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 $t, i \in [1, p]$

The AR(p) process is stationary i.e. $E[X_t] = \mu$ and $Cov(X_t, X_{t-i}) = \gamma_i$.

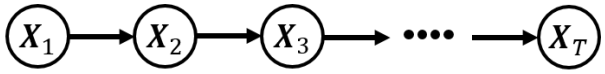
1. $\underbrace{E[X_t]}_{\mu} = c + \sum_{i=1}^{p} \phi_i \underbrace{E[X_{t-i}]}_{\mu} + \underbrace{E[\epsilon]}_{0} \Rightarrow \mu = \frac{c}{1-\sum_{i=1}^{p} \phi_i}$
2. $Cov(X_t, X_t) = \underbrace{Cov(c, X_t)}_{0} + \sum_{i=1}^{p} \phi_i \underbrace{Cov(X_{t-i}, X_t)}_{\gamma_{-i}} + \underbrace{Cov(\epsilon_t, X_t)}_{\sigma^2} \Rightarrow \gamma_0 = \sum_{j=1}^{p} \phi_j \gamma_{-j} + \sigma^2$

   We have that $Cov(\epsilon_t, X_t) = \sigma^2$, because only the noise is random and the noise between time steps is independent, meaning that $Cov(\epsilon_t, \epsilon_{t'}) = 0$ for $t \neq t'$.
3. $Cov(X_t, X_{t-i}) = \underbrace{Cov(c, X_{t-i})}_{0} + \sum_{j=1}^{p} \phi_i \underbrace{Cov(X_{t-j}, X_{t-i})}_{\gamma_{j-i}} + \underbrace{Cov(\epsilon_t, X_{t-i})}_{0} \Rightarrow \gamma_i = \sum_{j=1}^{p} \phi_j \gamma_{i-j}$

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

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

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?

The variable u_t depends on u_{t-1} and v_{t-1}. Moreover, v_{t-1} depends on u_{t-2}. So, u_t and u_{t-2} are not conditionally independent given u_{t-1} only. As a consequence, u_t is not a Markov chain.

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.

Since $(Y_t)_{t=1}^T$ is a markov chain, Y_t is independent of $Y_{t-2}, \ldots, Y_1$ conditioned on Y_{t-1}. This lets us decompose the conditional probability distribution over $[Y_t, Y_{t-1}]^T$ into

$$\Pr\left(\begin{bmatrix} Y_t \\ Y_{t-1} \end{bmatrix} \,\middle|\, Y_{t-1}, \ldots, Y_1\right) = \Pr\left(\begin{bmatrix} Y_t \\ Y_{t-1} \end{bmatrix} \,\middle|\, \begin{bmatrix} Y_{t-1} \\ Y_{t-2} \end{bmatrix}\right).$$

Combined with the fact that Y_t are discrete, binary values, the sequence in question is representable as 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 $\begin{bmatrix} Y_2 \\ Y_1 \end{bmatrix}, \begin{bmatrix} Y_3 \\ Y_2 \end{bmatrix}, \ldots, \begin{bmatrix} Y_T \\ Y_{T-1} \end{bmatrix}$.

We first compute $\boldsymbol{\pi}_x$:

$$\boldsymbol{\pi}_x(i,j) = \Pr\left(\begin{bmatrix} Y_2 \\ Y_1 \end{bmatrix} = \begin{bmatrix} j \\ i \end{bmatrix}\right) = \Pr(Y_2 = j \mid Y_1 = i) \Pr(Y_1 = i) = \mathbf{A}_y(i,j)\boldsymbol{\pi}_y(i)$$

Next we compute $\mathbf{A}_x$:

$$\mathbf{A}_x(i',j',i,j) = \Pr\left(\begin{bmatrix} Y_t \\ Y_{t-1} \end{bmatrix} = \begin{bmatrix} i \\ j \end{bmatrix} \,\middle|\, \begin{bmatrix} Y_{t-1} \\ Y_{t-2} \end{bmatrix} = \begin{bmatrix} i' \\ j' \end{bmatrix}\right) = \begin{cases} 0 & \text{if } j \neq i' \\ \mathbf{A}_y(j,i) & \text{otherwise.} \end{cases}$$

It is impossible for Y_{t-1} to take two different values, so a transition with $j \neq i'$ is impossible. Otherwise, we have indeed:

$$\begin{aligned} \Pr\left(\begin{bmatrix} Y_t \\ Y_{t-1} \end{bmatrix} = \begin{bmatrix} i \\ j \end{bmatrix} \,\middle|\, \begin{bmatrix} Y_{t-1} \\ Y_{t-2} \end{bmatrix} = \begin{bmatrix} j \\ j' \end{bmatrix}\right) &= \Pr(Y_t = i \mid Y_{t-1} = j, Y_{t-2} = j') \Pr(Y_{t-1} = j \mid Y_{t-1} = j, Y_{t-2} = j') \\ &= \Pr(Y_t = i \mid Y_{t-1} = j) \underbrace{\Pr(Y_{t-1} = j \mid Y_{t-1} = j)}_{=1} = \mathbf{A}_y(j,i) \end{aligned}$$

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:

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

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

1. Filtering: find the distribution $P(Z_3|X_{1:3})$.

We can use the forward algorithm to compute $\alpha_1, \alpha_2, \alpha_3$ and then normalize to obtain $P(Z_3|X_{1:3}) = \frac{\alpha_3}{\sum_{k'=1}^{2} \alpha_3(k')}$. Knowing the first observation $X_1 = c$, we compute α_1:

$$\alpha_1 = \pi \odot B_{:c} = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix} \odot \begin{bmatrix} 0.8 \\ 0 \end{bmatrix} = \begin{bmatrix} 0.4 \\ 0 \end{bmatrix}$$

Knowing the second observation $X_2 = a$, we compute α_2:

$$\alpha_2 = B_{:a} \odot (A^T \alpha_1) = \begin{bmatrix} 0.2 \\ 0.4 \end{bmatrix} \odot \left(\begin{bmatrix} 0.2 & 0.5 \\ 0.8 & 0.5 \end{bmatrix} \begin{bmatrix} 0.4 \\ 0 \end{bmatrix} \right) = \begin{bmatrix} 0.016 \\ 0.128 \end{bmatrix}$$

Knowing the third observation $X_3 = b$, we compute α_3:

$$\alpha_3 = B_{:b} \odot (A^T \alpha_2) = \begin{bmatrix} 0 \\ 0.6 \end{bmatrix} \odot \left(\begin{bmatrix} 0.2 & 0.5 \\ 0.8 & 0.5 \end{bmatrix} \begin{bmatrix} 0.016 \\ 0.128 \end{bmatrix} \right) = \begin{bmatrix} 0 \\ 0.04608 \end{bmatrix}$$

After normalization, we obtain $P(Z_3|X_{1:3}) = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$. A faster solution would be just to remark that it is impossible to observe $X_3 = b$ given that $Z_3 = 1$ by looking at the emission matrix. Hence, we should have $P(Z_3|X_{1:3}) = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$.

2. Smoothing: find the distribution $P(Z_3|X_{1:5})$.

We use the forward-backward algorithm to compute $\alpha_1, \alpha_2, \alpha_3$ and $\beta_5, \beta_4, \beta_3$ and then normalize $P(Z_3 = k|X_{1:5}) = \frac{\alpha_3(k)\beta_3(k)}{\sum_{k'=1}^{2} \alpha_3(k')\beta_3(k')}$. For the forward part, we re-use the results from filtering. For the backward algorithm, we can compute β_t:

$$\beta_5 = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \qquad \beta_4 = A(B_{:c} \odot \beta_5) = \begin{bmatrix} 0.16 \\ 0.4 \end{bmatrix} \qquad \beta_3 = A(B_{:a} \odot \beta_4) = \begin{bmatrix} 0.1344 \\ 0.096 \end{bmatrix}$$

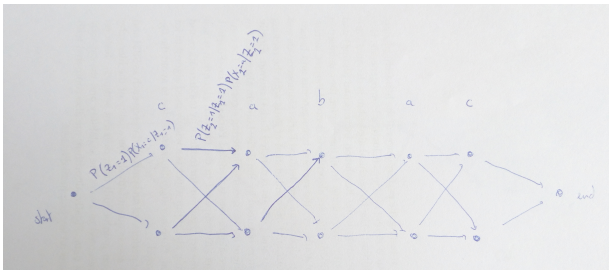
After normalization, we obtain $P(Z_3|X_{1:5}) = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$. A faster solution would be again to remark that it is impossible to observe $X_3 = b$ given that $Z_3 = 1$ by looking at the emission matrix.

3. Viterbi algorithm: find the most probable sequence $[Z_1, \ldots, Z_5]$.

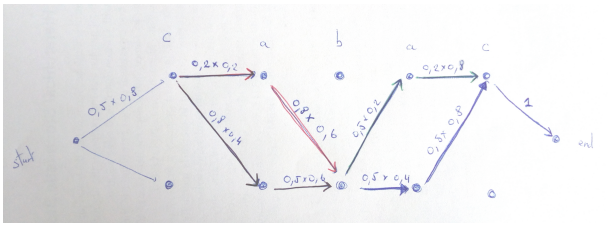
We want to solve:

$$[Z_1^*, \dots, Z_5^*] = \underset{[Z_1,\dots,Z_5]}{\arg\max}(P(Z_{1:5}|X_{1:5})) \Leftrightarrow [Z_1^*, \dots, Z_5^*] = \underset{[Z_1,\dots,Z_5]}{\arg\max}(P(Z_{1:5}, X_{1:5}))$$

We can formulate the problem as a longest path problem in the following graph:



In our case, we can remove all paths passing through $Z_t = 2, X_t = c$ or $Z_t = 1, X_t = b$ since these two configurations cannot happen according to the emission matrix. Hence, the probability of these paths are equal to 0.

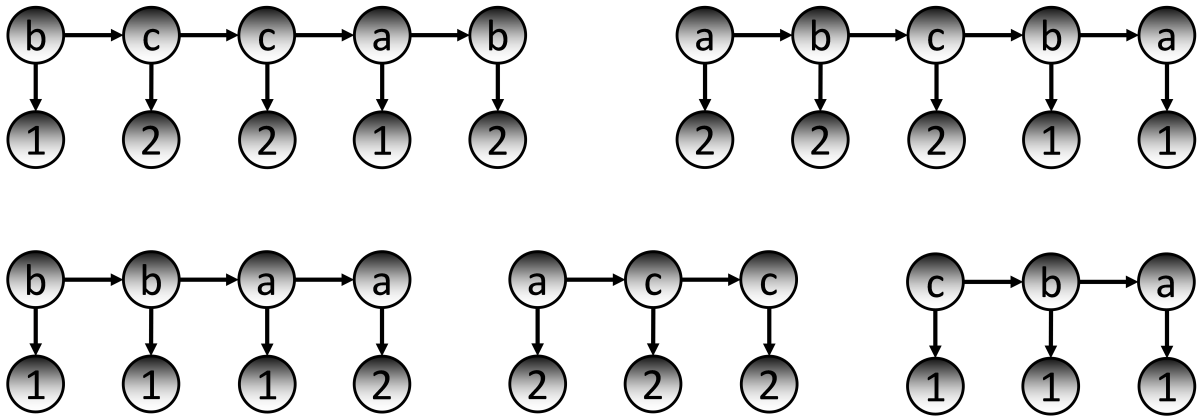


If we compare the four remaining paths, the optimal path is $[1, 2, 2, 2, 1]$ with length $0.5 \times 0.8 \times 0.8 \times 0.4 \times 0.5 \times 0.6 \times 0.5 \times 0.4 \times 0.5 \times 0.8$. Remark we can equivalently formulate the problem as a shortest path problem if we change edge weights to $-\log(P(Z_1)P(X_1|Z_1))$ and $-\log(P(Z_t|Z_{t-1})P(X_t|Z_t))$.

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.

We want to learn parameters of the HMM given 5 fully observed sequences. We can first compute the log-likelihood for one sequence s_1 of length T_1:

$$\log(P(s_1)) = \log P(Z_1) + \sum_{t=2}^{T_1} \log(P(Z_t|Z_{t-1})) + \sum_{t=1}^{T_1} \log(P(X_t|Z_t))$$



Using indicator functions and the parameters A, B and π, we can expand the log-likelihood:

$$= \sum_k \mathbf{I}(Z_1 = k) \log \pi(k) + \sum_{t=2}^{T_1} \sum_{i,j} \mathbf{I}(Z_t = j, Z_{t-1} = i) \log A(i,j)$$
$$+ \sum_{t=1}^{T_1} \sum_{i',j'} \mathbf{I}(X_t = j', Z_t = i') \log B(i',j')$$

For a set of independent observed sequences $D = \{s_i\}_{i=1}^5$, we can sum the log-likelihoods to get the complete data log-likelihood. It gives:

$$\log(P(D)) = \sum_{n=1}^5 \sum_k \mathbf{I}(Z_1 = k) \log \pi(k) + \sum_{n=1}^5 \sum_{t=2}^{T_n} \sum_{i,j} \mathbf{I}(Z_t = j, Z_{t-1} = i) \log A(i,j)$$
$$+ \sum_{n=1}^5 \sum_{t=1}^{T_n} \sum_{i',j'} \mathbf{I}(X_t = j', Z_t = i') \log B(i',j')$$
$$= \sum_k \underbrace{\sum_{n=1}^5 \mathbf{I}(Z_1 = k)}_{L(k)} \log \pi(k) + \sum_{i,j} \underbrace{\sum_{n=1}^5 \sum_{t=2}^{T_n} \mathbf{I}(Z_t = j, Z_{t-1} = i)}_{N(i,j)} \log A(i,j)$$
$$+ \sum_{i',j'} \underbrace{\sum_{n=1}^5 \sum_{t=1}^{T_n} \mathbf{I}(X_t = j', Z_t = i')}_{M(i',j')} \log B(i',j')$$

Now we maximize the data log-likelihood such that $\sum_k \pi(k) = 1$, $\sum_j A(i,j) = 1$, $\sum_{j'} B(i',j') = 1$. Encoding the equality constraints with Lagrange multipliers, we get $\pi^*(k) = \frac{L(k)}{\sum_k L(k)}$, $A^*(i,j) = \frac{N(i,j)}{\sum_j N(i,j)}$, $B^*(i,j) = \frac{M(i',j')}{\sum_j M(i',j')}$. So we can find the MLE parameters by counting observed initial states, transitions and emissions in the form of the matrices L, N and M and normalizing them

appropriately.

$$L = \begin{bmatrix} 2 \\ 2 \\ 1 \end{bmatrix} \quad N = \begin{bmatrix} 1 & 2 & 1 \\ 3 & 1 & 2 \\ 1 & 2 & 2 \end{bmatrix} \quad M = \begin{bmatrix} 4 & 3 \\ 5 & 2 \\ 1 & 5 \end{bmatrix}$$

Plugging these in we get the MLE parameters.

$$\pi^* = \begin{bmatrix} 2/5 \\ 2/5 \\ 1/5 \end{bmatrix} \quad A^* = \begin{bmatrix} 1/4 & 1/2 & 1/4 \\ 1/2 & 1/6 & 1/3 \\ 1/5 & 2/5 & 2/5 \end{bmatrix} \quad B^* = \begin{bmatrix} 4/7 & 3/7 \\ 5/7 & 2/7 \\ 1/6 & 5/6 \end{bmatrix}$$

If we had observed the X_t only, we could use the EM algorithm (see slide 26 + section 13.2.1 in Pattern Recognition and Machine Learning Book). In this case, we would consider γ_t and ξ_t (defined in slides 21, 22) which can be computed from the forward-backward algorithm (instead of L, N, M). We would have to alternate between (1) the forward-backward algorithm to compute γ_t, ξ_t given $\theta = \{\pi, A, B\}$, and (2) compute $\theta^{\text{new}} = \{\pi^{\text{new}}, A^{\text{new}}, B^{\text{new}}\}$ given γ_t, ξ_t.