SPR Day 17

day19 in 2023s

Markov models, Markov chains, and Hidden Markov Models

Reading: Bishop PRML Sec 13.1
Sec 13.2

Topics: - Motivation: Models w/ sequential dependence
- Markov assumptions (balance flexibility + tractability)
- Markov models (define distribution, parameters)
- Stationary distributions of Markov chains (BONUS content)
- Hidden Markov models
    - Definition of joint: $p(z_{1:T})p(x_{1:T}|z_{1:T})$
    - Analysis of marginal $p(x_{1:T})$ independence properties

Next time: How to train (estimate parameters via EM)

Unit 4 Motivation

Mixture models are very flexible distributions for explaining individual data examples $x_n \in \mathbb{R}^D$. However, mixtures we've used assume each example is independent of others given the mixture model parameters.

$$\underset{\substack{\uparrow \\ \text{dataset of } N \text{ examples}}}{p(X \mid \pi, \mu, \tau)} = \prod_{n=1}^{N} GMM(x_n \mid \pi, \mu, \tau)$$

order of $x_1, x_2, \dots x_N$ does NOT matter under iid assumption

Many real world analysis tasks would be questionable assuming x_n and x_{n+1} are independent. The order in which data are collected or observed often matters.

Examples:

Measuring weather at hourly intervals

Data: X_t is temp (°C) at hour t

Goal: Given $x_1, x_2, \dots x_T$, predict x_{T+1}

Predicting words in a texting app

Data: x_t is unigram of t-th word in sentence

Goal: Given "$\underset{x_1}{\text{The}}, \underset{x_2}{\text{state}}, \underset{x_3}{\text{of}}, \underset{x_4}{\text{Rhode}}$" predict x_5

Unit 4 Goal: Extend mixture models to case where sequence of observations matters

Model Assumptions Matter

Goal: Model for an ordered sequence of r.v.: $z_1, z_2, z_3 \cdots z_T$

We'll refer to each index t as a "timestep"

T is total length of sequence, $t \in \{1, 2, \ldots T\}$

A "model" is a joint distribution: $p(z_1, z_2, \cdots z_T)$

Consider a sequence of discrete random variables

$$z_1, \; z_2, \; z_3$$

e.g. $K=4$

$\Omega = \{a, b, c, d\}$

examples of length $T=3$

a, a, a ... d, d, a
a, a, b ... d, d, b
a, a, c ... d, d, c
a, a, d ... d, d, d

Each z_t is one of K possible values.

simple ⟵ spectrum of possible models ⟶ flexible

each outcome equally likely

$$p(z_1, z_2, z_3) = \frac{1}{K^3}$$

zero parameters to learn

each r.v. is iid

$$p(z_1 = j, z_2 = k, z_3 = l) = \theta_j \theta_k \theta_l$$

$K-1$ parameters to learn

$\theta \in \Delta^K$

each r.v. independent w/ own distribution

$$p(z_1 = j, z_2 = k, z_3 = l) = p(z_1 = j)\, p(z_2 = k)\, p(z_3 = l) = \theta_{1j} \theta_{2k} \theta_{3l}$$

$\theta_1 \in \Delta^K$
$\theta_2 \in \Delta^K$
$\theta_3 \in \Delta^K$

$3(K-1)$ params to learn

each outcome has own probability

$$p(z_1 = j, z_2 = k, z_3 = l) = \theta_{jkl}$$

$\theta \in \Delta^{K \cdot K \cdot K}$

$K^3 - 1$ parameters to learn

neither of these make sequential order matter

way too many parameters for long sequences

Compromise: Markov assumption

We want more flexibility than assuming each timestep is iid, but more simple than letting number of parameters grow with seq. length T.

First order Markov assumption:

z_{t+1} is conditionally independent of $z_1, z_2 \cdots z_{t-1}$ given z_t

For $T=3$, can always use product rule to write

$$p(z_1, z_2, z_3) = p(z_1)\, p(z_2 | z_1)\, p(z_3 | z_2, z_1)$$

Under Markov assumption,

$$p(z_3 | z_2, z_1) = p(z_3 | z_2) \;!$$

For large T, apply Markov assumption:

$$p(z_1, \dots z_T) = p(z_1) \underbrace{\prod_{t=2}^{T} p(z_t | z_{t-1})}_{\text{assume } 1^{st} \text{ order}}$$

(We'll focus on 1st order Markov, but note 2nd or higher order possible.)

Allowing separate parameters for each timestep, 1st order Markov assump. requires

$K-1$ params for $p(z_1)$

$K(K-1)$ params for $p(z_2 | z_1)$

$\vdots$

$K(K-1)$ params for $p(z_T | z_{T-1})$

$$\underbrace{(T-1)K(K-1)}_{t=2 \dots T} + \underbrace{K-1}_{t=1} \quad \text{total params.}$$

Make identical distribution assumption across time

$$p(z_2=k / z_1=j) = A_{jk}$$

$$p(z_3=k / z_2=j) = A_{jk}$$

same params A

$$\vdots$$

$$p(z_T=k / z_{T-1}=j) = A_{jk}$$

for all timesteps

total param count $= K(K-1)$ for $t \geq 2$

$+ K-1$ for $t=1$

Summary: Two key assumptions

(1) 1st order Markov: $p(z_{t+1} | z_t, z_{t-1}, \dots) = p(z_{t+1} | z_t$

(2) parameter sharing: $p(z_{t+1}=k / z_t=j) = p(z_2=k / z_1=j)$ $\forall t > 2$ $\forall j,k$

"all timesteps homogeneous"

achieve simple yet <u>tractable</u> model

1^{st} order Markov model

with homogeneous timesteps for discrete random variables

Random Variable: Sequence $z_1, z_2 \cdots z_T$
with each $z_t \in \{1, 2, \ldots K\}$
(Choose int indicators, not one-hot vectors)

Sample Space: All possible sequences of length T using the K symbols

Parameters:

$\pi \in \Delta^K$ initial timestep probabilities $\pi_k \triangleq p(z_1 = k)$

$A = \{A_j\}_{j=1}^K, \ A_j \in \Delta^K$ transition probabilities $A_{jk} \triangleq p(z_{t+1} = k \mid z_t = j)$

PMF:

$$p(z_1, z_2, \cdots z_T) = \pi_{z_1} \prod_{t=2}^{T} A_{z_{t-1}, z_t}$$

$$= p(z_1) \prod_{t=2}^{T} p(z_t \mid z_{t-1})$$

Exercises

Q: What is marginal $p(z_1)$?

$$p(z_1) = \sum_{z_T} \cdots \sum_{z_2} p(z_1, z_2, \cdots z_T)$$

by sum rule

Suppose $T=2$

$$p(z_1) = \sum_{z_2=1}^{K} p(z_1, z_2)$$

$$= \sum_{z_2} \pi_{z_1} A_{z_1 z_2} = \pi_{z_1} \sum_{k=1}^{K} A_{z_1,k}$$

$$= \pi_{z_1}$$

because sum of any row of A is one

Suppose $T=3$

$$p(z_1) = \sum_{j=1}^{K} \sum_{k=1}^{K} p(z_1, z_2=j, z_3=k)$$

$$= \sum_{j} \sum_{k} \pi_{z_1} A_{z_1,j} A_{j,k}$$

$$= \pi_{z_1} \cancelto{1}{\sum_{j} A_{z_1,j}} \cancelto{1}{\sum_{k} A_{jK}}$$

Sum of any row of A is one

$$= \pi_{z_1}$$

Exercises

Q: What is marginal $p(z_T)$?

Suppose $T=2$

$$p(z_2=k) = \sum_{j=1}^{K} p(z_1=j, z_2=k)$$

$$= \sum_{j=1}^{K} \pi_j A_{jk} = \pi^T A_{:k}$$

inner product of vector π and k^{th} column of A

Means that

$$z_2 \sim \text{Cat}(\pi^T A_{:1}, \dots \pi^T A_{:K})$$

Suppose $T=3$

$$p(z_3=k) = \sum_{a=1}^{K} \sum_{b=1}^{K} p(z_1=a, z_2=b, z_3=k)$$

$$= \sum_{b} \sum_{a} \pi_a A_{ab} A_{bk} = \pi^T A A_{:k}$$

General case

$$p(z_T=k) = \pi^T (A \cdot A \cdot A \cdots A) A_{:k} = \pi^{\overset{\text{transpose}}{T}} A^{\overset{\text{raise to power}}{(T)}} A_{:k}$$

Hidden Markov Model

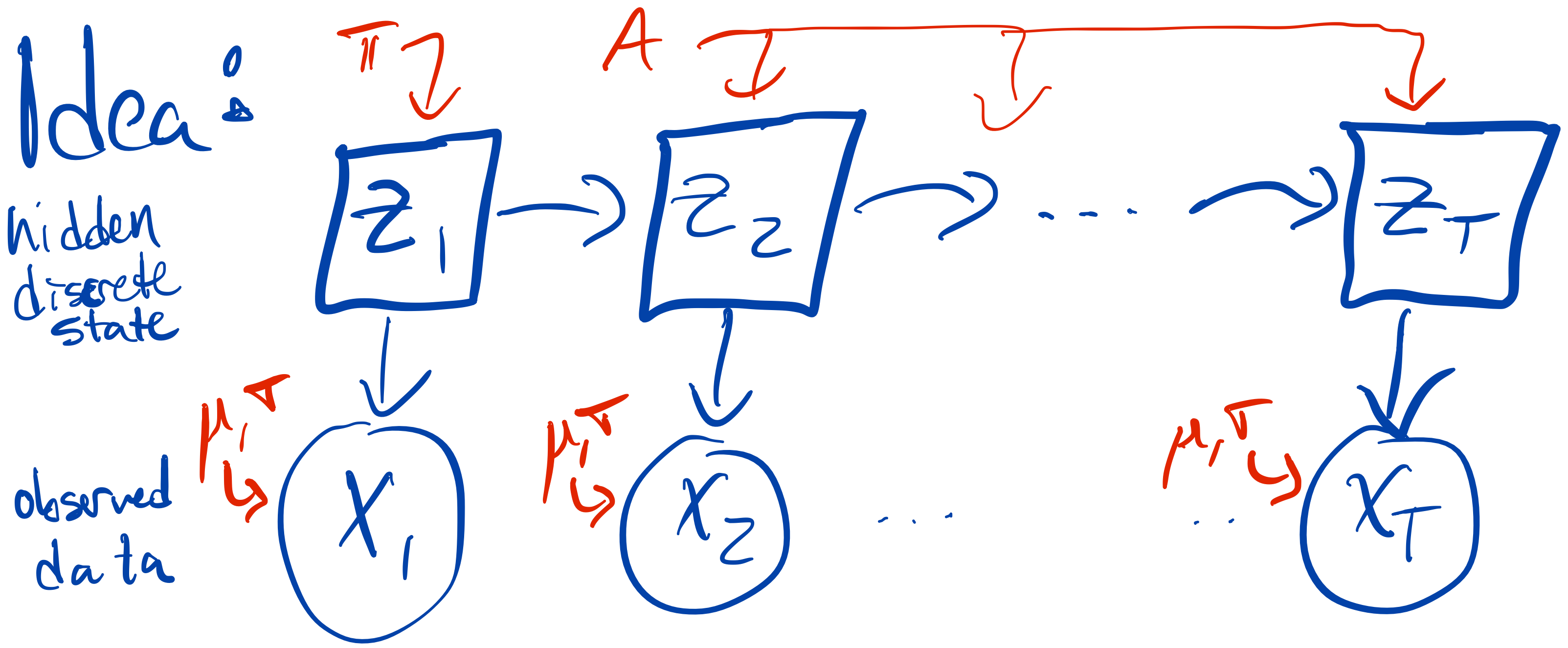
Goal: Tractable model for observed data sequence $x_1, x_2, \cdots x_T$ with $x_t \in \mathbb{R}$ that has <u>dependence</u> between x_t and $x_1, x_2 \cdots x_{t-1}$ but affordable to do 2 tasks:

(1) Compute data likelihoods

- joint $p(x_1, x_2 \cdots x_T | \theta)$
- conditional $p(x_T | x_1, x_2 \cdots x_{T-1}, \theta)$

(2) Estimate parameters θ via penalized Maximum likelihood

Idea:



HMM defines joint over x, z :

$$p(x_{1:T}, z_{1:T}) = \underbrace{p(z_{1:T})}_{\substack{\text{1st order} \\ \text{Markov} \\ \text{model}}} \underbrace{p(x_{1:T} | z_{1:T})}_{\substack{\text{each timestep iid given } z_t \\ \text{just like } p(x|z) \text{ term in mixture model}}}$$

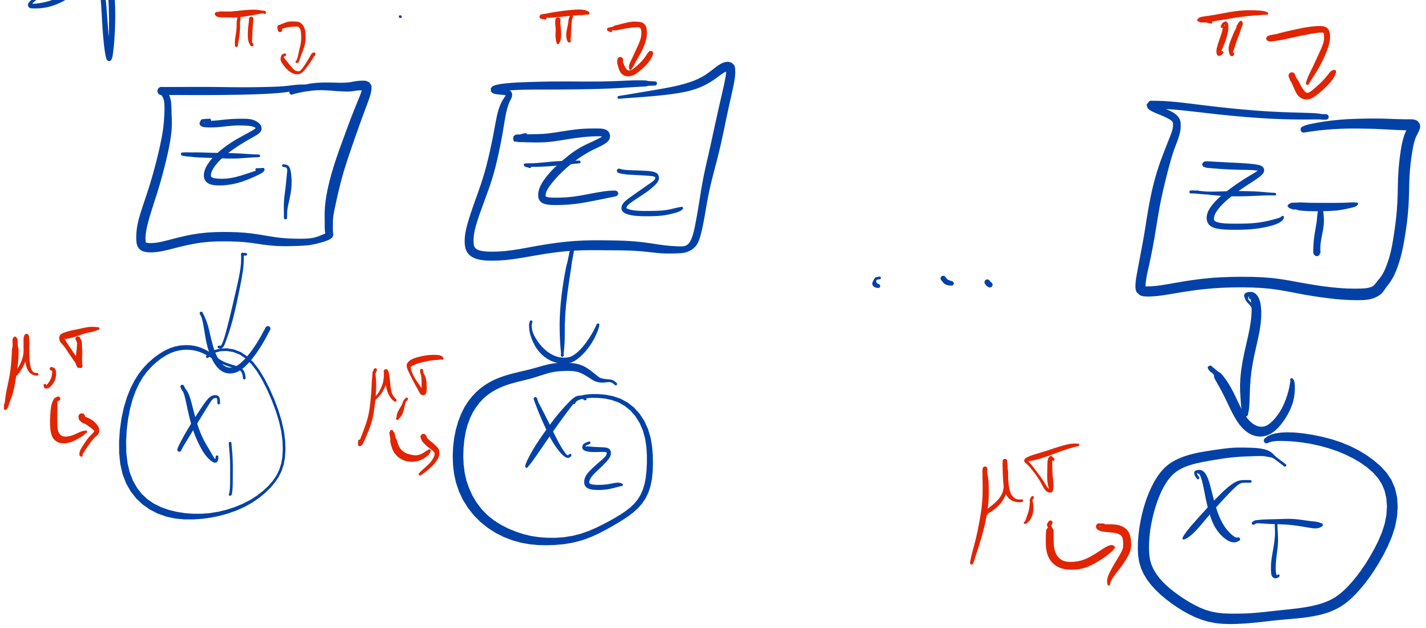
$$= \left[p(z_1) \prod_{t=2}^{T} p(z_2 | z_1) \right] \left[\prod_{t=1}^{T} p(x_t | z_t) \right]$$

$$= \left[\pi_{z_1} \prod_{t=2}^{T} A_{z_{t-1}, z_t} \right] \left[\prod_{t=1}^{T} \text{NormPDF}(x_t | \mu_{z_t}, \sigma_{z_t}^2) \right]$$

with parameters

$$\theta = \{ \underbrace{\pi, A}_{\text{used for } p(z)}, \underbrace{\mu, \sigma}_{\text{used for } p(x|z)} \}$$

Special case: mixture model



defines joint over x, z

$$p(x_{1:T}, z_{1:T}) = \prod_{t=1}^{T} \pi_{z_t} \prod_{t=1}^{T} \text{Norm}(x_t \mid \mu_{z_t}, \sigma^2_{z_t})$$

Can make an HMM into a mixture by setting

$$A_{1:} = \pi$$
$$A_{2:} = \pi$$
$$\vdots$$
$$A_{K:} = \pi$$

so $p(z_t = k \mid z_{t-1} = j) = A_{jk} = \pi_k$