Review of last lecture

① FNNs and its applications

- Sentiment classification
- Neural language model

② Training neural nets

- Computation graph
- Backward differentiation

③ Sequence labeling : NER & POS Tagging

④ • Markov chain

1. Set of states : $Q = \{q_1, q_2, \cdots, q_n\}$
2. Transition probability matrix A
   a_{ij} : prob. of moving state $i \to$ state j.
3. $\pi = \pi_1 \pi_2 \cdots \pi_n$ initial probability distribution

⑤ • Hidden Markov chain.

- observed events / hidden events

1. Set of states : $Q = \{q_1, q_2, \cdots, q_N\}$ $q_1 q_2 \cdots q_N = q_{1:N}$
2. Transition prob. mat. A where a_{ij} : prob of moving state $i \to$ state j.
3. $\pi = \pi_1 \pi_2 \cdots \pi_n$: initial prob. dist.
4. Observation likelihoods / Emission prob. mat. $B = b_i(O_t)$,
   where $b_i(O_t)$ is the prob. of an observation O_t being generated from a state q_i.
5. sequence of observations $O = O_1 O_2 \cdots O_T = O_{1:T}$

Two assumptions.

1. Markov assumption : $P(q_i \mid q_{1:i-1}) = P(q_i \mid q_{i-1})$
2. Output Independence :
   $P(O_i \mid q_{1:T}, O_{1:T}) = P(O_i \mid q_i)$

Three problems of HMM

1. likelihood calculation.
2. decoding
3. learning

Hidden Markov Model

5 components:

1. $Q = q_1 q_2 \cdots q_N = q_{1:N}$
2. $A = a_{11} \cdots a_{ij} \cdots a_{NN}$
3. $O = o_1 o_2 \cdots o_T = o_{1:T}$
4. $B = b_i(o_t)$
5. $\pi = \pi_1, \pi_2, \cdots, \pi_N = \pi_{1:N}$

Two assumptions:

① Markov assumption:

$$p(q_i | q_{1:i-1}) = p(q_i | q_{i-1})$$

② Output Independence:

$$p(o_i | q_{1:T}, o_{1:T}) = p(o_i | q_i).$$

☆ Likelihood calculation

$Q = \{q_1 = \text{cold}, q_2 = \text{hot}\}.$

$O = o_1$

$$p(o_1) = p(o_1, q_1) + p(o_1, q_2)$$

$$= \underbrace{p(o_1|q_1) \cdot p(q_1|\text{start})}_{\alpha_1(q_1)} + \underbrace{p(o_1|q_2) \cdot p(q_2|\text{start})}_{\alpha_1(q_2)}$$

$\alpha_1(q_1)$: observed o_1 and finally ending at q_1

$\alpha_1(q_2)$: observed o_1 and finall ending at q_2.

$$p(o_1 o_2) = \underbrace{p(o_{1:2}, q_1 q_1) + p(o_{1:2}, q_2 q_1)}_{\alpha_2(q_1)} + \underbrace{p(o_{1:2}, q_1 q_2) + p(o_{1:2}, q_2 q_2)}_{\alpha_2(q_2)}$$

$\alpha_2(q_1)$: observed $o_{1:2}$ and finally ending at q_1

$\alpha_2(q_2)$: observed $o_{1:2}$ and finally ending at q_2.

$$\alpha_2(q_1) = \underline{p(o_1|q_1)} \cdot \underline{p(o_2|q_1)} \cdot \underline{p(q_1|\text{start})} \cdot p(q_1|q_1) + \underline{p(o_1|q_2)} \cdot \underline{p(o_2|q_1)} \cdot \underline{p(q_2|\text{start})} \cdot p(q_2|q_1)$$

$$= [\alpha_1(q_1) \cdot p(q_1|q_1) + \alpha_1(q_2) \cdot p(q_2|q_1)] \cdot p(o_2|q_1)$$

$$\alpha_2(q_2) = [\alpha_1(q_1) \cdot p(q_2|q_1) + \alpha_1(q_2) \cdot p(q_2|q_2)] \cdot p(o_2|q_2).$$

$$= \underline{p(o_1|q_1)} \cdot p(o_2|q_2) \cdot \underline{p(q_1|\text{start})} \cdot p(q_2|q_1) + \underline{p(o_1|q_2)} \cdot p(o_2|q_2) \cdot p(q_2|\text{start}) \cdot p(q_2|q_2)$$

$$p(O=o_{1:T}) = \sum_{Q} p(O,Q) = \sum_{Q} p(O|Q)\cdot p(Q)$$

$$= \sum_{Q}\left[\prod_{i=1}^{T} p(o_i|q_i)\cdot \prod_{i=1}^{T} p(q_i|q_{i-1})\right].$$

q_0: initial state.

Let $\lambda = (A, B)$

$$\alpha_t(j) = p(o_{1:t}, q_t = j|\lambda)$$

$q_t = j$: t-th state in the sequence of states is state j.

$$\alpha_t(q_t) = p(o_{1:t}, q_t = j|\lambda) = \sum_{q_{t-1}\in Q} p(o_{1:t}, q_{t-1}, q_t = j|\lambda).$$

Since q_{t-1} is always finite and in Q.

Rewrite $o_{1:t} = o_{1:t-1}\,o_t$ and rewrite:

$$p(o_{1:t}, q_{t-1}, q_t) = p(o_{1:t-1}, o_t, q_{t-1}, q_t = j)$$

$$= p(o_t|o_{1:t-1}, q_{t-1}, q_t)\cdot p(o_{1:t-1}, q_{t-1}, q_t)$$

By the output independence assumption: o_t only depends on q_t.

So, $\alpha_t(q_t) = \sum_{q_{t-1}\in Q} p(o_t|q_t)\cdot p(o_{1:t-1}, q_{t-1}, q_t)$

Note: $p(o_{1:t-1}, q_{t-1}, q_t) = p(q_t|o_{1:t-1}, q_{t-1})\cdot p(o_{1:t-1}, q_{t-1})$

$$= p(q_t|q_{t-1})\cdot p(o_{1:t-1}, q_{t-1})$$

We reach:

$$\alpha_t(q_t) = \sum_{q_{t-1}\in Q} p(o_t|q_t)\cdot p(q_t|q_{t-1})\cdot \underbrace{p(o_{1:t-1}, q_{t-1})}_{\alpha_{t-1}(q_{t-1})}$$

Note that $b_j(o_t) = p(o_t|q_t = j)$, $p(q_t|q_{t-1}) = a_{ij}$. Finally Let $q_{t-1} = i$, $q_t = j$, then

$$\alpha_t(q_t) = \sum_{i\in Q} b_j(o_t)\cdot a_{ij}\cdot \alpha_{t-1}(i) = b_j(o_t)\cdot \sum_{i\in Q} \alpha_{t-1}(i)\cdot a_{ij}.$$

Remark:

①. Initial: $\alpha_1(j) = \pi_j \cdot b_j(o_1) \quad 1 \le j \le N.$

Recursion: $\alpha_t(j) = b_j(o_t) \sum_{i=1}^{N} \alpha_{t-1}(i) \cdot a_{ij} \quad 1 \le j \le N,\ 1 < t \le T.$

Termination: $p(O \mid \lambda = (A, B)) = \sum_{i=1}^{N} \alpha_T(i).$

②. Time complexity: $O(N^2 \cdot T).$

☆ Decoding

Given an input HMM, $\lambda = (A, B)$ and $O = o_{1:T}$,

find the most probable sequence of states $Q = q_{1:T}$.

Naive: $\max_{q_{1:T}} p(q_{1:T}, o_{1:T} \mid \lambda).$

Time complexity: $O(N^T).$

Subproblem: we seek to find a solution of a sub problem, define

$v_t(j)$: the prob. that the HMM model is in state j after seeing

the first t observations and passing through the most probable state

sequence $q_1 q_2 \cdots q_{t-1}$, given the mode $\lambda = (A, B)$. That is

$$v_t(j) = \max_{q_{1:t-1}} p(q_{1:t-1}, o_{1:t}, q_t = j \mid \lambda).$$

Try optimal sub structure:

$$v_t(j) = \max_{q_{t-1}} \left\{ \max_{q_{1:t-2}} p(q_{1:t-1}, o_{1:t}, q_t = j \mid \lambda) \right\}. \quad \text{why?}$$

Note that

$$p(\underline{q_{1:t-1}}, o_{1:t}, q_t = j) = p(\underline{q_{1:t-2}, q_{t-1}}, o_{1:t}, q_t = j).$$

Since we want to have $p(o_t | \cdots)$,

$$p(q_{1:t-2}, q_{t-1}, o_{1:t}, q_t = j)$$

$$= \underbrace{p(o_t | q_{1:t-2}, o_{1:t-1}, q_{t-1}, q_t = j)}_{\text{Output Independence}} \cdot p(q_{1:t-2}, o_{1:t-1}, q_{t-1}, q_t = j)$$

$$= p(o_t | q_t = j) \cdot \underbrace{p(q_{1:t-2}, o_{1:t-1}, q_{t-1}, q_t = j)}_{(*)}$$

We want to see $p(q_t = j | \cdots)$

$$(*) = \underbrace{p(q_t = j | q_{1:t-2}, o_{1:t-1}, q_{t-1})}_{\text{Markov Assum.}} \cdot p(q_{1:t-2}, o_{1:t-1}, q_{t-1})$$

$$= p(q_t = j | q_{t-1}) \cdot p(q_{1:t-2}, o_{1:t-1}, q_{t-1})$$

Back to our subproblem:

$$v_t(j) = \max_{q_{t-1} \in Q} \left\{ \max_{q_{1:t-2}} \underbrace{p(o_t | q_t = j) \cdot p(q_t = j | q_{t-1})}_{\text{constant w.r.t } q_{1:t-2}} \cdot p(q_{1:t-2}, o_{1:t-1}, q_{t-1}) \right\}.$$

$$= p(o_t | q_t = j) \cdot \max_{q_{t-1} \in Q} \left\{ p(q_t = j | q_{t-1}) \cdot \underbrace{\max_{q_{1:t-2}} \left(p(q_{1:t-2}, o_{1:t-1}, q_{t-1}) \right)}_{v_{t-1}(q_{t-1})} \right\}$$

$$= p(o_t | q_t = j) \cdot \max_{q_{t-1} \in Q} \left\{ p(q_t = j | q_{t-1}) \cdot v_{t-1}(q_{t-1}) \right\}.$$

As $b_j(o_t) = p(o_t | q_t = j)$, $q_{t-1} = i$: defined, then

$$v_t(j) = b_j(o_t) \cdot \max_{i \in [N]} a_{ij} \cdot v_{t-1}(i).$$

Remark: $v_1(j) = \pi_j b_j(o_1) \quad 1 \le j \le N.$

$bt_1(j) = 0 \quad 1 \le j \le N$

$bt_t(j) = \arg\max_{i \in [N]} v_{t-1}(i) \cdot a_{ij} b_j(o_t).$

Time complexity: $O(N^2 \cdot T)$. Space: $O(N \cdot T)$.

※ Training

Given $Q = o_{1:T}$, V, we want to know $\lambda = (A, B)$.

$[A]_{ij} = a_{ij}$. $[B]_{jk} = b_j(o_k)$.

For a_{ij}, we define the estimate $\hat{a}_{ij}$, from MLE, we know

$$\hat{a}_{ij} = \frac{E_T[\#\text{ transitions } i \to j]}{E_T[\#\text{ transitions from } i]} \qquad T: \text{tokens}$$

Define $\xi_t(i,j)$: prob. of being in state i at time t and state j at time $t+1$.

$$\xi_t(i,j) = p(q_t = i, q_{t+1} = j \mid O, \lambda)$$

We know $p(q_t, q_{t+1} \mid O) = \dfrac{p(q_t, q_{t+1}, O)}{p(O)}$

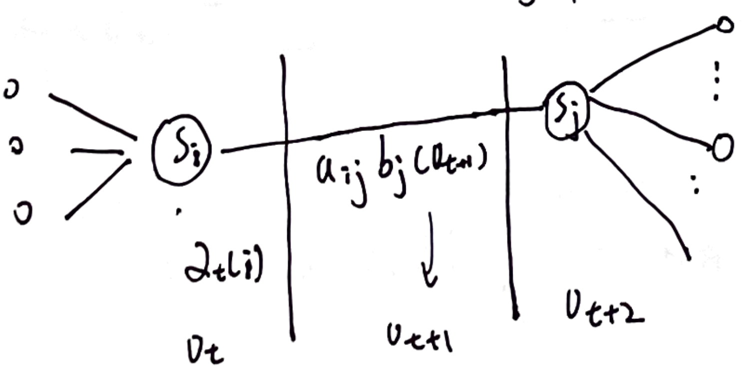
$p(O \mid \lambda)$: forward algo.

$p(q_t, q_{t+1}, O \mid \lambda)$ is difficulty. But, define backward prob.

$$\beta_t(i) = p(o_{t+1}, o_{t+2}, \dots o_T \mid q_t = i, \lambda)$$

prob. seeing $o_{t+1:T}$ given we are at state i at time t.

Verify by yourself: $\beta_t(i) = \sum_{j=1}^{N} a_{ij}\, b_j(o_{t+1}) \cdot \beta_{t+1}(j)$.



One of terms in $\beta_t(j)$ matches so:

$$p(q_t, q_{t+1}, O \mid \lambda) = \alpha_t(i) \cdot a_{ij}\, b_j(o_{t+1}) \cdot \beta_{t+1}(j). \quad (\text{why?})$$

$$p(O \mid \lambda) = \sum_{j=1}^{N} \alpha_t(j) \cdot \beta_t(j). \quad (\text{why?})$$

$$\xi_t(i,j) = \frac{\alpha_t(i)\, a_{ij}\, b_j(O_{t+1}) \cdot \beta_{t+1}(j)}{\sum_{j=1}^{N} \alpha_t(j)\, \beta_t(j)}$$, then

$$\hat{a}_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \sum_{k=1}^{N} \xi_t(i,k)}.$$

How to estimate $b_j(O_t)$. define $\hat{b}_j(v_k)$

$$\hat{b}_j(v_k) = \frac{E_T[\# \text{ in } j \text{ and observing } v_k)}{E_T[\# \text{ in state } j]}.$$

So, we need

$$\gamma_t(j) = p(q_t = j \mid O, \lambda) = \frac{p(q_t = j, O \mid \lambda)}{p(O \mid \lambda)}$$, where

$$p(q_t = j, O \mid \lambda) = p(O_{t+1:T} \mid q_t, O_{1:t}) \cdot p(q_t, O_{1:t})$$
$$= \alpha_t(j) \cdot \beta_t(j).$$

$$p(O \mid \lambda) = p(O_{1:t}, O_{t+1:T}) = \sum_{j \in Q} p(q_t = j, O_{1:t}, O_{t+1:T} \mid \lambda)$$
$$= \sum_{j=1}^{N} \alpha_t(j) \cdot \beta_t(j).$$

$$\hat{b}_j(v_k) = \frac{\sum_{t=1, O_t = v_k}^{T} \gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(j)}.$$

How to estimate $\hat{a}$, $\hat{b}$, restimate many times!

EM algo. in slides.