Machine Learning
CS690

Lecture 09

Razvan C. Bunescu

School of Electrical Engineering and Computer Science

bunescu@ohio.edu

Graphical Models

- In many supervised learning tasks, the entities to be labeled are related to each other:
  - hyperlinked web pages
  - cross-citations in scientific papers
  - social networks

- Standard approach: classify each entity independently

  => *flat models*

- Alternative approach: collective classification using undirected graphical models

  => *relational models*

Graphical Models

- An intuitive representation of conditional independence between domain variables:
  - **Directed Models** => well suited to represent temporal and causal relationships (*Bayesian Networks, NNs, HMMs*)
  - **Undirected Models** => appropriate for representing statistical correlation between variables (*Markov Networks*)

  - **Generative Models** => define a joint probability over observation and label sequences (*HMMs*)
  - **Discriminative Models** => specifies a probability over label sequences given an observation sequence (*CRFs*)

Markov Random Fields (MRF)

- V – a set of (discrete) random variables
- $G = (V, E)$ an undirected graph

Definition:

V is said to be a *Markov Random Field* with respect to G if:

$$P(V_i \mid V - V_i) = P(V_i \mid N(V_i)) \quad \text{, where } N(V_i) = \{V_j \mid (V_i, V_j) \in E\}$$

i.e. $N(V_i)$ is the *neighborhood* of V_i

Gibbs Random Fields (GRF)

- $G = (V, E)$ – an undirected graph
  - V is a set of (discrete) random variables
  - $C(G)$ is the set of all cliques of G
  - V_c is the set of vertices in a clique $c \in C(G)$

Definition:

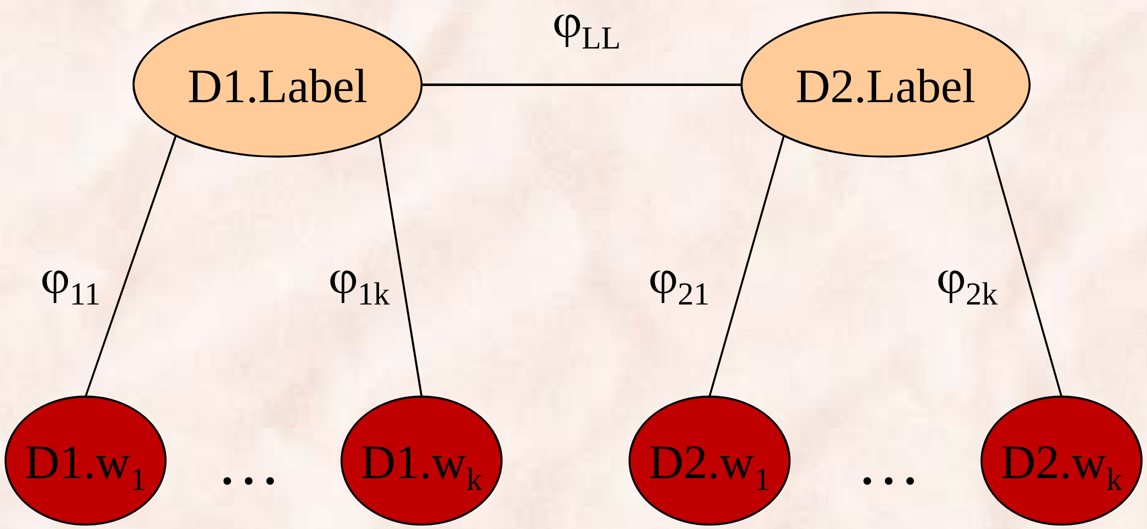
V is said to be a *Gibbs Random Field* with respect to G if:

$$P(V) = \frac{1}{Z} \exp \sum_{c \in C(G)} \varphi_c(V_c)$$

$\Phi = \{\varphi_c \mid \varphi_c : V_c \to R,\ c \in C(G)\}$ is the set of *clique potentials*

Z is the normalization constant

Gibbs Random Fields – Example



φ_{LL}	D1.Label	D2.Label
$\varphi_{LL}(0,0)$	0	0
$\varphi_{LL}(0,1)$	0	1
$\varphi_{LL}(1,0)$	1	0
$\varphi_{LL}(1,1)$	1	1

- D1, D2 are linked webpages
- D.Label $\in$ {0,1}
- D.w is true if word w $\in$D, otherwise false
- k is the size of the vocabulary

φ_{1j}	D1.Label	D1.w_j
φ_{1j}(0,false)	0	false
φ_{1j}(0,true)	0	true
φ_{1j}(1,false)	1	false
φ_{1j}(1,true)	1	true

Markov-Gibbs Equivalence

- A GRF is characterized by its global property
  => *the Gibbs distribution*
- An MRF is characterized by its local property
  => *the Markov assumption*

Theorem [Hammersley & Clifford, 1971]

V is an *MRF* w.r.t. G ⇔ V is a *GRF* w.r.t. G

Discriminative MRF (CRF)

- $V = X \cup Y$ is a set of discrete random variables:
  - X are *observed* variables
  - Y are *hidden* variables (labels)
- $G = (V, E)$ is an undirected graph.

Definition:

V is said to be a *Conditional Random Field* (*CRF*) w.r.t. G if:

$$P(Y_i \mid X, Y - Y_i) = P(Y_i \mid X, N(V_i)) \quad \text{, where } N(Y_i) = \{Y_j | (Y_i, Y_j) \in E\}$$

i.e. $N(Y_i)$ is the *neighborhood* of Y_i

[Lafferty, McCallum & Pereira 2000]

Discriminative GRF (CMN)

- $V = X \cup Y$ is a set of discrete random variables
  - X are *observed* variables
  - Y are *hidden* variables (labels)
- $G = (V, E)$ is an undirected graph:
  - $C(G)$ are the cliques of G
  - $V_c = X_c \cup Y_c$ is the set of vertices in a clique $c \in C(G)$

Definition:

V is said to be a *Conditional Markov Network* w.r.t. G if:

$$P(Y \mid X) = \frac{1}{Z(X)} \exp \sum_{c \in C(G)} \varphi_c(X_c, Y_c)$$

$Z(X)$ is the normalization constant

[Taskar, Abbeel & Koller 2002]

Markov-Gibbs Equivalence

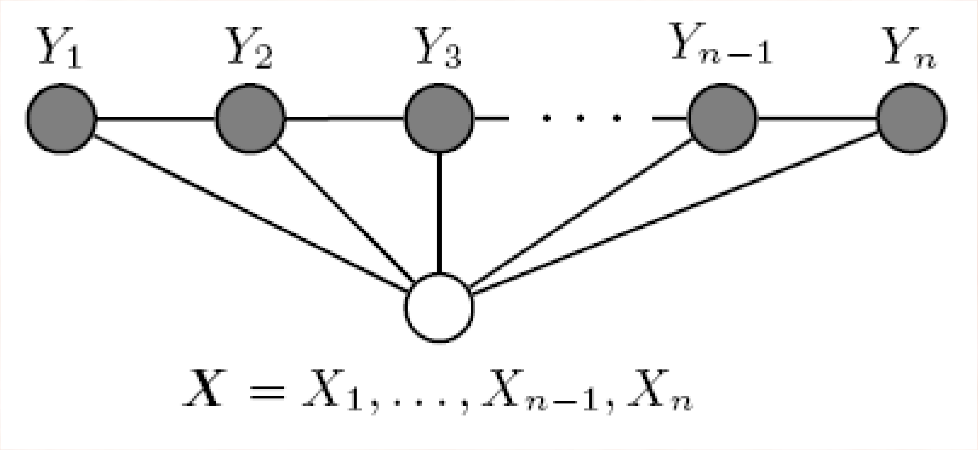
Theorem [Hammersley & Clifford, 1971] :

V is a *Conditional Random Field* w.r.t. G

$\Leftrightarrow$ V is a *Conditional Markov Network* w.r.t. G

[Lafferty, McCallum & Pereira 2001]

Linear-Chain CRFs



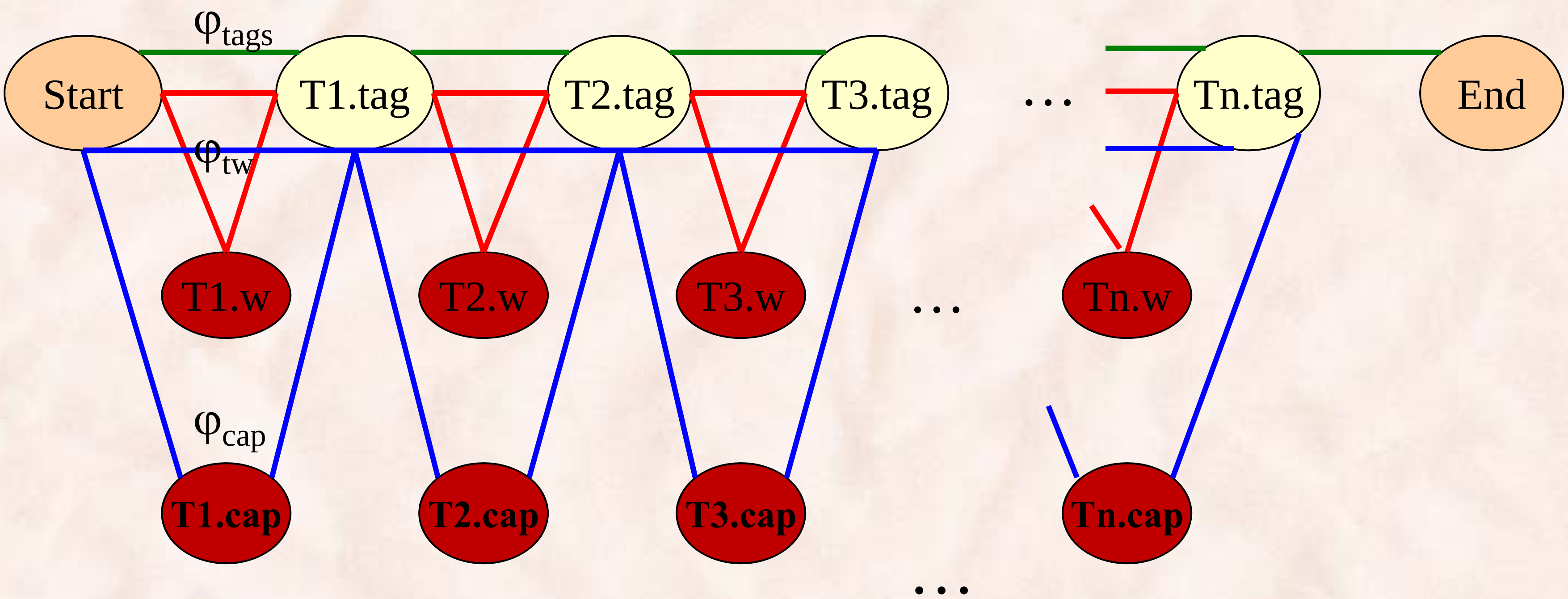
$$F_j(\mathbf{y}, \mathbf{x}) = \sum_{i=1}^{n} f_j(y_{i-1}, y_i, \mathbf{x}, i)$$

$$P(\mathbf{y} \mid \mathbf{x}, \lambda) = \frac{1}{Z(\mathbf{x})} \exp \sum_j \underbrace{\lambda_j F_j(\mathbf{y}, \mathbf{x})}_{\varphi_j(\mathbf{y}, \mathbf{x})}$$

[Lafferty, McCallum & Pereira 2001]

Part-of-speech Tagging

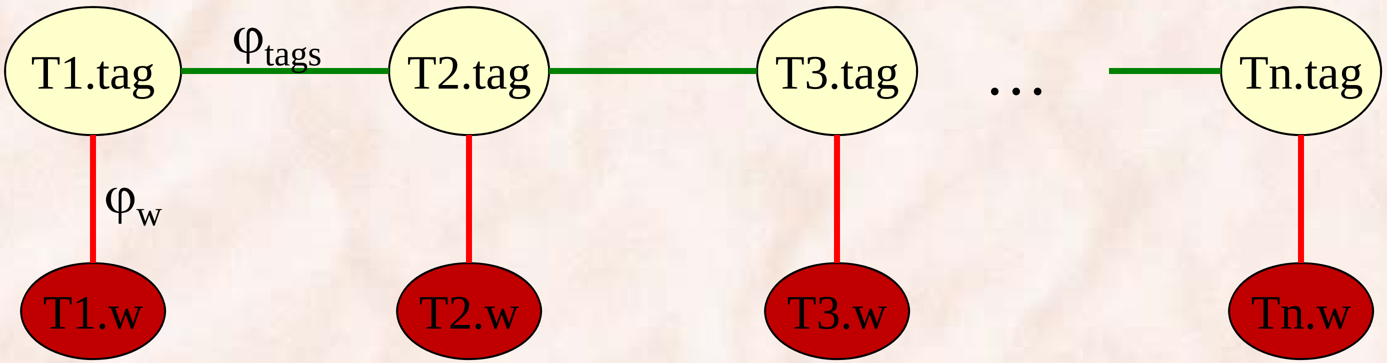
Sentence *S* = a sequence of tokens *T1*, …, *Tn (tokens as entities)*



- *Tj.tag* – the POS tag at position *j*
- *Tj.w* – *true* if word *w* occurs at position *j*
- *Tj.cap* – *true* if word at position *j* begins with capital letter
- …

[Lafferty, McCallum & Pereira 2000]

“Discriminative HMMs”



φ_{tags} and φ_w play a similar role to the (logarithms of the) usual HMM parameters $P(T_{j+1}.tag|T_j.tag)$ *and* $P(T.w|T.tag)$.

Inference in Linear Chain CRFs

Learning with Linear Chain CRFs