Natural Language Processing
CSE 325/425

Sihong Xie

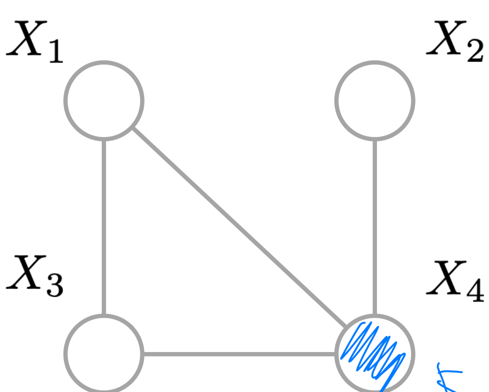
Lecture 11:

- Conditional random field
- Neural network revisit (forward propagation)

Graphical models

Conditional independence in a graphical model:

- $A \perp\!\!\!\perp B | C$ if any path from A to B have to pass some variables in C.



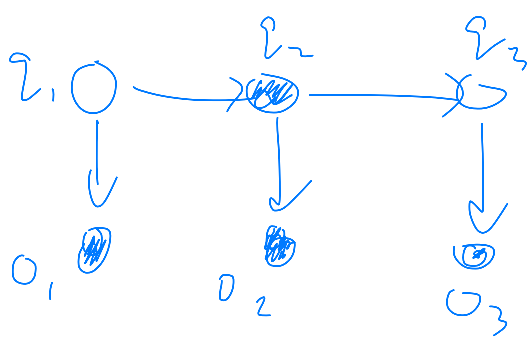
$X_1 \perp\!\!\!\perp X_2 \mid X_4$

$X_1 \not\perp\!\!\!\perp X_4 \mid X_3$

observed

"blocks" all paths from X_1 to X_2

HMM:



$\left(\sum_{q_1} Pr(q_1|q_2)\right)$

$\sum_{q_1}\sum_{q_3} Pr(q_1, q_3 | q_2) = \times$

$q_1 \perp\!\!\!\perp q_3 \mid q_2 \Rightarrow Pr(q_1, q_3 | q_2)$ $\left(\sum_{q_3} Pr(q_3|q_2)\right)$

"forward / backward" $= Pr(q_1|q_2) \times Pr(q_3|q_2)$

Cliques

Example of cliques:

① clique = $\{X_1, X_3, X_4\}$

(maximal clique)

② clique = $\{X_2\ X_4\}$

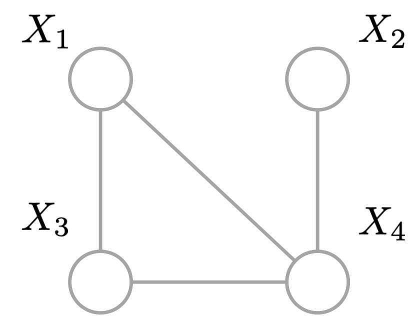
③ clique = $\{X_1\ X_3\}$ ← not a maximal clique

$X_1 \perp\!\!\!\perp X_2 | X_4$

Cliques and maximal cliques

- A clique ***c*** is a set of nodes that are fully connected
  - any two nodes in the clique are connected by an edge.
  - two nodes not in a clique **can** become conditional independent.

$$\Pr(X_i, X_j | X_{\backslash i,j}) = \Pr(X_i | X_{\backslash i,j}) \Pr(X_j | X_{\backslash i,j})$$



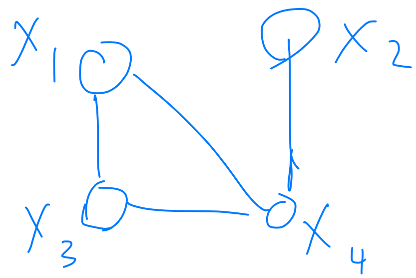
⇓ when can do "polynomial time" marginalization, such as, $Pr(X_1)$

- A maximal clique ***c*** is a clique that adding any new node will make their union not a clique.
  - the collection of maximal cliques on a graphical model encode all conditional independence properties.

$$\sum_{X_2X_3X_4} Pr(X_1X_3)Pr(X_1X_4)P(X_3X_4)Pr(X_2X_4)$$

$$= \sum_{X_2X_3X_4} Pr(X_1, X_2, X_3, X_4)$$

$$= \sum_{X_2X_3X_4} Pr(X_1X_3X_4) \times Pr(X_2X_4)$$



$c = 1 : \{X_1, X_3\ X_4\}$

$c = 2 : \{X_2, X_4\}$

$\mathcal{C} = \{1, 2\}$

Factorization

$X_c = \{X_i : X_i \in c\}$

Factorization using cliques.

$$\Pr(X_1, \ldots, X_n) = \frac{1}{Z} \prod_{c \in \mathcal{C}} \psi_c(\underline{X_c})$$

$\psi_1(X_1, X_3 X_4) : (x_1, x_3 x_4) \to (0, \infty)$

$\psi_2(X_2 X_4) : (x_2 x_4) \to (0, \infty)$

where

$\mathcal{C}$ = set of all cliques

- ***c*** is a clique.
- $\psi_c(X_c) \overset{>}{\cancel{\geq}} 0$ is a potential function of the variable in the clique ***c***.
  - this is not a joint distribution of the variables X_c
- The normalization factor is defined as

"partition function"

$$Z = \sum_{x_1, \ldots, x_n} \prod_{c \in \mathcal{C}} \psi_c(X_c)$$

$O(2^4)$

$X_1, X_2, X_3, \& X_4 \in \{0, 1\}$

$Z = \psi_1(X_1 = 0, X_3 = 0, X_4 = 0) \times \psi_2(x_2 = 0, x_4 = 0)$

$+ \quad + \psi_1(X_1 = 1, X_3 = 0, X_4 = 0) \times \psi_2(X_2 = 0, X_4 = 0)$

$\psi_1(X_1 = X_3 = X_4 = 1)$

$\psi(X_2 = X_4 = 1)$ . . . .

Discriminative (MEMM or CRF) : $P(Q \mid O, \theta)$

Generative (HMM) : $P(Q, O \mid \lambda)$

Conditional random fields

Conditional random fields:

- A "random field" refer to "a set of dependent random variables".
  - a specific type of "graphical models".
- "Conditional" means "conditioning on observed data" observed words from a sentence.
  - making CRF a discriminative model (vs. generative models such as HMM).
- Use the maximum entropy principle – a generalization of logistic regression.
  - each factor is in the form
    - ***c*** is a maximum clique
  - joint conditional distribution

$$\psi_c(X_c; O) = \exp\left\{\sum_{i=1}^{d} \theta_i f_i(X_c; O)\right\}$$

CRF parameter $\vec{\theta} = \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_d \end{bmatrix} \in \mathbb{R}^d$

$$\Pr(\underbrace{\cancel{X_1}, \ldots, \cancel{X_n}}_{q_1 \cdots q_T}; O) = \frac{1}{Z(O)} \boxed{\prod_{c \in \mathcal{C}}} \exp\left\{\sum_{i=1}^{d} \theta_i f_i(\underbrace{\cancel{X_c}}_{Q_c}; O)\right\}$$

d

$q_t \in \{1, \cdots, N\}$

$$Z(O) = \sum_{q_1 \cdots q_T} \prod_{c \in \mathcal{C}} \exp\left\{\sum_{i=1}^{d} \theta_i f_i(Q_c; O)\right\}$$

(q_t, q_{t-1})

Summation takes $O(N^T)$

O1 O2

q1 → q2

I went to church yesterday

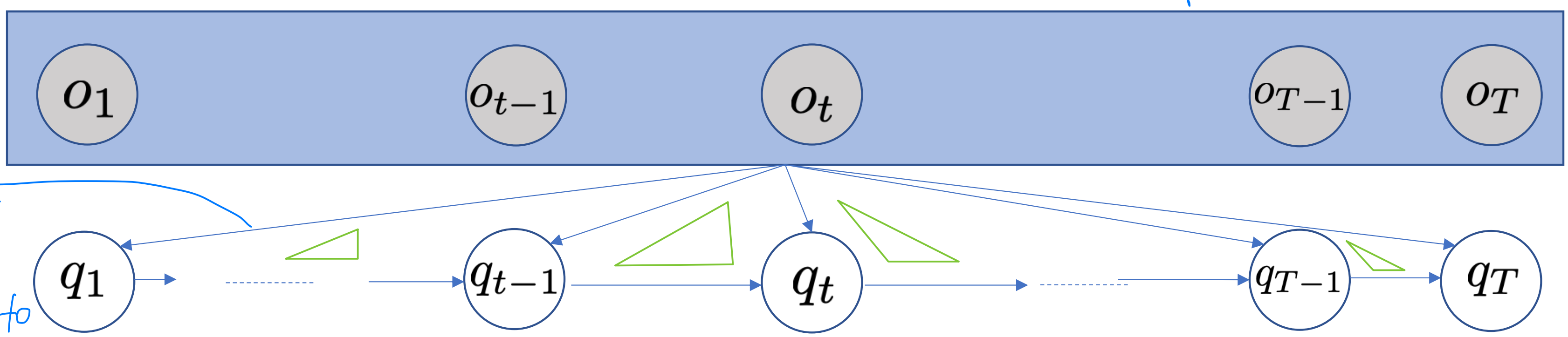
VBD

q2

CRF for POS tagging

Ot T

The graphical model is a linear chain



want to predict each tag using info of the entire sentence

All factors (or equivalently, maximum cliques) are pairwise

$$\psi_{t-1,t}(\underbrace{q_t, q_{t-1}}_{Q_t}; O) = \exp\left\{\sum_{i=1}^{d} \theta_i \boxed{f_i(q_t, q_{t-1}; O)}\right\}$$

$: (q_t, q_{t-1}) \to (0, \infty)$

$\mathcal{C} = \{\{q_1, q_2, O\}, \{q_2, q_3, O\}, \ldots, \{q_{T-1}, q_T, O\}\}$

The inner product = the compatibility score of the two tags:

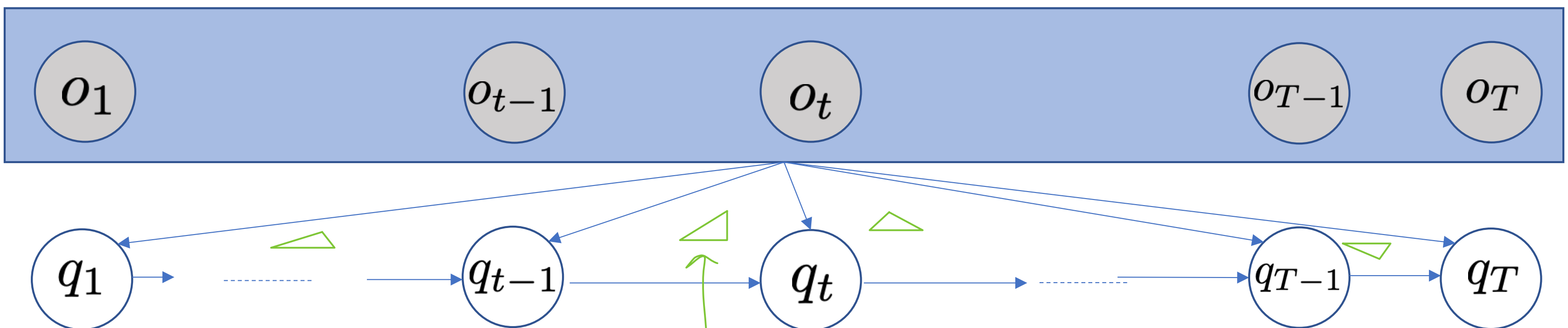
$$s_t(q_{t-1}, q_t; \boldsymbol{\theta}, O) = \langle \boldsymbol{\theta}, \mathbf{f}(q_{t-1}, q_t) \rangle \in \mathbb{R}$$

(replacing $a_{q_{t-1}q_t} b_{q_t}(o_t)$ in HMM, or $\Pr(q_t \mid q_{t-1}, O)$ in MEMM)

$\vec{\theta} = \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_d \end{bmatrix} \in \mathbb{R}^d$, $\quad \vec{f}(q_t, q_{t-1}, O) = \begin{bmatrix} f_1(q_t, q_{t-1}, O) \\ \vdots \\ f_d(q_t, q_{t-1}, O) \end{bmatrix}$

CRF for POS tagging

The graphical model is a linear chain



- The joint distribution of a tag sequence, conditioned on a word sequence is

$$\Pr(q_1, \ldots, q_T; O) = \frac{1}{Z(O)} \prod_{t=1}^{T} \boxed{\exp\left\{\sum_{i=1}^{d} \theta_i f_i(q_t, q_{t-1}; O)\right\}}$$

$$= \frac{1}{Z(O)} \exp\left\{\sum_{t=1}^{T} \boxed{\sum_{i=1}^{d} \theta_i f_i(q_t, q_{t-1}; O)}\right\} = \frac{1}{Z(O)} \exp\left\{\sum_{t=1}^{T} s_t(q_{t-1}, q_t; \boldsymbol{\theta}, O)\right\}$$

decompose over steps of the sequence:
=> conditional independence
=> polynomial inference alg.

$\langle \vec{\theta}, \vec{f}(q_t, q_{t-1}; O) \rangle$

Predicting sequence using CRF

Input:

- an input sentence $O = [o_1, \ldots, o_T], o_t \in V$
- and a trained CRF model $\boldsymbol{\theta}$

Output:

- optimal POS tag sequence $Q^* = \arg\max_Q \Pr(Q|O; \boldsymbol{\theta})$ $= \arg\max_Q \frac{1}{Z(O)} \exp\{\sum_{t=1}^{T} s_t(q_{t-1}, q_t; \theta, O)\}$

$$= \arg\max_Q \sum_{t=1}^{T} s_t(q_{t-1}, q_t; \boldsymbol{\theta}, O)$$

Adapt the Viterbi algorithm for HMM to CRF prediction:

- change the scores in HMM $s_t(q_{t-1} = i, q_t = j; \lambda, O) = a_{i,j} b_j(o_t)$

  to the scores defined for CRF.

? MEMM what you're replacing

$$\log Z(O) = \log \sum_{Q} \exp\left\{ \sum_{t=1}^{T} \langle \vec{\theta}, \vec{f}(q_t, q_{t-1}, \theta, O) \rangle \right\}$$

$$\frac{\partial}{\partial \theta} \log Z(O) = \frac{1}{Z(O)} \frac{\partial}{\partial \theta} Z(O) = \frac{1}{Z(O)} \sum_{Q} \exp\left\{ \sum_{t=1}^{T} \langle \vec{\theta}, \vec{f} \rangle \right\} \sum_{t=1}^{T} \frac{\partial}{\partial \theta} \langle \vec{\theta}, \vec{f}(q_t, q_{t-1}, \theta, O) \rangle$$

Learning a CRF model

$$= \sum_{Q} \Pr(Q \mid \theta, O) \sum_{t=1}^{T} \vec{f}(q_t, q_{t-1}, \theta, O)$$

$$= \sum_{t=1}^{T} \mathbb{E}_{Q \sim \Pr(Q \mid \theta, O)} \vec{f}(q_t, q_{t-1}, \theta, O)$$

Not much more difficult than training a logistic regression model!

Expectation of feature vector

$$\mathbb{E}_{\vec{x} \sim \Pr(\vec{x})} \vec{f}(\vec{x}) \triangleq \sum_{\vec{x}} \Pr(\vec{x}) f(\vec{x})$$

- Input: m POS-tagged sentences.
- MLE: $$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{m} \log \Pr(Q^{(i)} | O^{(i)}, \boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{m} \ell(Q^{(i)} | O^{(i)}, \boldsymbol{\theta})$$
- There is no closed form solution for the parameter, and gradient descent is needed.

Logistic regression "Glove"

$$\frac{\partial}{\partial \boldsymbol{\theta}} \ell(Q | O, \boldsymbol{\theta}) = \frac{\partial}{\partial \theta} \log \Pr(Q \mid O; \theta) = \frac{\partial}{\partial \theta} \log \frac{1}{Z(O)} \exp\left\{ \sum_{t=1}^{T} \langle \vec{\theta}, \vec{f}_t \rangle \right\}$$

$$= \frac{\partial}{\partial \theta} \left[\sum_{t=1}^{T} \langle \vec{\theta}, \vec{f}_t \rangle - \log Z(O) \right]$$

$$= \sum_{t=1}^{T} \vec{f}_t - \sum_{t=1}^{T} \mathbb{E}_{Q \sim \Pr(Q \mid O, \theta)} \vec{f}(q_t, q_{t-1}, \theta, O)$$

decompose over steps of the sequence:
=> conditional independence
=> polynomial inference alg.

- Recall the gradient for multi-class logistic regression …

matrix calculus:

$$\frac{\partial}{\partial \theta} \langle \vec{\theta}, \vec{f} \rangle = \vec{f}$$

learning rate 0.1

$$\vec{\theta} \leftarrow \vec{\theta} + \eta \frac{\partial}{\partial \theta} \ell(Q \mid O, \theta) = \vec{\theta} + \eta \left[\sum_{t=1}^{T} \vec{f}_t - \sum_{t=1}^{T} \mathbb{E}_{Q \sim \Pr} \vec{f} \right]$$

Learning a CRF model

$$\vec{f} = \begin{bmatrix} f_1 \\ \vdots \\ f_d \end{bmatrix} = \begin{bmatrix} f_1(\text{no cheating}, \text{cheat}; H,T,H) = 1 \\ f_2(\underset{q_t}{\text{cheating}}, \overset{q_{t-1}}{\text{cheating}}; H,T,H) \\ \vdots \end{bmatrix}$$

A running example (Cheating Casino)

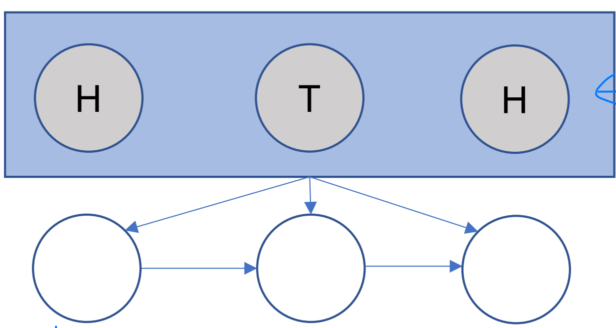
Alg: Random Init $\vec{\theta} = [0.1, 0.1, 0.1 \cdots 0.1] \in \mathbb{R}^d$

gradient ascent:

for $i = 1 \cdots n$

$$\vec{\theta}^{(i+1)} \leftarrow \vec{\theta}^{(i)} + \eta \left(\sum_{t=1}^{3} \vec{f}(q_{t-1}, q_t, O) - \sum_{t=1}^{3} E_{Q \sim Pr(Q|O,\theta)} \vec{f}(q_{t-1}, q_t, O) \right)$$

flips of coins



$\leftarrow O$

cheating — no cheating — cheating

q_1 — q_2 — q_3

$t=1$ — $t=2$ — $t=3$

$$E_{Q \sim Pr(Q|O,\theta)} \vec{f}(q_1, q_2, O)$$

$$= \sum_{Q=[q_1,q_2,q_3]} Pr(q_1, q_2, q_3 | O, \theta) \vec{f}(q_1, q_2, O)$$

$$= \underbrace{\sum_{q_3} Pr(q_3 | q_1, q_2, O, \theta)}_{=1} \sum_{q_1, q_2} Pr(q_1, q_2 | O, \theta) \vec{f}(q_1, q_2, O)$$

What is $Pr(q_1, q_2 | O, \theta)$? see the next page.

$$\Pr(q_1 q_2 ; \theta, O)$$

$$= \sum_{q_3} \frac{1}{Z(O)} \exp\left\{ \sum_{t=1}^{3} s_t(q_{t-1}, q_t ; \theta, O) \right\}$$

dummy fixed pos tag.

$$= \sum_{q_3} \frac{1}{Z(O)} \Big[\exp\{ s_1(q_0, q_1 ; \theta, O) \}$$

$$\times \exp\{ s_2(q_1, q_2 ; \theta, O) \}$$

$$\times \exp\{ s_3(q_2, q_3 ; \theta, O) \} \Big]$$

$$= \frac{1}{Z(O)} \underbrace{\exp\{ s_1(q_0, q_1 ; \theta, O) \}}_{} \quad \text{forward prob.} = \alpha_1(q_1)$$

$$\times \exp\{ s_2(q_1, q_2 ; \theta, O) \}$$

$$\times \underbrace{\sum_{q_3} \exp\{ s_3(q_2, q_3 ; \theta, O) \}}_{\text{backward prob.} \; = \beta_2(q_2)}$$

$$= \frac{1}{Z(O)} \alpha_1(q_1) \exp\{ s_2(q_1, q_2 ; \theta, O) \} \beta_2(q_2)$$

$$Z(O) = \sum_{q_1 q_2 q_3} \exp\{ s_1 + s_2 + s_3 \}$$

$$= \sum_{q_1 q_2 q_3} \exp\{ s_1(q_0, q_1) + s_2(q_1, q_2) + s_3(q_2 q_3) \}$$

$$= \sum_{q_3} \underbrace{\Bigg[\Bigg[\sum_{q_2} \overbrace{\sum_{q_1} \exp\{ s_1 + s_2 \}}^{= \alpha_2(q_2)} \Bigg] \exp\{ s_3(q_2 q_3) \} \Bigg]}_{\alpha_3(q_3)}$$