Applied Probabilistic Machine Learning

Markov models

Hugues Richard RichardH@rki.de
Ivan Tunov Ivan.Tunov@student.hpi.uni-potsdam.de

MF1 - Genome Competence Center Robert Koch Institute (RKI)
DACS Hasso Plattner Institute (HPI)

28 November 2024

Learning goals

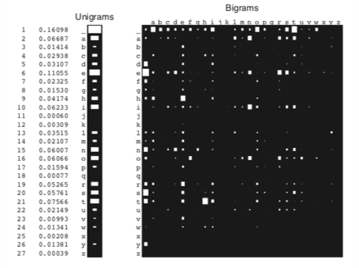
- Understand what are Markov Models
  - Model parameters
  - Representations of Markov Models
- Be able to manipulate Markov Models
- Learn about Markov Models properties
- Study example application of Markov Models
- Extensions of Markov models

Introduction

- Let's consider sequential data over discrete values
  - first hypothesis is that observations are independent, identically distributed (i.i.d.)

$$\mathrm{x}_1, \mathrm{x}_2, \dots, \mathrm{x}_n, \quad \mathrm{x}_i \sim \text{Categorical}(K)$$

  - $\mathrm{x}_i \perp\!\!\!\perp \mathrm{x}_j$
- Not always the case, we expect a **dependency** for time series:
  - Hand drawing
  - People/image tracking
  - Texts, genomes



- How to account for a local dependency?

The Markovian hypothesis

- We consider a time oriented process, using product rule the probability of a sequence is:

$$\mathbb{P}(x_1, \dots, x_n) = \mathbb{P}(x_n \mid x_1, \dots, x_{n-1}) \cdot \mathbb{P}(x_1, \dots, x_{n-1})$$

$$= \mathbb{P}(x_n \mid x_1, \dots, x_{n-1}) \cdot \mathbb{P}(x_{n-1} \mid x_1, \dots, x_{n-2}) \dots \mathbb{P}(x_3 \mid x_1, x_2) \cdot \mathbb{P}(x_2 \mid x_1)\mathbb{P}(x_1)$$

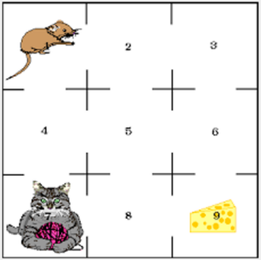
- We cannot estimate a conditional distribution from **all previous observations**
  - keep information about the current state (order 1):

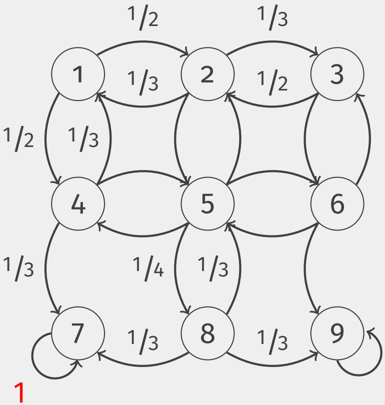
$$\mathbb{P}(x_n \mid x_1, \dots, x_{n-1}) = \mathbb{P}(x_n \mid x_{n-1})$$

  - Not much but still better than independence (order 0):

$$\mathbb{P}(x_n \mid x_1, \dots, x_{n-1}) = \mathbb{P}(x_n)$$

- Mouse in a maze:
  - 3×3 rooms, we monitor the mouse location between each of her room change
  - In each room, the mouse chooses one of the door randomly: $\mathbb{P}(x_{n+1} = 2 \mid x_n = 1) = \frac{1}{2}$, $\mathbb{P}(x_{n+1} = 5 \mid x_n = 2) = \frac{1}{3}$





Transition matrices

- Probability of a path? Similar to an automaton

$$\mathbb{P}(X_{1:5} = (1,2,5,8,9)) = \mathbb{P}(X_1 = 1)\mathbb{P}(X_2 = 2 \mid X_1 = 1) \dots \mathbb{P}(X_5 = 9 \mid X_4 = 8)$$
$$= 1 \cdot {}^1/_2 \cdot {}^1/_3 \cdot {}^1/_4 \cdot {}^1/_3 = \frac{1}{72}$$

- The weighted graph and the transition matrix are equivalent.
- Parameters of a homogeneous Markov chain over Σ:
  - Starting distribution $\pi = \mathbb{P}(X_1)$
  - Transition matrix $A_{i,j} = \mathbb{P}(X_{t+1} = j \mid X_t = i)$
    (**from** the rows **to** the columns)

$$A_{\text{maze}} = \begin{pmatrix} 0 & 1/2 & 0 & 1/2 & 0 & 0 & 0 & 0 & 0 \\ 1/3 & 0 & 1/3 & 0 & 1/3 & 0 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & 0 & 0 & 1/2 & 0 & 0 & 0 \\ 1/3 & 0 & 0 & 0 & 1/3 & 0 & 1/3 & 0 & 0 \\ 0 & 1/4 & 0 & 1/4 & 0 & 1/4 & 0 & 1/4 & 0 \\ 0 & 0 & 1/3 & 0 & 0 & 1/3 & 0 & 0 & 1/3 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1/3 & 0 & 1/3 & 0 & 1/3 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Properties of Markov chains

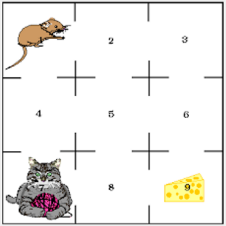
1. Probability of a sequence
2. Probability of two non consecutive events
3. What is the long term behaviour?
4. How to estimate the parameters of a Markov Chain?

Probability of a sequence

- Because of Markov property, the likelihood is a product over the consecutive observations

$$\mathbb{P}(x_1, \dots, x_n) = \pi_{x_1} \prod_{i=2}^{n} A[x_{i-1}, x_i]$$

- This conditional independence can be summarised with a graph and the probability of a sequence is like a walk on an automaton



$$\mathbb{P}(X_3 = 5 \mid X_1 = 1) = a_{1,2} \cdot a_{2,5} + a_{1,4} \cdot a_{4,5}$$

$$\begin{aligned}\mathbb{P}(X_3 = j \mid X_1 = i) &= \sum_{\ell \in \Sigma} \mathbb{P}(X_2 = j, X_1 = \ell \mid X_0 = i) \\ &= \sum_{\ell \in \Sigma} a_{i,\ell} \cdot a_{\ell,j} \\ &= A^2(i,j)\end{aligned}$$

- This generalizes to the k-step process (exercise)
  - It is a Markov chain
  - it transition Matrix is A^k: $\mathbb{P}(X_{n+k} = j \mid X_n = i) = A^k(i,j)$
  - Easy to compute the state of the system after t steps: $\mathbb{P}(X_n = i) = (\pi \cdot A^n)[i]$ (note that $\pi \cdot A^n$ is a vector of size $|\Sigma|$.)

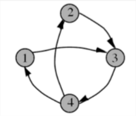
- We can use the powers of the transition matrix to look at the long term behavior
  - What is the probability that the mouse will end first in the cat room? In the cheese room?
  - We can compute powers of A, using the starting distribution $\pi = (1, 0, 0, 0, 0, 0, 0, 0, 0)$

$$\pi \cdot A^n = A^n[1, :]$$

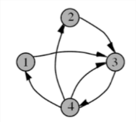
	1	2	3	4	5	6	cat	8	cheese
π*A^2	0.33	0	0.16	0	0.34	0	0.17	0	0
π*A^3	0	0.33	0	0.25	0	0.17	0.17	0.08	0
π*A^4	0.19	0	0.16	0	0.28	0	0.28	0	0.08
π*A^10	0.08	0	0.08	0	0.12	0	0.46	0	0.26
π*A^20	0.02	0	0.02	0	0.03	0	0.57	0	0.37
π*A^100	0	0	0	0	0	0	0.6	0	0.4

Long term behavior

- tl;dr: A *well behaved* Markov chain will converge to a **unique** stationary distribution
- Now, what are the component of a badly behaved Markov chain?
  - Absorbing states: dead ends in the chain (think of cat and cheese in the maze → 2 stationnary distributions)
  - if $\forall\, i, j \in \Sigma,\ \exists k \ / \ A^k(i, j) > 0$ then there are no absorbing states and the chain is **irreducible**
  - Periodic states: closed loops in the chain
    A state is periodic if we can get back to it only at a given multiple of k
  - a chain with no periodic states is called **aperiodic**.



All states have period 3



the chain is aperiodic

- If the chain is irreducible and aperiodic (well behaved)
- the stationary distribution μ is **unique** and $\mu \cdot A = \mu$
  - μ can be obtained by solving $\mu \cdot A = \mu \Leftrightarrow \mu(A - I) = 0$
  - μ is the eigenvector of A associated with the eigenvalue $\lambda_1 = 1$ (1 is also the largest eigenvalue)
  - Each row of A^k converges towards μ
    - In other words Markov chains have **short term memory** $\rightarrow X_t$ does not influence X_{t+k} when $k \nearrow$
    - Convergence is **exponentially fast**

$$\max_i \sum_{j \in \Sigma} |A^{k}[i,j] - \mu[j]| \leq C \cdot |\Sigma|^{r_2 - 1} \cdot |\lambda_2|^{k}$$

r_2 : multiplicity of λ_2

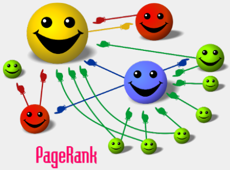
Advantage: Easy to approximate after spectral analysis
Drawback: cannot model long range effects

- Note: If the first state in the sequence is not specified, we usually set $\pi = \mu$.
  That way the chain already starts with the stationary distribution.

Famous stationary distributions

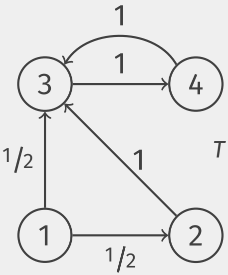
Google PageRank score

- How to decide the most relevant answers from a web search?
  - First web browsers (altavista...): number of pages linking to it (can easily be tricked with false websites)



- Consider a random (mouse) websurfer
  - Click all outgoing links on a page equally likely
    → Markov chain over webpages!
  - Which page would the surfer land more often?
    → This is the stationary distribution!
  - In practice, two cases can affect irreducibility and aperiodicity
    - dead-ends: pages with no outgoing links
    - disconnected components in the network
    - add a random page reset (Google used $p = 0.15$)

- Let's consider an internet with 5 pages



$$T = 0.15 \cdot \begin{pmatrix} 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \end{pmatrix} + 0.85 \cdot \begin{pmatrix} 0 & 1/2 & 1/2 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

$$T \approx \begin{pmatrix} 0.04 & 0.46 & 0.46 & 0.04 \\ 0.04 & 0.04 & 0.88 & 0.04 \\ 0.04 & 0.04 & 0.04 & 0.88 \\ 0.04 & 0.04 & 0.88 & 0.04 \end{pmatrix} \Rightarrow \text{PageRank } \mu^T = \begin{pmatrix} 0.0375 \\ 0.0534 \\ 0.4711 \\ 0.4379 \end{pmatrix}$$

PageRank example

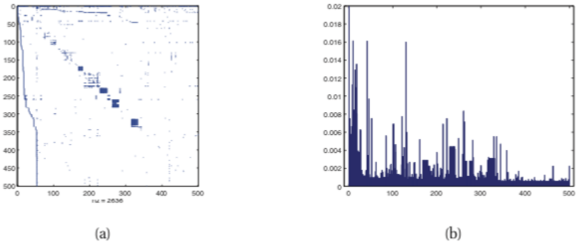
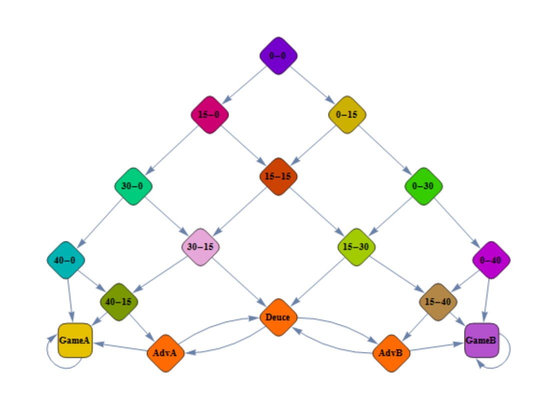


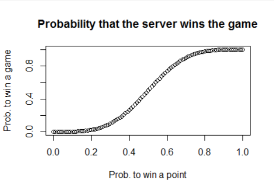
Figure 17.6 (a) Web graph of 500 sites rooted at `www.harvard.edu`. (b) Corresponding page rank vector. Figure generated by `pagerankDemoPmtk`, Based on code by Cleve Moler (Moler 2004).

[Murphy, 2022]

Simple example: Tennis match

- Consider a Tennis match where player A has a probability p to win a point on his/her serve
  - All scores configurations can be enumerated: state space Σ.
  - The sequence of scores is a Markov chain.

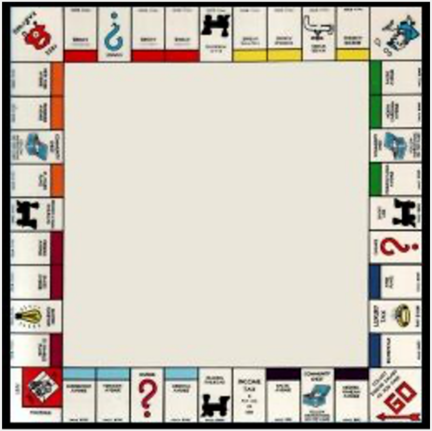




(source: wolfram.com)

Simple example: Monopoly project

- State space: All squares on the board (almost)
- Transition Probabilities of move are parameterised by 2-dice throws



(see project n. 3)

Parameters estimation

- The log-likelihood of a sequence $x_1, \dots x_n$ writes:

$$\log \ell(x_{1:n}, \theta) = \log \mathbb{P}(x_1) + \sum_{i=2}^{n} \log A_\theta[x_{i-1}, x_i]$$

- if we count the number of co-occurence of pairs of states $n_{a,b}$

$$n_{a,b} = \sum_{i=2}^{n} \mathbb{I}_{\{x_{i-1}=a, x_i=b\}}$$

then

$$\log \ell(x_{1:n}, \theta) = \log \mathbb{P}(x_1) + \sum_{a \in \Sigma} \sum_{b \in \Sigma} n_{a,b} \log A_\theta[a, b]$$

- Maximum likelihood estimators are like the ones for Multinoulli (neglecting sequence start)

$$\hat{A}_{ML}[a, b] = \frac{n_{a,b}}{n_{a,\bullet}}$$

Extensions of Markov models

- Order k Markov model increase the dependency:

$$\mathbb{P}(\mathrm{x}_n \mid \mathrm{x}_1, \dots, \mathrm{x}_{n-1}) = \mathbb{P}(\mathrm{x}_n \mid \mathrm{x}_{n-1}, \mathrm{x}_{n-2}, \mathrm{x}_{n-k})$$

  - But the number of parameters increases exponentially!
  - Order n Markov chains on Σ, can be viewed as order 1 Markov chains on Σ^k
    - Example $\Sigma = \{a, b\}$ and a Markov chain of order 2 with transitions $\alpha_{ij,k} = \mathbb{P}(\mathrm{x}_n = k \mid x_{n-1} = j, x_{n-2} = i)$
    - we can write the transition matrix A on Σ^2:

$$A = \begin{pmatrix} \alpha_{aa,a} & \alpha_{aa,b} & 0 & 0 \\ 0 & 0 & \alpha_{ab,a} & \alpha_{ab,b} \\ \alpha_{ba,a} & \alpha_{ba,b} & 0 & 0 \\ 0 & 0 & \alpha_{bb,a} & \alpha_{bb,b} \end{pmatrix}$$

- More parsimonious models were proposed:
  - Variable order Markov chains.

- Understand what are Markov Models
  - Models for sequential data with short range dependency
  - Fully parametrised with a transition Matrix + Init proba.
- Be able to manipulate Markov Models and models properties
  - Probability distributions can be computed with Linear algebra operations
  - Markov chains have short memory
- Study example application of Markov Models
  - Google PageRank
- Extensions of Markov models
  - Parameters of higher order Markov chains increase exponentially
- Application: sample complex probability distributions using the convergence to the stationary distribution.

References I

Hastie, T., Tibshirani, R., and Friedman, J. (2009).
The elements of statistical learning: data mining, inference, and prediction.
Springer Science & Business Media.

Murphy, K. P. (2022).
Probabilistic Machine Learning: An introduction.
MIT Press.