Hidden Markov Models

CISC 5800
Professor Daniel Leeds

Representing sequence data

- Spoken language
- DNA sequences
- Daily stock values

Example: spoken language

F?r plu? fi?e is nine

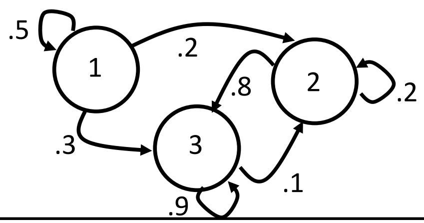
- Between F and r expect a vowel: “aw”, “ee”, “ah”; NOT “oh”, “uh”
- At end of “plu” expect consonant: “g”, “m”, “s”; NOT “d”, “p”

2

Markov Models

Start with:

- *n* states: $s_1, \ldots, s_n$
- Probability of initial start states: $\Pi_1, \ldots, \Pi_n$
- Probability of transition between states: $A_{i,j} = P(q_t{=}s_i | q_{t-1}{=}s_j)$

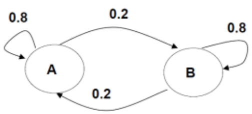


3

A dice-y example

$\Pi_A = 0.3, \Pi_B = 0.7$

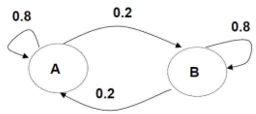
- Two colored die



- What is the probability we start at s_A? 0.3
- What is the probability we have the sequence of die choices: s_A, s_A? 0.3x0.8=0.24
- What is the probability we have the sequence of die choices: s_B, s_A, s_B, s_A? 0.7x0.2x0.2x0.2 = 0.0056

5

A dice-y example



- What is the probability we have the die choices s_B at time t=5

$$\Pi_A = 0.3, \Pi_B = 0.7$$

- Dynamic programming: find answer for q_t, then compute q_{t+1}

State\Time	t_1	t_2	t_3
s_A	0.3	0.38	0.428
s_B	0.7	0.62	0.572

$$p_t(i) = \sum_j p(q_t = s_i | q_{t-1} = s_j) p_{t-1}(j)$$

$p_t(i) = P(q_t = s_i)$ -- Probability state i at time t

7

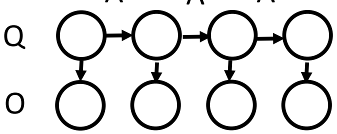
Hidden Markov Models

- Actual state q "hidden"
- State produces visible data o: $\phi_{i,j} = P(o_t = x_i | q_t = s_j)$

Probability observe value x_i when state is s_j

- Compute

$$P(\boldsymbol{O}, \boldsymbol{Q} | \boldsymbol{\theta}) = p(q_1|\pi) \prod_{t=2}^{T} p(q_t | q_{t-1}, \boldsymbol{A}) \prod_{t=1}^{T} p(o_t | q_t, \boldsymbol{\phi})$$



Probability of state sequence

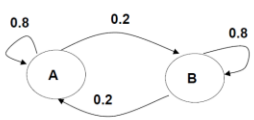
ϕ

Probability of observation sequence, given states

8

Deducing die based on observed "emissions"

Each color is biased



o	$P(o\|s_A)$	$P(o\|s_B)$
1	.3	.1
2	.2	.1
3	.2	.1
4	.1	.2
5	.1	.2
6	.1	.3

Intuition – balance transition and emission probabilities

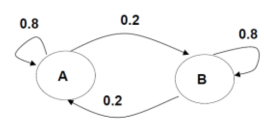
Observed numbers: 554565**2**54556 – the 2 is probably from s_B

Observed numbers: 554565**2**13321 – the 2 is probably from s_A

9

Deducing die based on observed "emissions"

Each color is biased



o	$P(o\|s_R)$	$P(o\|s_B)$
1	.3	.1
2	.2	.1
3	.2	.1
4	.1	.2
5	.1	.2
6	.1	.3

- We see: 5   What is probability of o=5 , q=B (blue)

$\Pi_B \phi_{5,B}$ = 0.7 x 0.2 = 0.14

- We see: 5, 3   What is probability of **o**=5,3 , **q**=B, B?

$\Pi_B \phi_{5,B} A_{B,B} \phi_{3,B}$ = 0.7 x 0.2 x 0.8 x 0.1 = 0.0112

11

Goal: calculate most likely states given observable data

$$\arg\max_Q P(Q \mid O) = \arg\max_Q \frac{P(O \mid Q)P(Q)}{P(O)}$$

$$= \arg\max_Q P(O \mid Q)P(Q)$$

Define and use $\delta_t(i)$

$$\delta_t(i) = \max_{q_1 \dots q_{t-1}} p(q_1 \dots q_{t-1} \wedge q_t = s_i \wedge O_1 \dots O_t)$$

$\delta_t(i)$: max possible value of $P(q_1,..,q_t,o_1,..,o_t)$ given we insist $q_t=s_i$

Find the most likely path from q_1 to q_t that

- $q_t=s_i$
- Outputs are $o_1, \dots, o_t$

Viterbi algorithm: $\delta_t(i)$

$\delta_1(i) = \Pi_i P(o_1 | q_1 = s_i) = \Pi_i \phi_{1,i}$

$\delta_t(i) = P(o_t | q_t = s_i) \max_j \delta_{t-1}(j) P(q_t = s_i | q_{t-1} = s_j) =$ $\boldsymbol{\phi_{o_t,i} \max_j \delta_{t-1}(j) A_{i,j}}$

$P(Q^*|O)=\text{argmax}_Q P(Q|O) = \text{argmax}_i \delta_t(i)$

Viterbi algorithm: bigger picture

Compute all $\delta_t(i)$'s

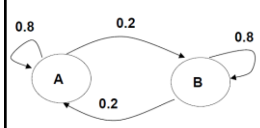
- At time t=1 compute $\delta_1(i)$ for every state i
- At time t=2 compute $\delta_2(i)$ for every state i (based on $\delta_1(i)$ values)
- …
- At time t=T compute $\delta_T(i)$ for every state i (based on $\delta_{T-1}(i)$ values)

Find states going from t=T back to t=1 to lead to max $\delta_T(i)$

- Now find state j that gives maximum value for $\delta_T(j)$
- Find state k at time T-1 used to maximize $\delta_T(j)$
- …
- Find state z at time 1 used to maximize $\delta_2(y)$

Viterbi in action: observe "5, 1"

$\Pi_A = 0.3, \Pi_B = 0.7$



o	$P(o\|s_A)$	$P(o\|s_B)$
1	.3	.1
2	.2	.1
3	.2	.1
4	.1	.2
5	.1	.2
6	.1	.3

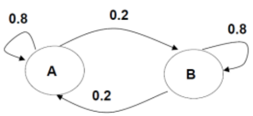
$\delta_2(A)$:
.3 x max(.8x.03 , **.2 x .14**)
= .3 x **.028** = .0084

$\delta_2(B)$:
.1 x max(.2x.03 , **.8 x .14**)
= .1 x **.112** = .0112

	t=1 (o_1=5)	t=2 (o_2=1)
$q_t=s_A$	.3x.1 = .03	.0084 (from B)
$q_t=s_B$	.7x.2 = .14	**.0112 (from B)**

Viterbi in action: observe "5, 1, 1, 1, 2"

$\Pi_A = 0.3, \Pi_B = 0.7$



o	$P(o\|s_A)$	$P(o\|s_B)$
1	.3	.1
2	.2	.1
3	.2	.1
4	.1	.2
5	.1	.2
6	.1	.3

$\delta_3(A)$:
.2 x max(**.8x.00054** , .2 x .00072)
= .2 x **.00043** = .000086

$\delta_3(B)$:
.1 x max(.2x.00054 , **.8 x .00072**)
= .1 x **.00058** = .000058

	t=1 (o_1=5)	t=2 (o_2=1)	t=3 (o_3=1)	t=4 (o_4=1)	t=5 (o_5=2)
$q_t=s_A$	.03	.0084 (<-B)	.00202 (<-A)	.00054 (<-B)	**.00009 (<-A)**
$q_t=s_B$	.14	.0112 (<-B)	**.00896 (<-B)**	**.00072 (<-B)**	.00006 (<-B)

State sequence: B, B, B, A, A

Parameters in HMM

Initial probabilities: π_i

Transition probabilities $A_{i,j}$

Emission probabilities $\phi_{i,j}$

How do we learn these values?

First, assume we know the states

Learning HMM parameters: π_i

x[1]: A,B,A,A,B
x[2]: B,B,B,A,A
x[3]: A,A,B,A,B
⋮

Compute MLE for each parameter

$$\pi^* = \underset{\pi}{\operatorname{argmax}} \prod_k \pi(q_1) \prod_{t=2}^{T} p(q_t|q_{t-1}) \prod_{t=1}^{T} p(o_t|q_t, \boldsymbol{\phi})$$

$$\pi_A = \frac{\#D(q_1 = s_A)}{\#D}$$

First, assume we know the states

Learning HMM parameters: $A_{i,j}$

x[1]: A,B,A,A,B
x[2]: B,B,B,A,A
x[3]: A,A,B,A,B
⋮

Compute MLE for each parameter

$$A^* = \underset{A}{\operatorname{argmax}} \prod_k \pi(q_1) \prod_{t=2}^{T} p(q_t|q_{t-1}) \prod_{t=1}^{T} p(o_t|q_t, \boldsymbol{\phi})$$

$$A_{i,j} = \frac{\#D(q_t=s_i, q_{t-1}=s_j)}{\#D(q_{t-1}=s_j)}$$

First, assume we know the states

Learning HMM parameters: $\phi_{i,j}$

x[1]: A,B,A,A,B
o[1]: 2,5,3,3,6

x[2]: B,B,B,A,A
o[2]: 4,5,1,3,2

x[3]: A,A,B,A,B
o[3]: 1,4,5,2,6

⋮

Compute MLE for each parameter

$$\phi^* = \underset{\phi}{\operatorname{argmax}} \prod_k \pi(q_1) \prod_{t=2}^{T} p(q_t|q_{t-1}) \prod_{t=1}^{T} p(o_t|q_t, \boldsymbol{\phi})$$

$$\phi_{i,j} = \frac{\#D(o_t = i, q_t = s_j)}{\#D(q_t = s_j)}$$

Challenges in HMM learning

Learning parameters (π, A, ϕ) with known states is not too hard

BUT usually states are unknown

If we had the parameters and the observations, we could figure out the states: Viterbi $P(Q^*|O)=\operatorname{argmax}_Q P(Q|O)$

Expectation-Maximization, or “EM”

Problem: Uncertain of y^i (class), uncertain of $\boldsymbol{\theta}^i$ (parameters)

Solution: Guess y^i, deduce $\boldsymbol{\theta}^i$, re-compute y^i, re-compute $\boldsymbol{\theta}^i$ … etc.
OR: Guess $\boldsymbol{\theta}^i$, deduce y^i, re-compute $\boldsymbol{\theta}^i$, re-compute y^i

Will converge to a solution

E step: Fill in expected values for missing labels y

M step: Regular MLE for $\boldsymbol{\theta}$ given known and filled-in variables

Also useful when there are holes in your data

Computing states q_t

Instead of picking one state: $q_t=s_i$, find $P(q_t=s_i|\mathbf{o})$

$$P(q_t = s_i|o_1, \cdots, o_T) = \frac{\alpha_t(i)\beta_t(i)}{\sum_j \alpha_t(j)\beta_t(j)}$$

Forward probability: $\boldsymbol{\alpha_t(i) = P(o_1 \ldots o_t \wedge q_t = s_i)}$

Backward probability: $\boldsymbol{\beta_t(i) = P(o_{t+1} \ldots o_T | q_t = s_i)}$

Details of forward probability

Forward probability: $\boldsymbol{\alpha_t(i) = P(o_1 \ldots o_t \wedge q_t = s_i)}$

$$\alpha_1(i) = \phi_{o_1,i}\pi_i = P(o_1|q_1 = s_i)P(q_1 = s_i)$$

$$\alpha_t(i) = \phi_{o_t,i} \sum_j A_{i,j}\alpha_{t-1}(j)$$

$$\alpha_t(i) = P(o_t|q_t = s_i) \sum_j P(q_t = s_i|q_{t-1} = s_j)\alpha_{t-1}(j)$$

27

Details of backward probability

Backward probability: $\boldsymbol{\beta_t(i) = P(o_{t+1} \ldots o_T|q_t = s_i)}$

$$\beta_t(i) = \sum_j A_{j,i}\phi_{o_{t+1},j}\beta_{t+1}(j)$$

$$\beta_t(i) = \sum_j P(q_{t+1} = s_j|q_t = s_i)P(o_{t+1}|q_{t+1} = s_j)\beta_{t+1}(j)$$

Final $\boldsymbol{\beta}$: $\boldsymbol{\beta_{T-1}(i)}$

$$\beta_{T-1}(i) = \sum_j A_{j,i}\phi_{o_{T-1},j}$$

$$= P(q_T = s_j|q_{T-1} = s_i)P(o_T|q_T = s_j)$$

28

E-step: State probabilities

One state:

$$P(q_t = s_i|o_1, \cdots, o_T) = \frac{\alpha_t(i)\beta_t(i)}{\sum_j \alpha_t(j)\beta_t(j)} = S_t(i)$$

Two states in a row:

$$P(q_t = s_j, q_{t+1} = s_i|o_1, \cdots, o_T) = \frac{\alpha_t(j)A_{i,j}\phi_{o_{t+1},i}\beta_{t+1}(i)}{\sum_i \sum_j \alpha_t(j)A_{i,j}\phi_{o_{t+1},i}\beta_{t+1}(i)}$$

$$= S_t(i,j)$$

29

Recall: when states known

$$\pi_A = \frac{\#D(q_1 = s_A)}{\#D}$$

$$A_{i,j} = \frac{\#D(q_t = s_i, q_{t-1} = s_j)}{\#D(q_{t-1} = s_j)}$$

$$\phi_{i,j} = \frac{\#D(o_t = i)}{\#D(q_t = s_j)}$$

30

M-step

$A_{i,j} = \frac{\sum_t S_t(i,j)}{\sum_t S_t(i)}$

$\phi_{obs,i} = \frac{\sum_{t|\boldsymbol{o_t=obs}} S_t(i)}{\sum_t S_t(\boldsymbol{i})}$

$\pi_i = S_1(i)$

Known states:

- $\pi_A = \frac{\#D(q_1=s_A)}{\#D}$
- $A_{i,j} = \frac{\#D(q_t=s_i, q_{t-1}=s_j)}{\#D(q_{t-1}=s_j)}$
- $\phi_{i,j} = \frac{\#D(o_t=i)}{\#D(q_t=s_j)}$

31

Review of HMMs in action

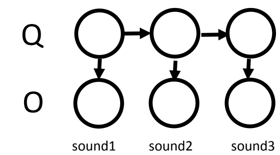
For classification, find highest probability class given features

Features for one sound:

- [q_1, o_1, q_2, o_2, …, q_T, o_T]

Conclude word:

Generates states:



33