy Softdrink Machine

with emission probabilities:

\[
\begin{array}{ccc}
\text{cola} & \text{ice-t} & \text{lem} \\
CP & 0.6 & 0.1 & 0.3 \\
IP & 0.1 & 0.7 & 0.2 \\
\end{array}
\]

from: Manning/Schütze; p. 321
Markov Models (2)

\[ p^{(t)}(s_i, \sigma_j) = p^{(t)}(s_i) \cdot p(\eta_t = \sigma_j | \xi_t = s_i) \]

where \( p^{(t)}(s_i) \) is the \( i \)th element of the vector \( \mathbf{vP}^{t-1} \)

- This means that the probability of emitting a particular signal depends only on the current state
- ... and not on the previous states
The probability that signal $\sigma_j$ will be emitted at time $t$ is then:

$$p^{(t)}(\sigma_j) = \sum_{i=1}^{n} p^{(t)}(s_i, \sigma_j) = \sum_{i=1}^{n} p^{(t)}(s_i) \cdot p(\eta_t = \sigma_j \mid \xi_t = s_i)$$

Thus if $p^{(t)}(\sigma_j)$ is the probability of the model emitting signal $\sigma_j$ at time $t$, i.e., after $t - 1$ steps, then

$$[p^{(t)}(\sigma_1), \ldots, p^{(t)}(\sigma_m)] = vP^{t-1}A$$