Welcome

COMP 136

Day 19: Viterbi for HMMs

Agenda:
* CP3 recap
* Midterm recap

---- midterm_solutions.pdf on Piazza

* Q&A on Viterbi algorithm

Upcoming Due Dates

- CP3 deadline extended to **Tue Apr 7 at NOON ET**
- Quiz3 released tomorrow Tue Apr 7, due **Thu** Apr 9 11:59pm ET

---- Will be entirely online via gradescope

---- Time limit: 20 minutes

---- all multiple choice

HW4 released later today (due next week)

CP4 released in next day or so (will be pushed back a few days, stay tuned)

CP3 recap

$$r_{nk} \triangleq p(z_n = k \mid x_n) = \frac{p(x_n, z_n = k)}{p(x_n)}$$

- Autograd debugging tips? See @168
- Getting nans in your estep? See **logsumexp** tips in @183
- Getting nans in your entropy calculation? See @170

'import scipy.special.logsumexp

$$r_{nk} \leftarrow \frac{\pi_k N(x_n \mid \mu_k, \Sigma_k)}{\sum_l \pi_l N(x_n \mid \mu_l, \Sigma_l^2)}$$

$$\log r_{nk} = a_{nk} - \log \sum_l e^{a_{nl}}$$

$$r_{nk} = e^{a_{nk} - \text{logsumexp}(a_n)}$$

$\text{sum}(r_{n1} \ldots r_{nK}) = 1$ ✓

$\log p(x_n, z_n = k)$

$a_{nk} \leftarrow \log \pi_k + \log N(x_n \mid \mu_k, \Sigma_k^2)$

$$r_{nk} = \frac{e^{a_{nk}}}{\sum e^{a_{nl}}}$$

$$r_{nk} = e^{[a_{nk} - \text{logsumexp}(a_n)]}$$

$$\text{logsumexp}(a_1 \dots a_K)$$

$$\log\left(\sum_k e^{a_k}\right) = \text{logsumexp}(a_1 \dots a_K)$$

$$M = \max_k a_k$$

$$\sum_k e^{-1000}$$

$K=3$

$a_1 = -10000$

$a_2 = -1000$

$a_3 = -1000$

$\text{logsumexp}(a_1 \dots a_K)$

$M = \max_k a_k$

$0 \cdot 0$

$$e^{M}\left(e^{a_1-M} + e^{a_2-M} + \dots + e^{a_K-M}\right)$$

$$\log(3)$$

$$-1000 + \log(3)$$

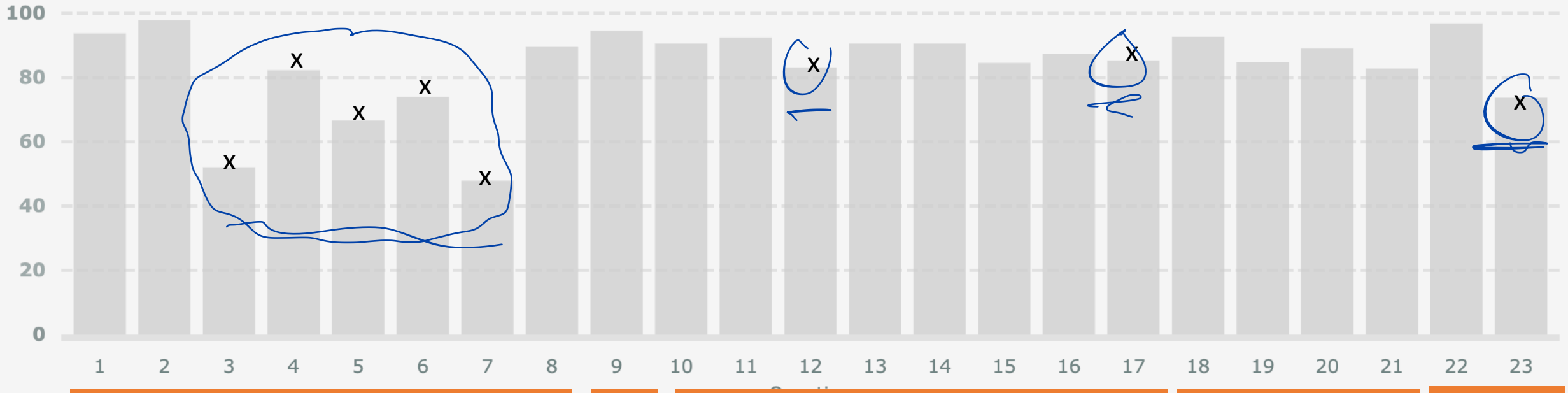
$$\rightarrow \log\left(e^{-1000} + e^{-1000} + e^{-1000}\right)$$

$$\rightarrow \log\left(e^{-1000}\left(e^{0} + e^{0} + e^{0}\right)\right) = \log e^{-1000} + \log\left(e^{0}+e^{0}+e^{0}\right)$$

Midterm

Percentiles:
- 10th : 0.77
- 50th : 0.87
- 90th : 0.93

Pretty good! ***Pretty good!***



Question 1. True/False	Q2	Question 3. Estimators with *k* and *r*	Q4	Q5

midtermsolutions.pdf

Piazza → Resources

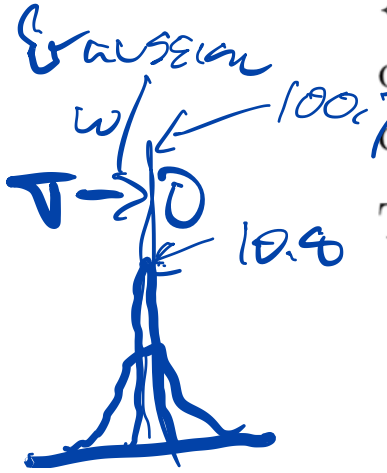
1c

$p(x)$ is a PDF for real x

PDF ≥ 0 (density)

(c) You compute the log marginal likelihood $\log p(\mathbf{x})$ of a dataset $\mathbf{x} = \{x_n\}_{n=1}^N$. Each observation is a real number: $x_n \in \mathbb{R}$. For a specific dataset, you compute $\log p(\mathbf{x}) = 2.345$. There must be a bug in your code; this value is impossible for a valid probabilistic model.

$0 \leq \text{Mass} \leq 1$



True False

PDFs of real-valued variables can be any value ≥ 0.

$e^{2.345}$ is a valid PDF value.

(d) Consider a random variable S with K possible values. $\{2, 4, 8, 16, \ldots 2^K\}$.

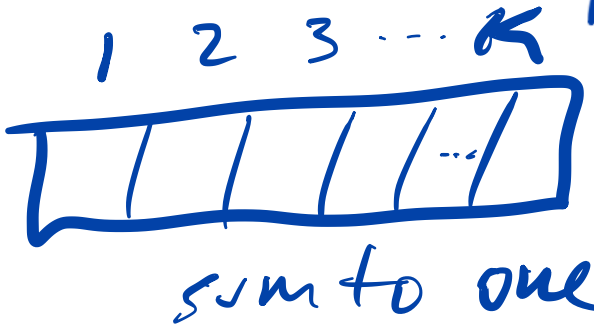
$p(x) = e^{2.345} > 0$

1d

(d) Consider a random variable S with K possible values, $\{2, 4, 8, 16, \ldots 2^K\}$. We wish to define the marginal distribution $p(S)$. We can define this distribution in terms of a parameter vector that requires exactly $K-1$ distinct scalar values ($K-1$ degrees of freedom).

True False

Yes, by the sum to one requirement, if we know K-1 values we can determine the proba. of the K^{th} outcome.



K = 2

0.6 | 0.4 = 1

1e

(e) When doing regression, if the inverse of $\Phi^T\Phi$ does not exist, this means there is no possible weight vector w that maximizes the likelihood.

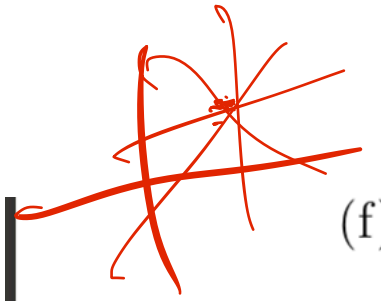
True False

There are infinitely many w that maximize lik.

how many lines go thru this point?

$y = mx + b$

1f



(f) When doing regression, if the number of examples N is much larger than the number of features M, there is always a unique weight vector w that maximizes the likelihood.

True False

(g) When doing regression, if we place a Normal prior on the precision β,

1g

w ~ Normal prior

Normal lik

$N(y|w^T x, b)$ → Normal post w/ $y, x \sim N$

(g) When doing regression, if we place a Normal prior on the precision β, the posterior on β is also Normal.

N times

True False

Normal prior on $\beta > 0$ doesn't make sense.

Not conjugate.

(h) We could create a generalized linear model for binary outcomes y_n

3c

$X = \{x_n\}_{n=1}^{N}$ $k(x_n|\theta)$ likelihood

$\theta \in \{\theta_l\}$ $r(\theta|\lambda)$ λ hyperparam

(c) Define the evidence, the marginal probability of the observed data (integrating away θ), in terms of the functions r and k.

log evidence

$$\log p(X|\lambda) = \int p(X,\theta)\,d\theta \quad \text{sum rule}$$

$$= \int \left[\prod_{n=1}^{N} k(x_n|\theta)\right] r(\theta|\lambda)\,d\theta$$

Common mistakes:

- log
-

3g: Note what **consistency** actually means

(g) List one advantage and one disadvantage of maximum likelihood estimation.

Consistent as $N \to \infty$

PRO:
- Possible to do for any probabilistic model
- Easier than alternatives (faster, simpler)

CON: Tend to overfit when data small

Consistency

If data drawn from $k(x_n|\theta)$

- then as $N \to \infty$

$\theta_{ML} \to \theta_{true}$

If model wrong

as $N \to \infty$

θ_{ML} will be the true par

3h

(h) Describe two reasonable ways to select hyperparameter λ from a grid of possible values $\lambda_1, \ldots \lambda_G$. For each one, specify the performance metric and how you would use the N training examples.

(1) Pick value with largest evidence

see expression from 3c

$$\max_{\lambda_g \in \{\lambda_1 \ldots \lambda_G\}} \log p(x \mid \lambda_g)$$

x: whole dataset

(2) Pick value w/ best heldout likelihood on validation set, using MAP estimate

Divide data into
80% train: x^{tr}
20% valid: x^{va}

$$\max_{\lambda_g \in \{\lambda_1 \ldots \lambda_G\}} \sum_{n=1}^{N} \log k(x_n^{va} \mid \theta_{MAP}(x^{tr}, \lambda_g))$$

Why can't use θ_{ML}?

N x_n
$k(\cdot)$
$r(\cdot)$

$\sum_n \log k(x_n \mid \theta)$

θ_{ML} does not depend on λ

private chat your answer

5b: Lagrange

$$\max_{w \in \mathbb{R}^D} \sum_{d=1}^{D} w_d c_d, \quad \text{subject to:} \quad \sum_d w_d^2 = 1$$

$w^T c$; w is unit vector

$\mathcal{L}(w,\lambda) = \sum_d w_d c_d + \lambda\left(1 - \sum_d w_d^2\right)$

first deriv = 0 does not alone make maximizer

sanity check!

$D = 1 \quad c = 1$

$w = \cancel{+2} \; +1 \cdot 1 = +1$

$-1 \cdot 1 = -1$

(b) Solve for w (a length-D vector) and $\lambda \neq 0$, via the system of $D+1$ equations: $\frac{d}{dw_1}\mathcal{L}(w,\lambda) = 0, \ldots \frac{d}{dw_D}\mathcal{L}(w,\lambda) = 0$, and $\nabla_\lambda \mathcal{L}(w,\lambda) = 0$

System of D+1 eqns

$\frac{\partial}{\partial w_1}\mathcal{L} = 0 \rightarrow \frac{\partial}{\partial w_1}\left[w^T c + \lambda\left(1 - \sum_d w_d^2\right)\right] = 0$

take deriv

$c_1 - \lambda 2 w_1 = 0$

add $\lambda 2 w_1$

$c_1 = \lambda 2 w_1 \; (1)$

$\frac{\partial}{\partial w_2}\mathcal{L} = 0 \rightarrow$ (as above) $c_2 = \lambda 2 w_2 \; (2)$

$\vdots$

$\frac{\partial}{\partial w_D}\mathcal{L} = 0 \rightarrow$ (as above) $c_D = \lambda 2 w_D \; (D)$

$\frac{\partial}{\partial \lambda}\mathcal{L} = 0 \rightarrow \frac{\partial}{\partial \lambda}\left[w^T c + \lambda\left(1 - \sum_d w_d^2\right)\right] = 0$

$1 = \sum_d w_d^2 \quad (D+1)$

Now take (1) − (D) & add squares of both sides

$c_1^2 = \lambda^2 4 w_1^2$

$\vdots$

$c_D^2 = \lambda^2 4 w_D^2$

$\sum_d c_d^2 = \lambda^2 4 \left(\sum_d w_d^2\right)$

Sub in $\sum_d w_d^2 = 1$ from (D+1) we get,

$\sum_d c_d^2 = 4\lambda^2 \quad \lambda \neq 0$

$\lambda = \frac{\pm\sqrt{c_1^2 + c_2^2 + \cdots + c_D^2}}{2}$

Now plug λ^* back into (1) … (D) to solve for $w_1^* \cdots w_D^*$

$w_1 = \frac{c_1}{2\lambda} = \frac{c_1}{\sqrt{c_1^2 + \cdots c_D^2}}$

$w_2 = \frac{c_2}{\sqrt{c_1^2 + \cdots c_D^2}}$

$w_D = \frac{c_D}{\sqrt{c_1^2 + \cdots c_D^2}}$

is w^* a unit vec?

That's it!

w should be the unit vector that points in same direction as c

$\lambda = \frac{\pm\sqrt{c_1^2 + \cdots c_D^2}}{2}$

$w_1 = \frac{\pm c_1}{\sqrt{c_1^2 + \cdots + c_D^2}}$, etc.

$w_D = \frac{c_D}{\sqrt{c_1^2 + \cdots + c_D^2}}$

Q&A Viterbi

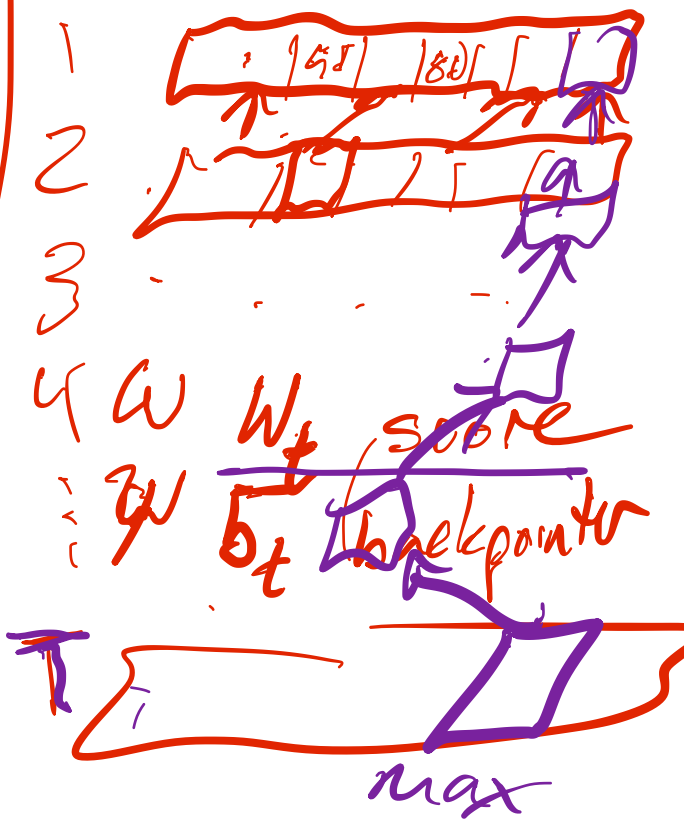
① Task: infer "best" state seq

$$z_1, z_2, \dots z_T = \underset{z_{1:T}}{\operatorname{argmax}}\ p(z_{1:T} \mid x_{1:T}, \theta)$$

note $[\hat{z}_1 \dots \hat{z}_T]$ with $x_{1:T}$

compared to $\hat{z}_1 \dots \hat{z}_T\ \hat{z}_{T+1}$ with $x_{1:T+1}$

② solve w/ dyn prog.



Big Picture

$x_{1:T}$

(1) Estimate HMM params $\theta = \{\pi, A, \mu, \Sigma\}$ max likelihood (E-M)

(2) given $x_{1:T}, \theta$ what is $\hat{z}_{1:T}$? (Viterbi)

(3) given $x_{1:T}, \theta$, what is $p(x_{1:T} | \theta)$

EM

alternate

$q(z) \leftarrow p(z_{1:T} | x_{1:T}, \theta)$

$\pi, A, \mu, \Sigma \leftarrow$ updated

optimizing

$\mathcal{L}(x, s, \pi, A, \mu, \Sigma)$

After E step

$q(z|s) = p(z_{1:T} | x_{1:T}, \theta)$

$= \mathbb{E}_q[\log p(x,z) - \log q(z)]$

$\alpha = \log p(x_{1:T}|\theta) = \mathbb{E}_q[\log p(x|\theta) + \log p(z|x,\theta) - \log p(z|x,\theta)]$

$= \log p(x|\theta)$

EM algorithm actually computes $\log p(x|\theta)$

correct EM algo

forward algo

for t
  s ∈ Estep
  θ ∈ Mstep α eval $\log p(x_{1:T}|\theta)$

---

† Forward algorithm

α_{tk} $P(z_t = k | x_{1:t}, \theta) = \frac{\text{joint } x_{1:t},\ z_t = k}{x_{1:t}}$