Language Acquisition and Parameters: Part II

Matilde Marcolli

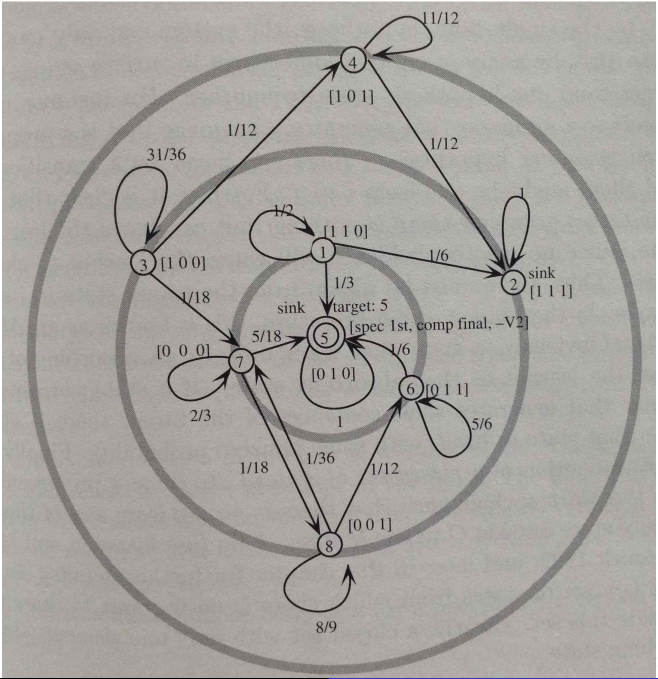
CS101: Mathematical and Computational Linguistics

Winter 2015

Transition Matrices in the Markov Chain Model

- absorbing states correspond to local maxima (unique absorbing state at the target gives learnability)
- probability matrix of a Markov Chain $T = (T_{ij})$ with $T_{ij} =$ probability $\mathbb{P}(s_i \to s_j)$ of moving from state s_i to state s_j
- absorbing states: rows of T with only one 1 entry and all zeros
- powers T^m of probability matrix: entry T^m_{ij} = probability of going from state s_i to state s_j in exactly m steps
- probabilities of reaching s_j in the limit with initial state s_i:

$$T_\infty = \lim_{m\to\infty} T^m$$


11/12
4
[1 0 1]
1/12
1/12
31/36
1/2
[1 1 0]
1
1
1/6
3
[1 0 0]
2
sink
[1 1 1]
1/3
1/18
sink
target: 5
5
[spec 1st, comp final, –V2]
5/18
[0 0 0]
7
1/6
[0 1 0]
6
[0 1 1]
2/3
1
5/6
1/18
1/36
1/12
[0 0 1]
8
8/9

Transition Matrix for the 3-parameter example

$$T = \begin{pmatrix} 1/2 & 1/6 & & & 1/3 & & & \\ & 1 & & & & & & \\ & & 3/4 & 1/12 & & & 1/6 & \\ & 1/12 & & 11/12 & & & & \\ & & & & 1 & & & \\ & & & & 1/6 & 5/6 & & \\ & & & & 5/18 & & 2/3 & 1/18 \\ & & & & & 1/12 & 1/36 & 8/9 \end{pmatrix}$$

- local maxima problem: two absorbing states

Limiting probabilities matrix for the 3-parameter example

$$T_\infty = \begin{pmatrix} & 2/5 & & & 3/5 & & & \\ & 1 & & & & & & \\ & 2/5 & & & 3/5 & & & \\ & 1 & & & & & & \\ & & & & 1 & & & \\ & & & & 1 & & & \\ & & & & 1 & & & \\ & & & & 1 & & & \end{pmatrix}$$

- for learnability irrespective of initial state would need column of 1's at the target state
- here if starting at s_2 or s_4 end up at s_2 (local maximum) instead of target s_5; initial states s_5, s_6, s_7, s_8 converge to correct target s_5
- starting at s_1 or s_3 will reach true target s_5 with probability $3/5$ and false target s_2 with probability $2/5$

• the last case in the example shows that there are initial states for which there is a "triggered path to the target", but the learner does not take that path with probability 1, only with a smaller probability

• if take same 3-parameter model but with target state s_1 and transition matrix

$$T = \begin{pmatrix} 1 & & & & & & & \\ 1/6 & 5/6 & & & & & & \\ 5/18 & & 2/3 & 1/18 & & & & \\ & 3/36 & 1/36 & 8/9 & & & & \\ 1/3 & & & & 23/36 & 1/36 & & \\ & 5/36 & & & & 31/36 & & \\ & & 1/18 & & & & 11/12 & 1/36 \\ & & & 1/18 & & & & 17/18 \end{pmatrix}$$

then no other local maxima and T_∞ has first column of 1's

Eigenvalues and eigenvectors of transition matrices

- matrix T is stochastic: $T_{ij} \geq 0$ and $\sum_j T_{ij} = 1$ for all i

- Perron–Frobenius theorem: if T is irreducible (some power T^m has all entries $T^m_{ij} > 0$) then
- spectral radius $\rho(T) = 1 =$ PF eigenvalue
- PF eigenvalue is simple
- PF (left) eigenvector v with all $v_i > 0$ (uniform: $v_i = 1$)
- period $h = \mathrm{lcd}\{m : T^m_{ii} > 0\}$ number of eigenvalues $|\lambda| = 1$

- but irreducible condition means graph strongly connected: every vertex is reachable from every other vertex... in general does not happen with Markov chains: in general T not irreducible non-irreducible transition matrices T of a Markov Chain

- $\lambda = 1$ is always an eigenvalue

- all other eigenvalues have $|\lambda| < 1$

- multiplicity of $\lambda = 1$ is number of closed classes C_i in decomposition of the Markov Chain

- if T has a basis of linearly independent left eigenvectors $\mathbf{v}_i$ with $\mathbf{v}_i T = \lambda_i \mathbf{v}_i$ (and $\mathbf{w}_i$ right eigenvectors $T\mathbf{w}_i = \lambda_i \mathbf{w}_i$)

$$T^m = \sum_i \lambda_i^m \, \mathbf{w}_i \mathbf{v}_i$$

linear combination of matrices $\mathbf{w}_i \mathbf{v}_i$ (independent of m) with coefficients λ_i^m

- initial probabilities $\pi_i \geq 0$, with $\sum_i \pi_i = 1$, of having state s_i as initial state
- after m steps: $\pi_i^{(m)} = \sum_j \pi_j T_{ji}^m$
- limiting distribution:

$$\pi_i^{(\infty)} = \sum_j \pi_j T_{\infty,ji} = \lim_{m \to \infty} \sum_j \pi_j T_{ji}^m$$

probability of learner approaching state s_i in the limit

- if target state (say s_1) is learnable, then $\pi_1^{(\infty)} = 1$ and $\pi_i^{(\infty)} = 0$ for $i \neq 1$

Rate of convergence

- rate of convergence of $\pi_i^{(m)}$ to $\pi_i^{(\infty)}$ is rate of convergence of T^m to T_∞, which is rate of convergence of $\lambda_i \to 0$ for $|\lambda_i| < 1$

$$\|\pi^{(m)} - \pi^{(\infty)}\| = \| \sum_{i \geq 2} \lambda_i^m \pi \mathbf{w}_i \mathbf{v}_i \| \leq \max_{i \geq 2} \{ |\lambda_i|^m \} \sum_{i \geq 2} \| \pi \mathbf{w}_i \mathbf{v}_i \|$$

- estimate rate of decay of second largest eigenvalue

Summary

1. construct Markov Chain for Parameter space
2. compute transition matrix T
3. compute eigenvalues
4. if multiplicity of eigenvalue $\lambda = 1$ is more than one: target is unlearnable (local maxima problem)
5. if multiplicity one, check if basis of independent eigenvectors
6. if yes, find rate of decay of second largest eigenvalue: learnability at that speed
7. if not, project onto subspaces of lower dimension

Markov Chains and Learning Algorithms

• how broad is the Markov Chain method in modeling learning algorithms?

• suppose given $\mathcal{A} : \mathcal{D} \to \mathcal{H}$; hypotheses $\mathfrak{h}_n = \mathcal{G}$ and $\mathfrak{h}_{n+1} = \mathcal{G}'$

• probability of passing from $\mathcal{A}(\tau_n) = \mathfrak{h}_n = \mathcal{G}$ to $\mathcal{A}(\tau_{n+1}) = \mathfrak{h}_{n+1} = \mathcal{G}'$ at $n+1$-st input is measure of set

$$A_{n,\mathcal{G}} \cap A_{n+1,\mathcal{G}'} = \{\tau \,|\, \mathcal{A}(\tau_n) = \mathcal{G}\} \cap \{\tau \,|\, \mathcal{A}(\tau_{n+1}) = \mathcal{G}'\}$$

• measure with respect to μ^∞ on $\mathfrak{A}^\omega$ determined by μ on $\mathfrak{A}^\star$ (supported on target language $\mathcal{L}$ for positive examples only)

$$\mathbb{P}(\mathfrak{h}_{n+1} = \mathcal{G}' \,|\, \mathfrak{h}_n = \mathcal{G}) = \frac{\mu_\infty(A_{n,\mathcal{G}} \cap A_{n+1,\mathcal{G}'})}{\mu_\infty(A_{n,\mathcal{G}})}$$

assuming $\mu_\infty(A_{n,\mathcal{G}}) > 0$

Inhomogeneous Markov Chain

- state space = set of possible grammars $\mathcal{H}$ = set of possible (binary) syntactic parameters

- Transition matrix at n-th step:

$$T_n(s,s') = \mathbb{P}(s \to s') = \mathbb{P}(\mathcal{A}(\tau_{n+1}) = \mathcal{G}_{s'} \mid \mathcal{A}(\tau_n) = \mathcal{G}_s)$$

$$= \frac{\mu_\infty(A_{n,\mathcal{G}_s} \cap A_{n+1,\mathcal{G}_{s'}})}{\mu_\infty(A_{n,\mathcal{G}_s})}$$

- these satisfy $\sum_{s'} T_n(s,s') = 1$ for all s

- to define $T_n(s,s')$ also when $\mu_\infty(A_{n,\mathcal{G}_s}) = 0$ take a set of $\alpha_s > 0$ with $\sum_s \alpha_s = 1$ and set

$$T_n(s,s') = \alpha_{s'}, \quad \text{when} \quad \mu_\infty(A_{n,\mathcal{G}_s}) = 0.$$

- the transition matrix $T = T_n$ is time dependent

- Conclusion: any deterministic learning algorithm $\mathcal{A} : \mathcal{D} \to \mathcal{H}$ can be modeled by an inhomogeneous Markov Chain

- the inhomogeneous Markov Chain depends on $\mathcal{A}$, on the target language $\mathcal{L}_{\mathcal{G}}$ and on the measure μ

- memoryless learner hypothesis: $\mathcal{A}(\tau_{n+1}) = F(\mathcal{A}(\tau_n), \tau(n+1))$

$$\Rightarrow \quad T_n(s, s') = T(s.s') = \mu(\{x \in \mathfrak{A}^\star \,|\, F(s,x) = s'\}$$

- memory limited learning algorithms: m-memory limited if $\mathcal{A}(\tau_n)$ only depends on last m sentences in text τ_m and the previous grammatical hypothesis

- Fact: if $\mathcal{H}$ learnable by a memory limited algorithm, in fact learnable by a memoryless algorithm