HW5: Hidden Markov Models

Last modified: 2023-04-17 21:56

Status: **RELEASED.**

Due date: Thu Apr 20 at 11:59pm ET **with 4 free late days**

How to turn in: Submit PDF to https://www.gradescope.com/courses/496674/assignments/2830695/

Jump to: Problem 1 Problem 2

Questions?: Post to the **hw5** topic on the Piazza discussion forums.

Instructions for Preparing your PDF Report

What to turn in: PDF of typeset answers via LaTeX. No handwritten solutions will be accepted, so that grading can be speedy and you get prompt feedback.

Please use provided LaTeX Template: https://github.com/tufts-ml-courses/cs136-23s-assignments/blob/main/unit5_HW/hw5_template.tex

Your PDF should include (in order):

- Cover page with your full name, estimate of hours spent, and Collaboration statement
- Problem 1 answer
- Problem 2 answer

When you turn in the PDF to gradescope, mark each part via the in-browser Gradescope annotation tool)

Problem 1: Independence Assumptions for HMMs

Background: Assume we have T timesteps, with each timestep indexed by $t \in \{1, 2, \dots T\}$. Consider a Hidden Markov Model with K discrete states.

This HMM defines a distribution over two *sequences* of random variables:

- a discrete state sequence $z_{1:T} = [z_1, z_2, \dots z_T]$, where each value is an integer indicator $z_t \in \{1, 2, \dots K\}$
- a observed data sequence $x_{1:T} = [x_1, x_2, \dots x_T]$, where each value is a measured feature vector x_t (could be univariate or multivariate, discrete or continuous)

In order to specify the joint distribution over $z_{1:T}, x_{1:T}$, the HMM makes two key *conditional independence* assumptions:

$$\begin{aligned} &\text{HMM assumption A:} && p(z_{t+1} | z_t, z_{t-1}, \dots z_1) = p(z_{t+1} | z_t) \qquad \text{for } t \in 1, 2, 3, \dots T-1 \\ &\text{HMM assumption B:} && p(x_t | z_t, z_{1:t-1}, z_{t+1:T}, x_{1:t-1}, x_{t+1:T}) = p(x_t | z_t), \qquad \text{for } t \in 1, 2, \dots T \end{aligned}$$

In words, the first assumption (A) says the state at time $t+1$, given the state at time t, is conditionally independent of all other state variables before time t. This is the first-order Markov assumption.

The second assumption (B) says that given the hidden state at time t, the observation at time t is conditionally independent of all other variables in the model.

1a: Prove the following property for all timesteps $t \geq 1$. Remember to provide a short verbal justification for every step.

$$p(z_{t+1} | x_t, z_t) = p(z_{t+1} | z_t)$$

You can only use the following transformations: sum rule, product rule, Bayes rule, property A above, and property B above.

1b: Prove the following property for all timesteps $t \geq 1$. Remember to provide a short verbal justification for every step.

$$p(x_{t+1} | x_{1:t}, z_{1:t}) = p(x_{t+1} | z_t)$$

You can only use the following transformations: sum rule, product rule, Bayes rule, property A above, and property B above.

Problem 2: Understanding EM for HMMs with binary observations

Suppose have a sequence $x_{1:T} = x_1, x_2, \ldots, x_T$ of T binary vectors, where x_t is a D-length binary vector $x_t = [x_{t1}, x_{t2}, \ldots, x_{td} \ldots x_{tD}]$. At each timestep t and feature dimension d, you have a scalar binary value: $x_{td} \in \{0, 1\}$.

You wish to model this sequence using a hidden Markov model with K states. This HMM has parameters $\theta = \{\pi, A, \varphi\}$, where π is a K-length vector that sums to one, and A is a $K \times K$ matrix whose rows sums to one. Your model assumes the following joint distribution:

$$p(z_{1:T}, x_{1:T}|\theta) = p(z_{1:T}|\pi, A)p(x_{1:T}|z_{1:T}, \varphi)$$

Probabilistic model

The $z_{1:T}$ are generated by a Markov model:

$$\begin{aligned} p(z_{1:T}|\pi, A) &= \text{CatPMF}(z_1|\pi) \cdot \prod_{t=2}^{T} \text{CatPMF}(z_t|A_{z_{t-1}}) \\ &= \prod_{k=1}^{K} \pi_k^{\delta(z_1,k)} \cdot \prod_{t=2}^{T} \prod_{j=1}^{K} \prod_{k=1}^{K} A_{jk}^{\delta(z_{t-1},j)\delta(z_t,k)} \end{aligned}$$

Here, we'll use the notation $\delta(a, b)$ to be a binary indicator that is 1 if $a == b$ is true, and 0 otherwise.

Remember that the vector $[\delta(z_t, 1)\ \delta(z_t, 2)\ \ldots \delta(z_t, K)]$ is **one hot**, meaning exactly one of the K entries will be 1.

Given the $z_{1:T}$, each x_t is drawn iid from a multivariate Bernoulli given that timestep t's assigned state z_t

$$p(x_{1:T}|z_{1:T}, \varphi) = \prod_{t=1}^{T} \prod_{d=1}^{D} \prod_{k=1}^{K} \text{BernPMF}(x_{td}|\varphi_{kd})^{\delta(z_t,k)}$$

Here, the parameter φ_{kd} is the probability that the binary value of dimension d of vector x_t will be "on" or "1", if generated when time t assigned to state k. The value of φ_{kd} must be a valid Bernoulli parameter: $0 \le \varphi_{kd} \le 1$.

Approximate posterior

We have defined an "approximate posterior" distribution $q(z_{1:T}|s)$ over our state sequence, with learnable parameters s. The parameters $s = \{s_t\}_{t=1}^{T-1}$ specify joint distributions over each adjacent pair z_t, z_{t+1}, and of course must satisfy the constraints that neighboring marginals are the same.

For full details, see the notes from in-class about HMMs: https://www.cs.tufts.edu/cs/136/2023s/notes/day20.pdf#page=4

But for this problem, you just need to be able to use s to evaluate the following expectations:

$$\mathbb{E}_{q(z_{1:T}|s)}[\delta(z_t, k)] = r_{tk}(s), \qquad r_{tk} = \sum_{j=1}^{K} s_{tjk} = \sum_{\ell=1}^{K} s_{tk\ell}$$

$$\mathbb{E}_{q(z_{1:T}|s)}[\delta(z_t, j)\delta(z_{t+1}, k)] = s_{tjk}$$

where again, the notation $\delta(a, b)$ is a binary indicator that is 1 if $a == b$ is true, and 0 otherwise.

Problems

2a: Expected log likelihood Write out an expression for the expected complete log likelihood:

$$\mathbb{E}_{q(z_{1:T}|s)}\left[\log p(z_{1:T}, x_{1:T}|\theta)\right]$$

Use the HMM probabilistic model $p(z_{1:T}, x_{1:T}|\theta)$ and the approximate posterior $q(z_{1:T}|s)$ defined above.

Your answer should be a function of the data x, the local sequence parameters s and $r(s)$, as well as the HMM parameters π, A, φ.

2b: Deriving the M-step for data-per-state parameters Using your objective function from 2a above, show that for the M-step optimal update to the Bernoulli parameters φ_{kd}, the optimal update is given by:

$$\varphi_{kd} = \frac{\sum_{t=1}^{T} r_{tk} x_{td}}{\sum_{t=1}^{T} r_{tk}}$$

2c: Explaining the M-step for transition parameters You can find out (by looking up in your textbook) that the optimal update for each entry of A is given by:

$$A_{jk} = \frac{\sum_{t=1}^{T-1} s_{tjk}}{\sum_{t=1}^{T-1} \sum_{k=1}^{K} s_{tjk}}, \quad \text{for } j \in \{1, \dots, K\}, k \in \{1, \dots, K\}$$

Provide a short verbal summary of the update for A. How should we interpret the numerator? The denominator?

Hint: In 2c, we're not looking for any proof, just your ability to interpret the provided math in plain English.
