Machine Learning for Graphs and Sequential Data

Sequential Data – Markov Chains

Lecturer: Prof. Dr. Stephan Günnemann
cs.cit.tum.de/daml

Summer Term 2023

Roadmap

- Chapter: Temporal Data / Sequential Data

    1. Autoregressive Models

    2. **Markov Chains**

    3. Hidden Markov Models

    4. Neural Network Approaches

    5. Temporal Point Processes

Data Analytics and
Machine Learning

Markov Chains - Definition

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

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

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

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

- The joint distribution of a Markov Chain is:

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

Data Analytics and
Machine Learning

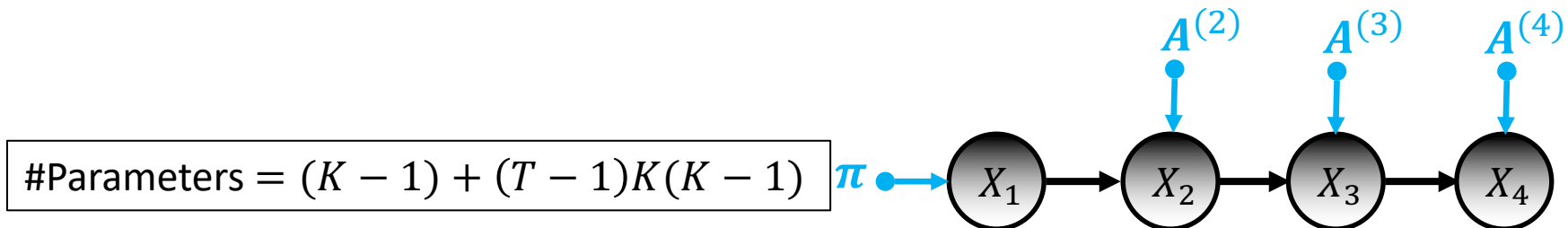# Markov Chain – General case

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

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

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:

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

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

Markov Chain – Stationary case

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

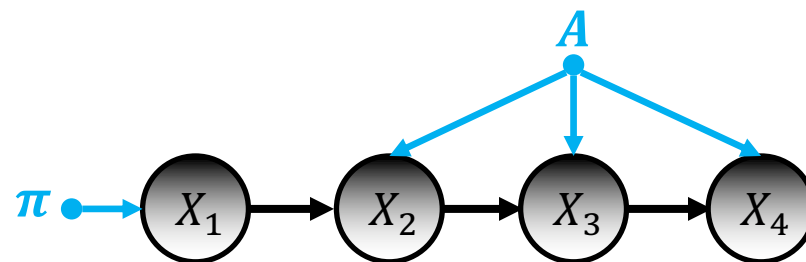$$P(X_1 = i) = \pi_i \text{ and } P(X_{t+1} = j | X_t = i) = A_{ij}$$

- 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.

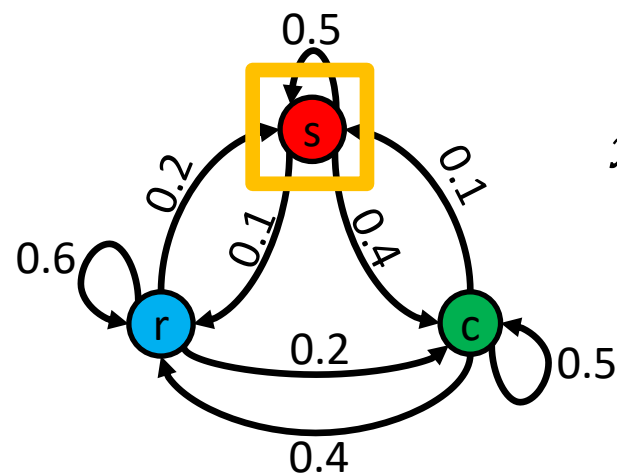- The joint probability and the graphical model become:

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

$$\boxed{\#\text{Parameters} = (K-1) + K(K-1)}$$

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.

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

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

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

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

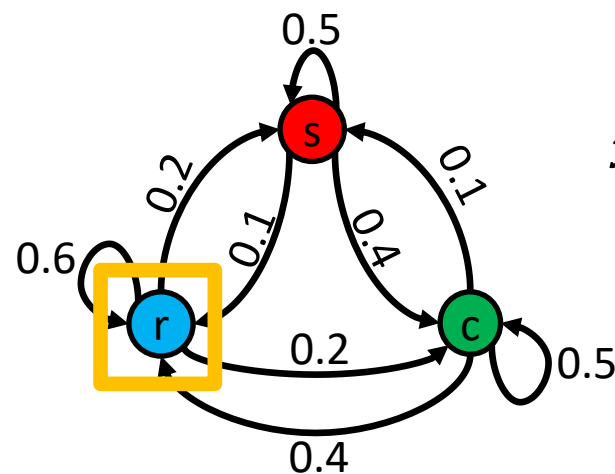 - We can think of a sequence (i.e. a sample from the MC) as a random walk.

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

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

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

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.

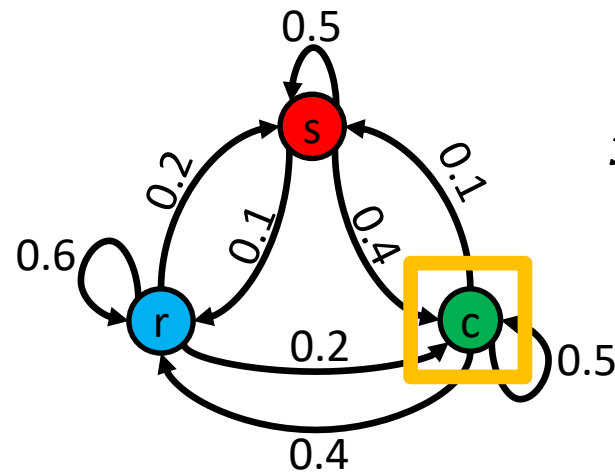$$x_{1:T} = s \quad r \quad \boxed{c} \quad r \quad r \quad c \quad s$$

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

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

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

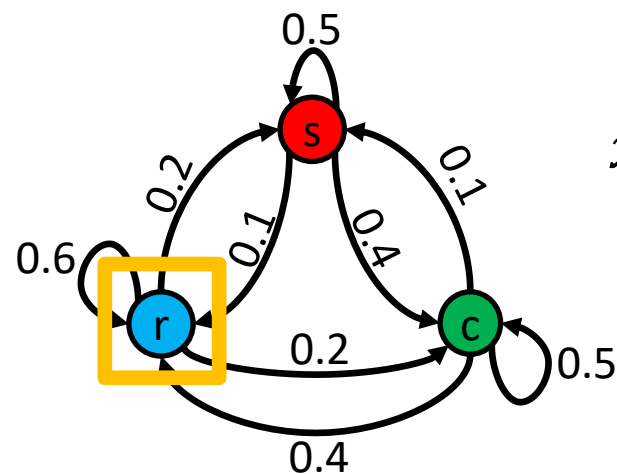 - We can think of a sequence (i.e. a sample from the MC) as a random walk.

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

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

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

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

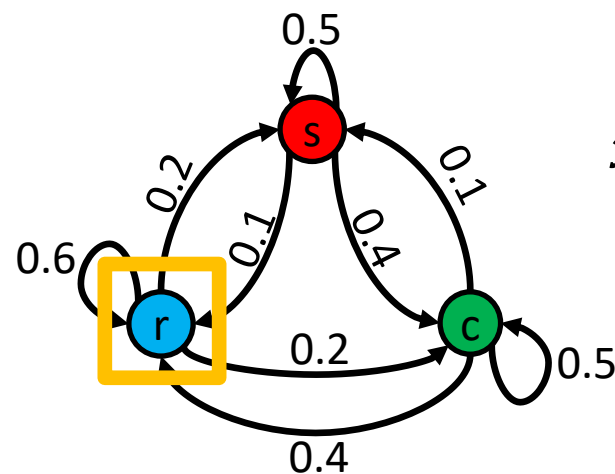 - We can think of a sequence (i.e. a sample from the MC) as a random walk.

$$x_{1:T} = \begin{array}{ccccccc} s & r & c & r & \boxed{r} & c & s \end{array}$$

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

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

Data Analytics and
Machine Learning

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

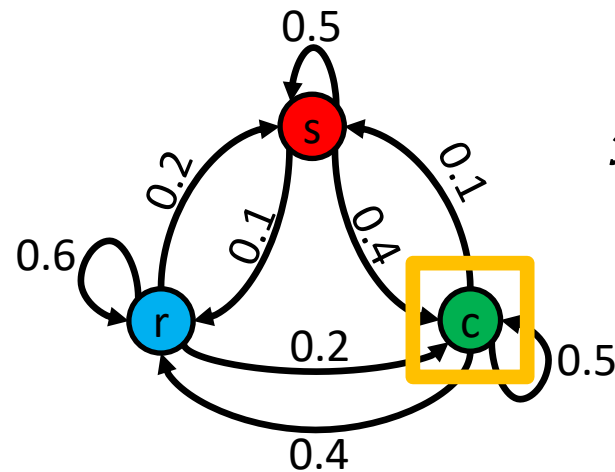 - We can think of a sequence (i.e. a sample from the MC) as a random walk.

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

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

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

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

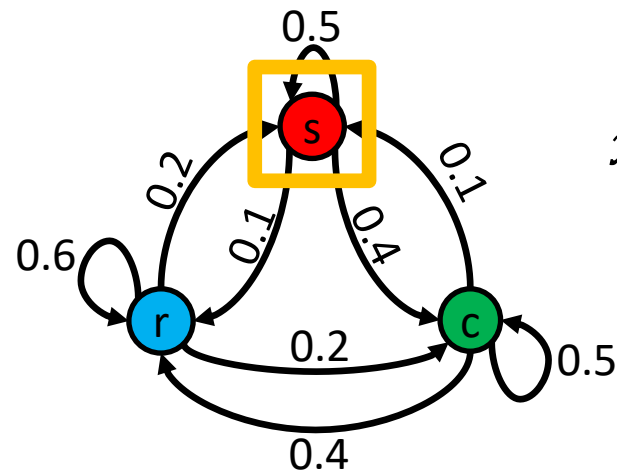 - We can think of a sequence (i.e. a sample from the MC) as a random walk.

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

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

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

Markov Chain – Learning of Model Parameters

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

$$\boxed{\begin{aligned} {\color{cyan}L(k) = \#(X_1 = k)} \\ {\color{red}N(i,j) = \#(X_t = i, X_{t+1} = j)} \end{aligned}}$$

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

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

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

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

Data Analytics and
Machine Learning

Markov Chain – More Insights

- 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 $A_{ij}(n)$?

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

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

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

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

- Chapman-Kolmogorov equations:

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

Data Analytics and
Machine Learning

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 $\pi_j(t)$?

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

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

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

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

Data Analytics and
Machine Learning

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 $\boldsymbol{A} \in \mathbb{R}^{3 \times 3}$?

Data Analytics and
Machine Learning

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

Data Analytics and
Machine Learning

Machine Learning for Graphs and Sequential Data

Sequential Data – Markov Chains

Lecturer: Prof. Dr. Stephan Günnemann
cs.cit.tum.de/daml

Summer Term 2023

Data Analytics and
Machine Learning

TUM