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
  \[\Rightarrow\textit{flat models}\]

- Alternative approach: collective classification using undirected graphical models
  \[\Rightarrow\textit{relational models}\]
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:

$$P(V_i \mid V - V_i) = P(V_i \mid N(V_i)),$$

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:

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

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

$Z$ is the normalization constant
Gibbs Random Fields – Example

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

- A GRF is characterized by its global property
  $\Rightarrow$ the Gibbs distribution
- An MRF is characterized by its local property
  $\Rightarrow$ the Markov assumption

**Theorem [Hammersley & Clifford, 1971]**

$V$ is an MRF w.r.t. $G \Leftrightarrow V$ 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:

$$P(Y_i \mid X, Y - Y_i) = P(Y_i \mid X, N(Y_i))$$

where $N(Y_i) = \{Y_j \mid (Y_i, Y_j) \in E\}$

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] Lecture 09
Markov-Gibbs Equivalence

**Theorem** [Hammersley & Clifford, 1971] :

\[ V \text{ is a Conditional Random Field w.r.t. } G \]

\[ \Leftrightarrow V \text{ is a Conditional Markov Network w.r.t. } G \]
Linear-Chain CRFs

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

\[ P(y \mid x, \lambda) = \frac{1}{Z(x)} \exp \sum_{j} \lambda_j F_j(y, x) \phi_j(y, x) \]

[Lafferty, McCallum & Pereira 2001]
Part-of-speech Tagging

Sentence $S = \text{a sequence of tokens } T_1, \ldots, T_n$ (tokens as entities)

- $T_j.tag$ – the POS tag at position $j$
- $T_j.w$ – true if word $w$ occurs at position $j$
- $T_j.cap$ – true if word at position $j$ begins with capital letter
- $\ldots$

[Lafferty, McCallum & Pereira 2001]
"Discriminative HMMs"

\( \phi_{\text{tags}} \) and \( \phi_w \) play a similar role to the (logarithms of the) usual HMM parameters \( P(T_{j+1}.\text{tag}|T_j.\text{tag}) \) and \( P(T.w|T.\text{tag}) \).

[Lafferty, McCallum & Pereira 2000]
Inference in Linear Chain CRFs
Learning with Linear Chain CRFs