Machine Learning for Graphs and Sequential Data Exercise Sheet 06

Autoregressive Models, Markov Chains, Hidden Markov Models

Exercises marked with a $(*)$ will be discussed in the in-person exercise session.

Problem 1: Consider the stationary AR(p) process $X_t = c + \sum_{i=1}^{p} \phi_i X_{t-i} + \epsilon$ with $\epsilon \sim \mathcal{N}(0, \sigma^2)$. We denote by μ the mean $E[X_t]$ and by γ_i the autocovariance $Cov(X_t, X_{t-i})$. Show:

1. $\mu = \frac{c}{1-\sum_{i=1}^{p} \phi_i}$, for all t
2. $\gamma_0 = \sum_{j=1}^{p} \phi_j \gamma_{-j} + \sigma^2$
3. $\gamma_i = \sum_{j=1}^{p} \phi_j \gamma_{i-j}$, for all $t, i \in [1, p]$

Problem 2: $(*)$ Let $\mathbf{X}_t$ be a 2-D random vector:

$$\mathbf{X}_t = \begin{bmatrix} u_t \\ v_t \end{bmatrix} \quad \text{where } u_t, v_t \in \{1, 2, \ldots, K\}. \tag{1}$$

Consider the following Markov chain.

Model parameters are as follows:

- initial distribution $\boldsymbol{\pi}_x \in \mathbb{R}^{K \times K}$ that parametrizes $\Pr(\mathbf{X}_1)$:

$$\Pr\left(\mathbf{X}_1 = \begin{bmatrix} i \\ j \end{bmatrix}\right) = \boldsymbol{\pi}_x(i, j). \tag{2}$$

- transition probability matrix $\mathbf{A}_x \in \mathbb{R}^{K \times K \times K \times K}$ that parametrizes $\Pr(\mathbf{X}_{t+1} | \mathbf{X}_t)$:

$$\Pr\left(\mathbf{X}_{t+1} = \begin{bmatrix} i_{t+1} \\ j_{t+1} \end{bmatrix} \,\middle|\, \mathbf{X}_t = \begin{bmatrix} i_t \\ j_t \end{bmatrix}\right) = \mathbf{A}_x(i_t, j_t, i_{t+1}, j_{t+1}). \tag{3}$$

Because of the Markov property of $\mathbf{X}_t$, the joint probability can be factorized as

$$\Pr\left(\mathbf{X}_1, \ldots, \mathbf{X}_T\right) = \Pr\left(\mathbf{X}_1\right) \prod_{t=1}^{T-1} \Pr\left(\mathbf{X}_{t+1} | \mathbf{X}_t\right).$$

In this task, we refer to this model as *"2-D first-order Markov chain"*.

a) Does the sequence $[u_1, \ldots, u_T]$ (where $u_t \in \{1, 2, \ldots, K\}$ is defined in Eq. (1)) have the first-order Markov property? Why or why not?

b) Let $[Y_1, \ldots, Y_T] \in \{1, 2\}^T$ be a first-order Markov chain with initial probability distribution $\boldsymbol{\pi}_y \in \mathbb{R}^2$ and transition probabilities $\mathbf{A}_y \in \mathbb{R}^{2 \times 2}$.

Recap

AR: Learn using Pesudo - In // Yule - walke

ML: - Markov property $p(z_t | z_{1:t-1}) = p(z_t | z_{t-1})$

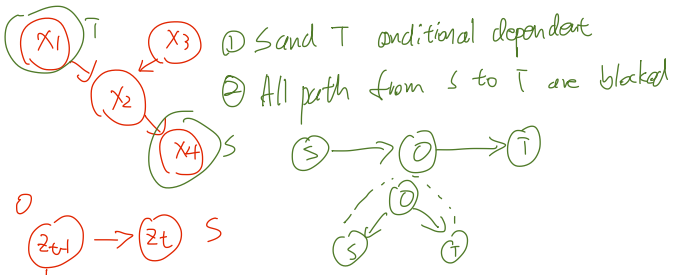
- MLE counting Transition prequential

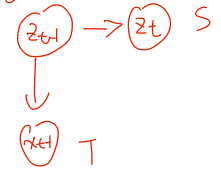
HMM: - Filtering $p(z_t | x_{1:t})$ = forward

- Smoothing $p(z_t | x_{1:T})$ = forward - backward

- MAP $\underset{z_{1:T}}{\text{argmax}}\ p(z_{1:T} | x_{1:T})$ vertebi

- Learning => EM

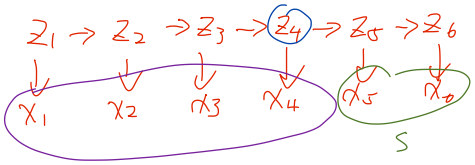




why $p(x_{t+1:T} | z_t = k, x_{1:t})$

$= p(x_{t+1:T} | z_t = k)$

$\neq p(x_{t+1:T} | x_{1:t})$



Problem 1: Consider the stationary AR(p) process $X_t = c + \sum_{i=1}^{p} \phi_i X_{t-i} + \epsilon$ with $\epsilon \sim \mathcal{N}(0, \sigma^2)$. We denote by μ the mean $E[X_t]$ and by γ_i the autocovariance $Cov(X_t, X_{t-i})$. Show:

1. $\mu = \frac{c}{1-\sum_{i=1}^{p} \phi_i}$, for all t
2. $\gamma_0 = \sum_{j=1}^{p} \phi_j \gamma_{-j} + \sigma^2$
3. $\gamma_i = \sum_{j=1}^{p} \phi_j \gamma_{i-j}$, for all $t, i \in [1, p]$

Stationary $\Rightarrow$
- $E[X_t] = E[X_{t-i}]$
- $cov(X_t, X_{t-i}) = \gamma_i = \gamma_{-i}$
- $E[|X_t|^2] < \infty$

1. For all t, $E(X_t) = E(X_{t-i})$

$E(X_t) = E(c) + \sum_{i=1}^{p} \phi_i E(X_{t-i}) + E(\epsilon_t)$

$= E(X_t)$

$E(X_t)\left[1 - \sum_{i=1}^{p} \phi_i\right] = c$

$E(X_t) = \mu = \frac{c}{1 - \sum_{i=1}^{p} \phi_i}$

2. $\gamma_0 = Cov(X_t, X_t) = Var(X_t) = \underbrace{var(c)}_{0} + \underbrace{var(\epsilon)}_{\sigma^2} + var\left(\sum_{i=1}^{p} \phi_i X_{t-i}\right)$

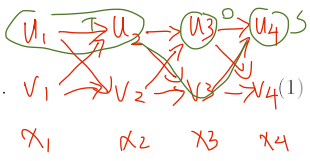
$= \sigma^2 + \sum_{j=1}^{p} \phi_j \underbrace{\sum_{i=1}^{p} \phi_i \frac{cov(X_i, X_{t-i})}{\gamma_j}}_{=\gamma_{-j}}$

3. $\gamma_i = Cov(X_t, X_{t-i}) = \underbrace{cov(c, X_{t-i})}_{0} + \underbrace{cov(\epsilon, X_{t-i})}_{0} + \sum_{i=1}^{p} \phi_i \frac{cov(X_{\tau-j}, X_{t-i})}{\gamma_{i-j}}$

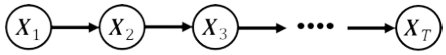
$= \sum_{j=1}^{p} \phi_j \gamma_{i-j}$

Problem 2: (*) Let $\mathbf{X}_t$ be a 2-D random vector:

$$\mathbf{X}_t = \begin{bmatrix} u_t \\ v_t \end{bmatrix} \quad \text{where } u_t, v_t \in \{1, 2, \ldots, K\}. \tag{1}$$



Consider the following Markov chain.



Model parameters are as follows:

- initial distribution $\boldsymbol{\pi}_x \in \mathbb{R}^{K \times K}$ that parametrizes $\Pr(\mathbf{X}_1)$:

$$\Pr\left(\mathbf{X}_1 = \begin{bmatrix} i \\ j \end{bmatrix}\right) = \boldsymbol{\pi}_x(i, j). \tag{2}$$

- transition probability matrix $\mathbf{A}_x \in \mathbb{R}^{K \times K \times K \times K}$ that parametrizes $\Pr(\mathbf{X}_{t+1} | \mathbf{X}_t)$:

$$\Pr\left(\mathbf{X}_{t+1} = \begin{bmatrix} i_{t+1} \\ j_{t+1} \end{bmatrix} \,\middle|\, \mathbf{X}_t = \begin{bmatrix} i_t \\ j_t \end{bmatrix}\right) = \mathbf{A}_x(i_t, j_t, i_{t+1}, j_{t+1}). \tag{3}$$

Because of the Markov property of $\mathbf{X}_t$, the joint probability can be factorized as

$$\Pr(\mathbf{X}_1, \ldots, \mathbf{X}_T) = \Pr(\mathbf{X}_1) \prod_{t=1}^{T-1} \Pr(\mathbf{X}_{t+1} | \mathbf{X}_t).$$

In this task, we refer to this model as *"2-D first-order Markov chain"*.

a) Does the sequence $[u_1, \ldots, u_T]$ (where $u_t \in \{1, 2, \ldots, K\}$ is defined in Eq. (1)) have the first-order Markov property? Why or why not?

b) Let $[Y_1, \ldots, Y_T] \in \{1, 2\}^T$ be a first-order Markov chain with initial probability distribution $\boldsymbol{\pi}_y \in \mathbb{R}^2$ and transition probabilities $\mathbf{A}_y \in \mathbb{R}^{2 \times 2}$.

- Briefly explain why the sequence $\begin{bmatrix} Y_2 \\ Y_1 \end{bmatrix}, \begin{bmatrix} Y_3 \\ Y_2 \end{bmatrix}, \ldots, \begin{bmatrix} Y_T \\ Y_{T-1} \end{bmatrix}$ is a 2-D first-order Markov chain.
- Compute initial and transition probabilities, $\boldsymbol{\pi}_x$ and $\mathbf{A}_x$ (defined in Eqs. (2) and (3)) for the sequence $\begin{bmatrix} Y_2 \\ Y_1 \end{bmatrix}, \begin{bmatrix} Y_3 \\ Y_2 \end{bmatrix}, \ldots, \begin{bmatrix} Y_T \\ Y_{T-1} \end{bmatrix}$.

1-D

$X_t = [u_t]$

$\Pr(X_1 = [u_1]) = \pi_x(u_1)$

$\Pr(X_2 = [u_2] \mid X_1 = [u_1]) = A_x$

$A_x \in R^{K \times K}$.

$\begin{bmatrix} u_{11} & u_{12} & & \\ & & u_{23} & \\ & & & \ddots \end{bmatrix}$

2-D,

$X_{t+1} = A \cdot X_t$.

$\begin{bmatrix} u_{t+1} \\ v_{t+1} \end{bmatrix} = A \cdot \begin{bmatrix} u_t \\ v_t \end{bmatrix}$

$u_{t+1} = A \cdot u_t + A \cdot v_t$.

not fulfill Markov property

$Y_i = \{[1], [2], \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \begin{bmatrix} 2 \\ 1 \end{bmatrix}, \begin{bmatrix} 2 \\ 2 \end{bmatrix}\}$

$Y_1 \to Y_2 \to Y_3 \to Y_4 \to Y_5$

$Y_{i=1}^{T}$ is first-order Markov chain

$p(Y_T \mid Y_{T-1}, \ldots, Y_1) = p(Y_T \mid Y_{T-1})$

$p\left(\begin{bmatrix} Y_T \\ Y_{T-1} \end{bmatrix} \middle| \begin{bmatrix} Y_{T-1} \\ Y_{T-2} \end{bmatrix} \cdots \begin{bmatrix} Y_2 \\ Y_1 \end{bmatrix}\right)$

because Y_T is only dependent on Y_{T-1}

$= p\left(\begin{bmatrix} Y_T \\ Y_{T-1} \end{bmatrix} \middle| \begin{bmatrix} Y_{T-1} \\ Y_{T-2} \end{bmatrix}\right) \Longleftrightarrow$ 2D-Markov chain

---

$\Pr\left(X_1 = \begin{bmatrix} Y_2 \\ Y_1 \end{bmatrix} = \begin{bmatrix} i \\ j \end{bmatrix}\right) = \pi_x(Y_2, Y_1) = \pi_x(i, j)$

$= \Pr(Y_2 = i \mid Y_1 = j) \cdot p(Y_1 = j)$

$= A_y(i, j) \cdot \pi_y(j)$

$\Pr\left(X_{t+1} = \begin{bmatrix} Y_T \\ Y_{T-1} \end{bmatrix} = \begin{bmatrix} i \\ j \end{bmatrix} \middle| X_t = \begin{bmatrix} Y_{T-1} \\ Y_{T-2} \end{bmatrix} = \begin{bmatrix} j \\ k \end{bmatrix}\right) = A_x$

$= \Pr(Y_T = i \mid Y_{T-1} = j, Y_{T-2} = k) \Pr(Y_{T-1} = j \mid Y_{T-1} = j, Y_{T-2} = k)$

$= A_y(i, j) \cdot A_y(j, k)$

$= 1$

- Briefly explain why the sequence $\begin{bmatrix} Y_2 \\ Y_1 \end{bmatrix}, \begin{bmatrix} Y_3 \\ Y_2 \end{bmatrix}, \ldots, \begin{bmatrix} Y_T \\ Y_{T-1} \end{bmatrix}$ is a 2-D first-order Markov chain.
- Compute initial and transition probabilities, $\boldsymbol{\pi}_x$ and $\mathbf{A}_x$ (defined in Eqs. (2) and (3)) for the sequence $\begin{bmatrix} Y_2 \\ Y_1 \end{bmatrix}, \begin{bmatrix} Y_3 \\ Y_2 \end{bmatrix}, \ldots, \begin{bmatrix} Y_T \\ Y_{T-1} \end{bmatrix}$.

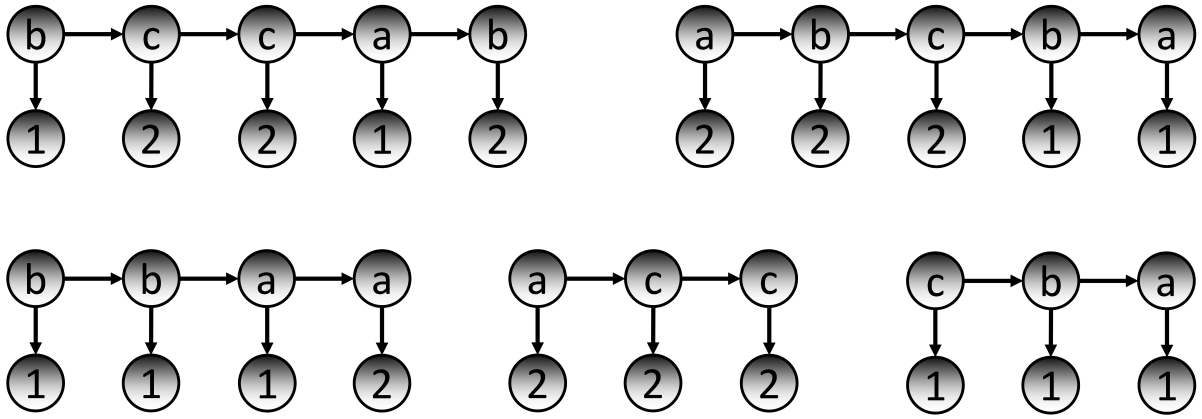
Problem 3: $(*)$ Consider an HMM where hidden variables are in $\{1, 2\}$ and observed variables are in $\{a, b, c\}$. Let the model parameters be as follows:

$$A = \begin{array}{c} \\ 1 \\ 2 \end{array} \begin{array}{c} \begin{array}{cc} 1 & 2 \end{array} \\ \begin{bmatrix} 0.2 & 0.8 \\ 0.5 & 0.5 \end{bmatrix} \end{array} \qquad B = \begin{array}{c} \\ 1 \\ 2 \end{array} \begin{array}{c} \begin{array}{ccc} a & b & c \end{array} \\ \begin{bmatrix} 0.2 & 0 & 0.8 \\ 0.4 & 0.6 & 0 \end{bmatrix} \end{array} \qquad \pi = \begin{array}{c} \\ 1 \\ 2 \end{array} \begin{array}{c} \\ \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix} \end{array}$$

Assume that the sequence $X_{1:5} = [cabac]$ is observed.

1. Filtering: find the distribution $P(Z_3|X_{1:3})$.
2. Smoothing: find the distribution $P(Z_3|X_{1:5})$.
3. Viterbi algorithm: find the most probable sequence $[Z_1, \ldots, Z_5]$.

Problem 4: Consider an HMM where states Z_t are in $\{a, b, c\}$ and emissions X_t are in $\{1, 2\}$. Given is the following set of fully-observed instances (two sequences of length 5, one sequence of length 4, and two sequences of length 3):



Learn the parameters of the HMM (i.e. $\pi \in \mathbb{R}^3, \mathbf{A} \in \mathbb{R}^{3\times3}$, and $\mathbf{B} \in \mathbb{R}^{3\times2}$) using maximum-likelihood estimation.

Problem 3: (*) Consider an HMM where hidden variables are in $\{1, 2\}$ and observed variables are in $\{a, b, c\}$. Let the model parameters be as follows:

$$A = \begin{array}{c} \\ 1 \\ 2 \end{array} \begin{array}{c} \begin{array}{cc} 1 & 2 \end{array} \\ \begin{bmatrix} 0.2 & 0.8 \\ 0.5 & 0.5 \end{bmatrix} \end{array} \quad B = \begin{array}{c} \\ 1 \\ 2 \end{array} \begin{array}{c} \begin{array}{ccc} a & b & c \end{array} \\ \begin{bmatrix} 0.2 & 0 & 0.8 \\ 0.4 & 0.6 & 0 \end{bmatrix} \end{array} \quad \pi = \begin{array}{c} \\ 1 \\ 2 \end{array} \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}$$

$Z_i \in \{1,2\} \Leftarrow$ Markov property

$X_i \in \{a,b,c\} \Leftarrow$ time only depend

Assume that the sequence $X_{1:5} = [cabac]$ is observed.

1. Filtering: find the distribution $P(Z_3|X_{1:3})$.
2. Smoothing: find the distribution $P(Z_3|X_{1:5})$.
3. Viterbi algorithm: find the most probable sequence $[Z_1, \ldots, Z_5]$.

$\alpha_1 = \pi \odot B_{:(x_1)}$

$\alpha_{t+1} = B_{:(x_{t+1})} \odot (A' \alpha_t)$

1. $P(Z_3 | X_{1:3}) = \frac{P(Z_3, X_{1:3})}{\sum_{j=1}^{k} P(X_{1:3}, Z_3 = j)}$ $\quad \alpha_3 = \begin{bmatrix} \alpha_3(1) \\ \alpha_3(2) \end{bmatrix}$

$\alpha_1(1) = P(Z_1 = 1, X_1)$

$= P(X_1 | Z_1 = 1) \cdot P(Z_1 = 1)$

$= P(X_1 = c | Z_1 = 1) P(Z_1 = 1)$

$= 0.8 \cdot 0.5 = 0.4$

$\alpha_1(2) = P(X_1 = c | Z_1 = 2) P(Z_1 = 2)$

$= 0$

$\alpha_2(1) = P(Z_2 = 1, X_{1:2})$

$= P(X_2 | Z_2 = 1, X_1) \cdot P(Z_2 = 1, X_1)$

$= P(X_2 = a | Z_2 = 1) \cdot \sum_{j=1}^{k} P(Z_2 = 1, Z_1 = j, X_1)$

$= 0.2 \cdot \sum_{j=1}^{k} P(Z_2 = 1 | Z_1 = j) \cdot P(Z_1 = j, X_1)$

$= 0.2 \cdot [P(Z_2 = 1 | Z_1 = 1) \cdot P(Z_1 = 1, X_1 = c) + P(Z_2 = 1 | Z_1 = 2) P(Z_1 = 2, X_1 = c)]$

$= 0.2 \cdot [0.2 \cdot 0.4 + 0]$

$= 0.016$

$\alpha_2(2) = P(X_2 = a | Z_2 = 2) \sum P(Z_2 = 2 | Z_1 = j) P(Z_1 = j, X_1)$

$= 0.4 [0.8 \cdot 0.4 + 0]$

$= 0.128$

$\alpha_3 = B_{:(X_3)} \odot A' \alpha_2$ $\quad A' \neq A$

$= \begin{bmatrix} 0 \\ 0.6 \end{bmatrix} \odot \left(\begin{bmatrix} 0.2 & 0.5 \\ 0.8 & 0.5 \end{bmatrix} \begin{bmatrix} 0.016 \\ 0.128 \end{bmatrix} \right) = \begin{bmatrix} 0 \\ 0.6 \end{bmatrix} \odot \begin{bmatrix} 0.672 \\ 0.768 \end{bmatrix} = \begin{bmatrix} 0 \\ 0.04608 \end{bmatrix}$

$P(Z_3 | X_{1:3}) = \frac{\alpha_3(1)}{\alpha_3(1) + \alpha_3(2)} = 0 \;//\; \frac{\alpha_3(2)}{---} = 1$

---

cabac

2. $P(Z_3 | X_{1:5})$ $\quad \beta_t = A (B_{:(x_{t+1})} \odot \beta_{t+1})$

$P(Z_t = k | X_{1:T}) \propto \alpha_t(k) \beta_t(k)$ $\quad A = \begin{array}{c} \\ 1 \\ 2 \end{array} \begin{array}{c} \begin{array}{cc} 1 & 2 \end{array} \\ \begin{bmatrix} 0.2 & 0.8 \\ 0.5 & 0.5 \end{bmatrix} \end{array}$ $\quad B = \begin{array}{c} \\ 1 \\ 2 \end{array} \begin{array}{c} \begin{array}{ccc} a & b & c \end{array} \\ \begin{bmatrix} 0.2 & 0 & 0.8 \\ 0.4 & 0.6 & 0 \end{bmatrix} \end{array}$ $\quad \pi = \begin{array}{c} \\ 1 \\ 2 \end{array} \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}$

$\beta_5 = 1$

$\beta_4 = \begin{bmatrix} 0.2 & 0.8 \\ 0.5 & 0.5 \end{bmatrix} \left(\begin{bmatrix} 0.8 \\ 0 \end{bmatrix} \odot \begin{bmatrix} 1 \\ 1 \end{bmatrix} \right)$

$= \begin{bmatrix} 0.16 \\ 0.4 \end{bmatrix}$

$\beta_3 = \begin{bmatrix} 0.2 & 0.8 \\ 0.5 & 0.5 \end{bmatrix} \left(\begin{bmatrix} 0.2 \\ 0.4 \end{bmatrix} \odot \begin{bmatrix} 0.16 \\ 0.4 \end{bmatrix} \right)$

$= \begin{bmatrix} 0.2 & 0.8 \\ 0.5 & 0.5 \end{bmatrix} \begin{bmatrix} 0.032 \\ 0.016 \end{bmatrix}$

$= \begin{bmatrix} 0.0192 \\ 0.024 \end{bmatrix}$

$P(Z_3 | X_{1:5}) = \frac{\alpha_t(k) \beta_t(k)}{\sum_s \alpha_t(s) \beta_t(s)} = \frac{0}{--} + \frac{0.04608 \cdot 0.024}{0 + \cdots} = 1$

$= \begin{bmatrix} 0 \\ 1 \end{bmatrix}$

3. $\arg\max_z p(z_{1:T} | x_{1:T}) = \arg\max_z \log p(z_{1:T}, x_{1:T}) = \arg\max_z \log p(x_1|z_1)p(z_1) + \sum \log p(x_t|z_t)p(z_t|z_{t-1})$

cabac

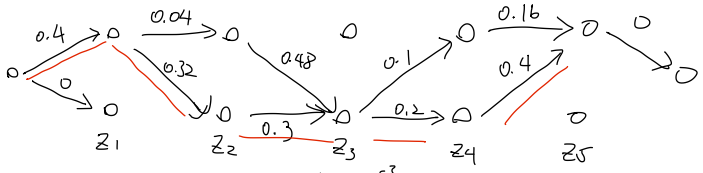
$-\log p(z_1=j)\, p(x_1|z_1=j)$

$= -\log\left[\frac{1}{2} \cdot 0.8\right]$

$-\log p(z_t=j|z_{t-1}=i)\, p(x_t|z_t=j)$

0

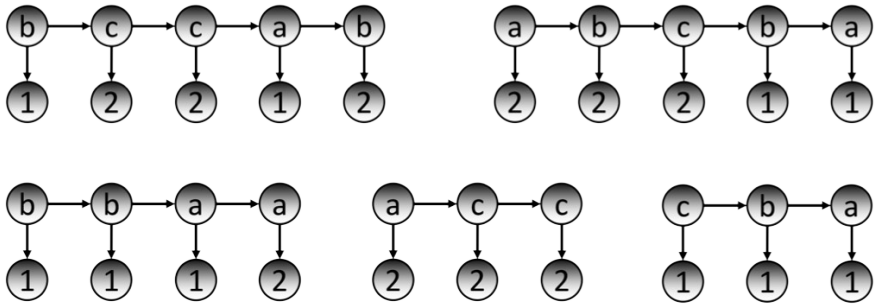
$$A = \begin{array}{c} \\ 1 \\ 2 \end{array}\begin{array}{c} \begin{array}{cc} 1 & 2 \end{array} \\ \begin{bmatrix} 0.2 & 0.8 \\ 0.5 & 0.5 \end{bmatrix} \end{array} \quad B = \begin{array}{c} \\ 1 \\ 2 \end{array}\begin{array}{c} \begin{array}{ccc} a & b & c \end{array} \\ \begin{bmatrix} 0.2 & 0 & 0.8 \\ 0.4 & 0.6 & 0 \end{bmatrix} \end{array} \quad \pi = \begin{array}{c} \\ 1 \\ 2 \end{array}\begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}$$



$0.4 \times 0.04 \times 0.48 = 7.68 \times 10^{-3}$ $\quad 0.1 \times 0.16 = 0.016$

$0.4 \times 0.32 \times 0.3 = 3.84 \times 10^{-2}$ $\quad 0.2 \times 0.4 = 0.08$

Problem 4: Consider an HMM where states Z_t are in $\{a, b, c\}$ and emissions X_t are in $\{1, 2\}$. Given is the following set of fully-observed instances (two sequences of length 5, one sequence of length 4, and two sequences of length 3):



Learn the parameters of the HMM (i.e. $\pi \in \mathbb{R}^3, \mathbf{A} \in \mathbb{R}^{3\times3}$, and $\mathbf{B} \in \mathbb{R}^{3\times2}$) using maximum-likelihood estimation.

$\max p(z_{1:T}, x_{1:T})$

$\max \log p(z_{1:T}, x_{1:T})$

$= \log p(z_1) + \sum \log p(z_t|z_{t-1}) + \sum \log p(x_t|z_t)$

$= \sum_{n=1}^{5} \sum_k \mathbb{I}(z_1 = k) \log \pi_k + \sum_{t=2}^{T} \sum_{i,j} \mathbb{I}(z_t = j, z_{t-1} = i) \cdot \log A_{ij} + \sum_{t=1}^{T} \sum_{i,j} \mathbb{I}(x_t = j', z_t = i') \cdot \log B_{i'j'}$

$\sum_k \sum_{n=1}^{5} \mathbb{I}(z_1 = k) \cdot \log \pi_k + \cdots$

① s.t. $\pi \in \mathbb{R}^K$ $\pi \in \Delta_{K-1}$ $\sum \pi = 1$ $\pi_k \geq 0$.

$\max_\pi \max_\lambda \sum_k \sum_{n=1}^{5} \mathbb{I}(z_1 = k) \log \pi_k + \lambda\left(\sum_k \pi_k - 1\right)$

$\frac{\partial}{\partial \pi_k} = \sum_{n=1}^{5} \mathbb{I}(z_1 = k) \cdot \frac{1}{\pi_k} + \lambda = 0. \Rightarrow \pi_k = -\frac{\sum_{n=1}^{5} \mathbb{I}(z_1 = k)}{\lambda} \propto \begin{pmatrix} \sum \mathbb{I}(z_1=1)/N \\ \sum \mathbb{I}(z_1=2)/N \\ \sum \mathbb{I}(z_1=3)/N \end{pmatrix}$

② $\sum_j A(i,j) = 1$

③ $\sum_{j'} B(i',j') = 1$