Machine Learning for Graphs and Sequential Data

Sequential Data – Markov Chains

lecturer: Prof. Dr. Stephan Günnemann
www.daml.in.tum.de

Summer Term 2020

Data Analytics and
Machine Learning

Roadmap

- Chapter: Temporal Data / Sequential Data
  1. Autoregressive Models
  2. **Markov Chains**
  3. Hidden Markov Models
  4. Neural Network Approaches
  5. Temporal Point Processes

Markov Chains - Definition

A B A A C

- Definition: A **Markov Chain** is a sequence of r.v. $X_1, X_2, \dots, X_T$ which fulfills the **Markov property** :

$$P(X_t | X_1, \dots, X_{t-1}) = P(X_t | X_{t-1})$$

- The values taken by the time index t are discrete i.e. $t \in \{1,2, \dots, T\}$

- We assume that the r.v. X_t are discrete i.e. $X_t \in \{1,2, \dots, K\}$

- The joint distribution of a Markov Chain is:

$$\mathrm{P}(X_1 = i_1, \dots, X_T = i_T) = \mathrm{P}(X_1 = i_1) \prod_{t=1}^{T-1} \mathrm{P}(X_{t+1} = i_{t+1} | X_t = i_t)$$

Markov Chain – General case

$\pi : [0.9 \quad 0.01 \quad 0.01]$

A, B, C

$K = 3$

- In the general case, the distribution of each r.v. can be different:

$$\mathrm{P}(X_1 = i) = \pi_i \text{ and } \mathrm{P}(X_{t+1} = j | X_t = i) = A_{ij}^{(t+1)}$$

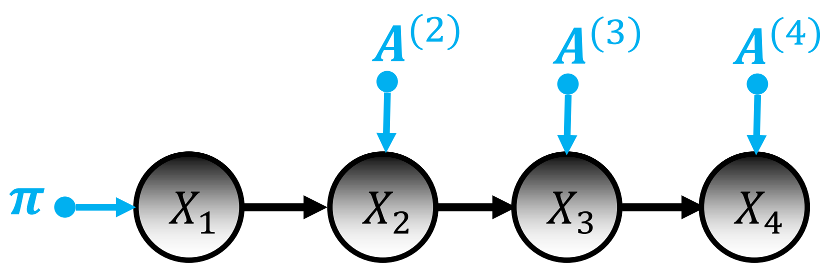
$$\begin{array}{c|ccc} & A & B & C \\ \hline A & \cdot & \cdot & \cdot \\ B & \cdot & \cdot & \cdot \\ C & \cdot & \cdot & \cdot \end{array}$$

where $\pi \in \mathbb{R}^K$ is a **prior probability** on the initial state, and $\boldsymbol{A}^{(t)} \in \mathbb{R}^{K \times K}$ are the **transition matrices**.

- Consequently the joint probability and the graphical model are:

$$\mathrm{P}(X_1 = i_1, \dots, X_T = i_T) = \pi_{i1} \times A_{i1,i2}^{(2)} \times \cdots \times A_{i_{T-1},i_T}^{(T)}$$

$$\boxed{\#\text{Parameters} = K + (T-1)\,K^2}$$



Markov Chain – Stationary case

$\tilde{A}$ B $\tilde{A}$ C C

- To simplify, we assume a **time-homogeneous** or **stationary** Markov Chain:

$$\mathrm{P}(X_1 = i) = \pi_i \text{ and } \mathrm{P}(X_{t+1} = j | X_t = i) = A_{ij}$$

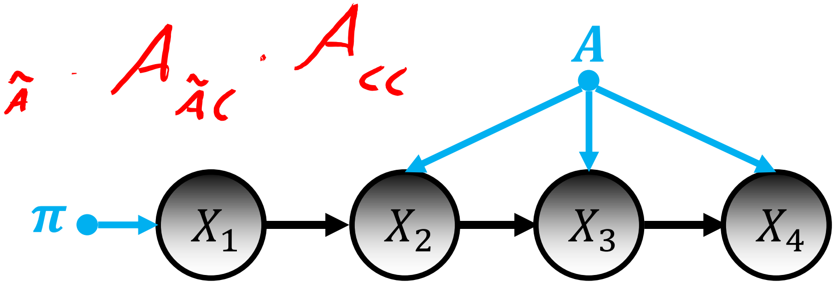
  – The transition matrix $\boldsymbol{A}^{(t)} = \boldsymbol{A}$ does not depend on t.
  All r.v. $X_2, \dots, X_T$ follow the same conditional distribution.

- The joint probability and the graphical model become:

$$\mathrm{P}(X_1 = i_1, \dots, X_T = i_T) = \pi_{i1} \times A_{i1,i2} \times \cdots \times A_{i_{T-1},i_T}$$

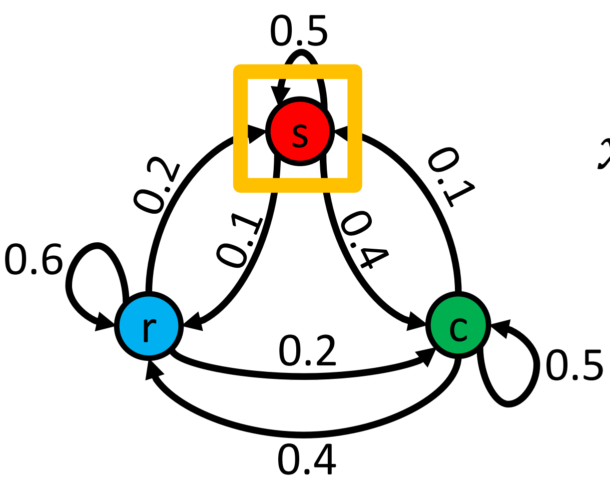
$$\pi_{\tilde{A}} \cdot A_{\tilde{A}B} \cdot A_{B\tilde{A}} \cdot A_{\tilde{A}C} \cdot A_{CC}$$

$\boxed{\#\text{Parameters} = K + K^2}$



Markov Chain – As a Random Walk

- Time-homogeneous discrete MCs can be interpreted as state machines
- Example: a model for weather condition
  - $X_t \in \{rainy, sunny, cloudy\}$ weather condition on t-th day
  - We can think of a sequence (i.e. a sample from the MC) as a random walk.



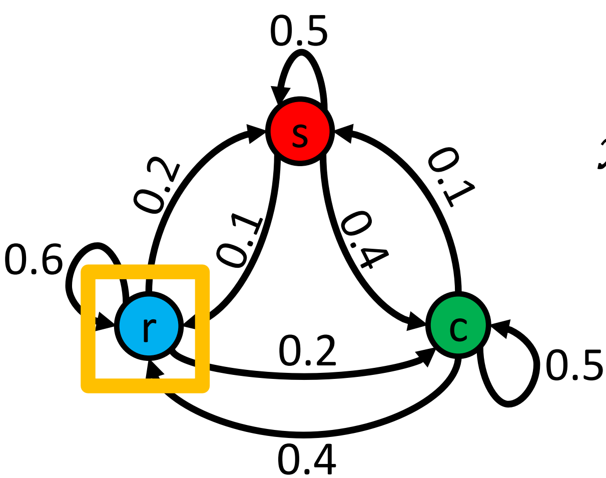
$$x_{1:T} = \boxed{s}\ r\ c\ r\ r\ c\ s$$

$$A = \begin{array}{r c} & \begin{array}{ccc} rainy & sunny & cloudy \end{array} \\ \begin{array}{r} rainy \\ sunny \\ cloudy \end{array} & \begin{bmatrix} 0.6 & 0.2 & 0.2 \\ 0.1 & 0.5 & 0.4 \\ 0.4 & 0.1 & 0.5 \end{bmatrix} \end{array}$$

$$\mathrm{P}(X_{1:T} = x_{1:T}) = \mathrm{P}(X_1 = s) \times 0.1 \times 0.2 \times 0.4 \times 0.6 \times 0.2 \times 0.1$$

Markov Chain – As a Random Walk

- Time-homogeneous discrete MCs can be interpreted as state machines
- Example: a model for weather condition
  - $X_t \in \{rainy, sunny, cloudy\}$ weather condition on t-th day
  - We can think of a sequence (i.e. a sample from the MC) as a random walk.



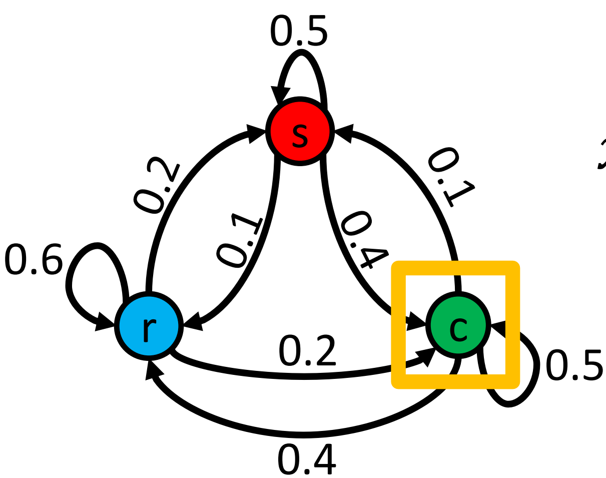
$$x_{1:T} = s \;\; \boxed{r} \;\; c \;\; r \;\; r \;\; c \;\; s$$

$$A = \begin{array}{c} \\ rainy \\ sunny \\ cloudy \end{array} \begin{array}{c} \begin{array}{ccc} rainy & sunny & cloudy \end{array} \\ \begin{bmatrix} 0.6 & 0.2 & 0.2 \\ 0.1 & 0.5 & 0.4 \\ 0.4 & 0.1 & 0.5 \end{bmatrix} \end{array}$$

$$\mathrm{P}(X_{1:T} = x_{1:T}) = \mathrm{P}(X_1 = s) \times 0.1 \times 0.2 \times 0.4 \times 0.6 \times 0.2 \times 0.1$$

Markov Chain – As a Random Walk

- Time-homogeneous discrete MCs can be interpreted as state machines
- Example: a model for weather condition
  - $X_t \in \{rainy, sunny, cloudy\}$ weather condition on t-th day
  - We can think of a sequence (i.e. a sample from the MC) as a random walk.



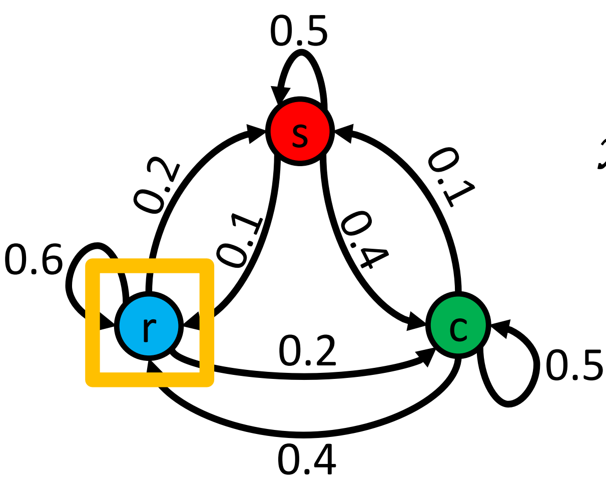
$$x_{1:T} = s \quad r \quad \boxed{c} \quad r \quad r \quad c \quad s$$

$$A = \begin{array}{r} \\ rainy \\ sunny \\ cloudy \end{array} \begin{array}{c} \begin{array}{ccc} rainy & sunny & cloudy \end{array} \\ \begin{bmatrix} 0.6 & 0.2 & 0.2 \\ 0.1 & 0.5 & 0.4 \\ 0.4 & 0.1 & 0.5 \end{bmatrix} \end{array}$$

$$\mathrm{P}(X_{1:T} = x_{1:T}) = \mathrm{P}(X_1 = s) \times 0.1 \times 0.2 \times 0.4 \times 0.6 \times 0.2 \times 0.1$$

Markov Chain – As a Random Walk

- Time-homogeneous discrete MCs can be interpreted as state machines
- Example: a model for weather condition
  - $X_t \in \{rainy, sunny, cloudy\}$ weather condition on t-th day
  - We can think of a sequence (i.e. a sample from the MC) as a random walk.



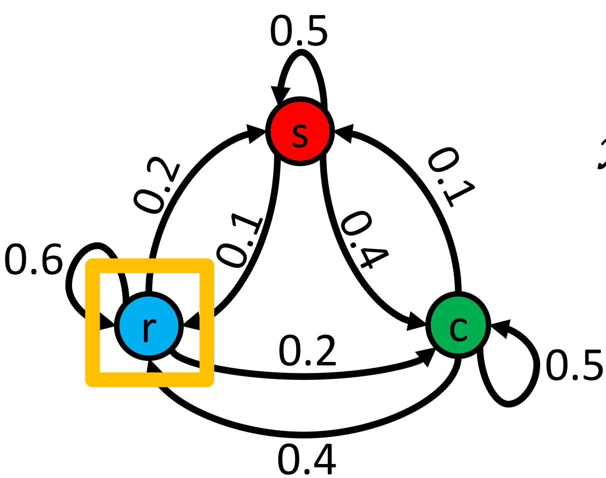
$$x_{1:T} = s \quad r \quad c \quad \boxed{r} \quad r \quad c \quad s$$

$$A = \begin{array}{r c} & \begin{array}{ccc} rainy & sunny & cloudy \end{array} \\ \begin{array}{r} rainy \\ sunny \\ cloudy \end{array} & \begin{bmatrix} 0.6 & 0.2 & 0.2 \\ 0.1 & 0.5 & 0.4 \\ 0.4 & 0.1 & 0.5 \end{bmatrix} \end{array}$$

$$\mathrm{P}(X_{1:T} = x_{1:T}) = \mathrm{P}(X_1 = s) \times 0.1 \times 0.2 \times 0.4 \times 0.6 \times 0.2 \times 0.1$$

Markov Chain – As a Random Walk

- Time-homogeneous discrete MCs can be interpreted as state machines
- Example: a model for weather condition
  - $X_t \in \{rainy, sunny, cloudy\}$ weather condition on t-th day
  - We can think of a sequence (i.e. a sample from the MC) as a random walk.



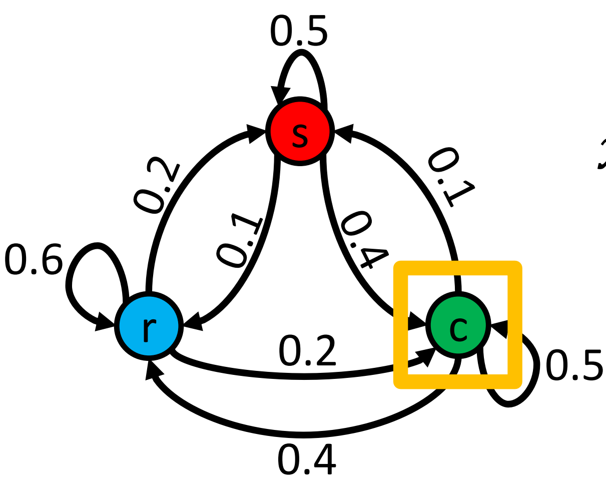
$$x_{1:T} = s \;\; r \;\; c \;\; r \;\; \boxed{r} \;\; c \;\; s$$

$$A = \begin{array}{r} \\ rainy \\ sunny \\ cloudy \end{array} \begin{array}{c} \begin{array}{ccc} rainy & sunny & cloudy \end{array} \\ \begin{bmatrix} 0.6 & 0.2 & 0.2 \\ 0.1 & 0.5 & 0.4 \\ 0.4 & 0.1 & 0.5 \end{bmatrix} \end{array}$$

$$\mathrm{P}(X_{1:T} = x_{1:T}) = \mathrm{P}(X_1 = s) \times 0.1 \times 0.2 \times 0.4 \times 0.6 \times 0.2 \times 0.1$$

Markov Chain – As a Random Walk

- Time-homogeneous discrete MCs can be interpreted as state machines
- Example: a model for weather condition
  - $X_t \in \{rainy, sunny, cloudy\}$ weather condition on t-th day
  - We can think of a sequence (i.e. a sample from the MC) as a random walk.



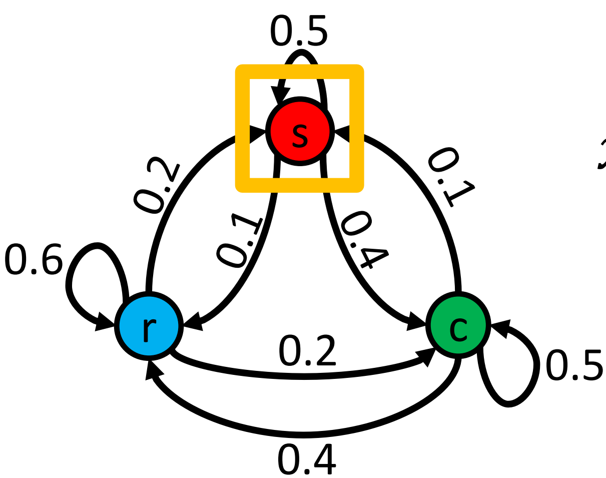
$$x_{1:T} = s \quad r \quad c \quad r \quad r \quad \boxed{c} \quad s$$

$$A = \begin{array}{r} \\ rainy \\ sunny \\ cloudy \end{array} \begin{array}{c} \begin{array}{ccc} rainy & sunny & cloudy \end{array} \\ \begin{bmatrix} 0.6 & 0.2 & 0.2 \\ 0.1 & 0.5 & 0.4 \\ 0.4 & 0.1 & 0.5 \end{bmatrix} \end{array}$$

$$\mathrm{P}(X_{1:T} = x_{1:T}) = \mathrm{P}(X_1 = s) \times 0.1 \times 0.2 \times 0.4 \times 0.6 \times 0.2 \times 0.1$$

Markov Chain – As a Random Walk

- Time-homogeneous discrete MCs can be interpreted as state machines
- Example: a model for weather condition
  - $X_t \in \{rainy, sunny, cloudy\}$ weather condition on t-th day
  - We can think of a sequence (i.e. a sample from the MC) as a random walk.



$$x_{1:T} = s \quad r \quad c \quad r \quad r \quad c \quad \boxed{s}$$

$$A = \begin{array}{r} \\ rainy \\ sunny \\ cloudy \end{array} \begin{array}{c} \begin{array}{ccc} rainy & sunny & cloudy \end{array} \\ \begin{bmatrix} 0.6 & 0.2 & 0.2 \\ 0.1 & 0.5 & 0.4 \\ 0.4 & 0.1 & 0.5 \end{bmatrix} \end{array}$$

$$\mathrm{P}(X_{1:T} = x_{1:T}) = \mathrm{P}(X_1 = s) \times 0.1 \times 0.2 \times 0.4 \times 0.6 \times 0.2 \times 0.1$$

$n(A,B)=1$ ABBAC

Markov Chain – Learning of Model Parameters

BBAACCCA

$$L(k) = \#(X_1 = k)$$
$$N(i,j) = \#(X_t = i, X_{t+1} = j)$$

$N(B,A)=2$

- Given a set $\{X_{1:T_n}^{(n)}\}$ of N observed sequences, we can learn $\boldsymbol{\pi}$ and $\boldsymbol{A}$ using maximum-likelihood.

$$\mathrm{P}(all) = \prod_{n=1}^{N} \mathrm{P}\left(X_1^{(n)}\right) \prod_{t=1}^{T_n-1} \Pr(X_{t+1}^{(n)} | X_t^{(n)}) \quad = \left(\prod_{k=1}^{K} \pi_k^{L(k)}\right)\left(\prod_{i=1}^{K}\prod_{j=1}^{K} A_{ij}^{N(i,j)}\right)$$

$$\Rightarrow \log P(all) = \sum_{k=1}^{K} L(k)\log(\pi_k) + \sum_{i=1}^{K}\sum_{j=1}^{K} N(i,j)\log(A_{ij})$$

- Minimizing $\log P(all)$ subject to $\sum_k \pi_k = 1$ and $\sum_j A_{ij} = 1$, we get:

$$A_{ij} = \frac{N(i,j)}{\sum_{j'} N(i,j')} \qquad \pi_k = \frac{L(k)}{\sum_{k'} L(k')} = \frac{L(k)}{N}$$

Markov Chain – More Insights

(1) $P(A) = \sum_b P(A, B=b)$

(2) $P(A,B|C) = P(A|B,C) \cdot P(B|C)$

- Task 1: Determine $A_{ij}(n) = \mathrm{P}(X_{t+n} = j \mid X_t = i)$
  - In words, $A_{ij}(n)$ = probability of getting from state i to state j in n steps
- How to compute $A_{ij}(n)$?

$$\mathrm{P}(X_{t+n} = j|X_t = i) \overset{(1)}{=} \sum_{k=1}^{K} \mathrm{P}(X_{t+n} = j, X_{t+n-1} = k \,|X_t = i)$$

$$\overset{(2)}{=} \sum_{k=1}^{K} \mathrm{P}(X_{t+n} = j \,|X_{t+n-1} = k \,, X_t = i)\, \mathrm{P}(X_{t+n-1} = k|X_t = i)$$

$$= \sum_{k=1}^{K} \mathrm{P}(X_{t+n} = j \,|X_{t+n-1} = k)\, \mathrm{P}(X_{t+n-1} = k|X_t = i) = \sum_{k=1}^{K} A_{kj}\, A_{ik}(n-1)$$

$$\Rightarrow \boldsymbol{A}(n) = \boldsymbol{A}(n-1)\boldsymbol{A} \xrightarrow{\boldsymbol{A}(1) = \boldsymbol{A}} \boldsymbol{A}(n) = \boldsymbol{A}^n$$

$= A(n-2)\,A\,A$

- Chapman-Kolmogorov equations:

$$A_{ij}(m+n) = \sum_{k=1}^{K} \mathrm{A}_{ik}(m) A_{kj}(n) \Rightarrow \quad \boldsymbol{A}(m+n) = \boldsymbol{A}(m)\boldsymbol{A}(n)$$

Markov Chain – More Insights

- Task 2: Determine $\pi_j(t) = \Pr(X_t = j)$
  - In words, $\pi_j(t)$ = probability of reaching state j in step t.
- How to compute $\pi_j(t)$?

$$\Pr(X_t = j) = \sum_{i=1}^{K} \Pr(X_t = j | X_{t-1} = i) \Pr(X_{t-1} = i) = \sum_{i=1}^{K} \mathrm{A}_{ij} \pi_i(t-1)$$

$$\Rightarrow \boldsymbol{\pi}(t) = \boldsymbol{\pi}(t-1)\boldsymbol{A}$$

$\boldsymbol{\pi}(t)$ and $\boldsymbol{\pi}$ are row vectors

$$\Rightarrow \boldsymbol{\pi}(t) = \boldsymbol{\pi}\, \boldsymbol{A}^{(t-1)}$$

Questions – MC

1. We assume that $X_t \in \{1, 2, 3\}$. We consider $\boldsymbol{\pi} = \begin{bmatrix} 0.0 \\ 0.5 \\ 0.5 \end{bmatrix}$ and $\boldsymbol{A} = \begin{bmatrix} 0.6 & 0.2 & 0.2 \\ 0.1 & 0.5 & 0.4 \\ 0.4 & 0.1 & 0.5 \end{bmatrix}$.
   a) What is the probability to observe the sequence $\boldsymbol{X}^{(1)} = [1, 2, 3]$?
   b) What is the probability to observe the sequence $\boldsymbol{X}^{(2)} = [2, 2, 3]$?

2. We assume that $X_t \in \{1, 2, 3\}$ and we observed three sequences:
   - $\boldsymbol{X}^{(1)} = [1, 3, 2]$
   - $\boldsymbol{X}^{(2)} = [3]$
   - $\boldsymbol{X}^{(3)} = [1, 1, 3, 2]$

   What is the MLE of the transition matrix $\boldsymbol{A} \in \mathbb{R}^{3\times 3}$?

Reading Material

- [1] Pattern Recognition and Machine Learning, section 13.1: https://www.microsoft.com/en-us/research/uploads/prod/2006/01/Bishop-Pattern-Recognition-and-Machine-Learning-2006.pdf