General Formulations for Structures: Conditional Random Fields

CS 6355: Structured Prediction

THE UNIVERSITY OF UTAH

So far…

- Binary and multiclass classifiers
- Graphs as structured output
- Sequence labeling

This lecture

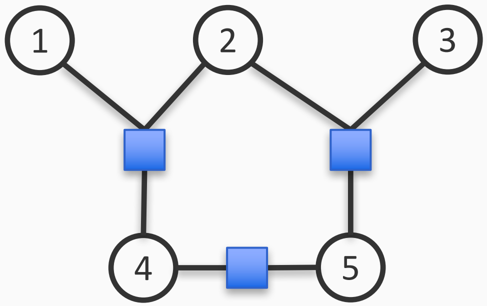
- Graphical models
  - Bayesian Networks
  - Markov Random Fields

- Formulations of structured output
  - Joint models
    - Markov Logic Network
  - Conditional models
    - Conditional Random Fields (again)
    - Constrained Conditional Models

Where are we?

- Graphical models
  - Bayesian Networks
  - Markov Random Fields
- Formulations of structured output
  - Joint models
    - Markov Logic Network
  - Conditional models
    - Conditional Random Fields (again)
    - Constrained Conditional Models

Recall: Markov Random Fields



We have seen Markov random fields

- Joint probability over a set of random variables
- Factors connect random variables

- Each factor is associated with a potential function
  - Typically, but not necessarily, exponential functions

- A way of decomposing joint probability
  - Product of potentials over factors

$$P(\mathbf{x}) \propto f(x_1, x_2, x_4)\, f(x_2, x_3, x_5)\, f(x_4, x_5)$$

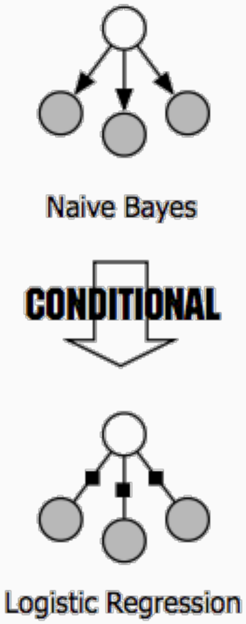
- (Markov Logic Networks provide a concise language for defining MRFs)

Another look at conditional random fields

- A MRF models a joint distribution
  - Think P(**x**, **y**)
    - In fact, no separation of variables into inputs and outputs
  - A generative model

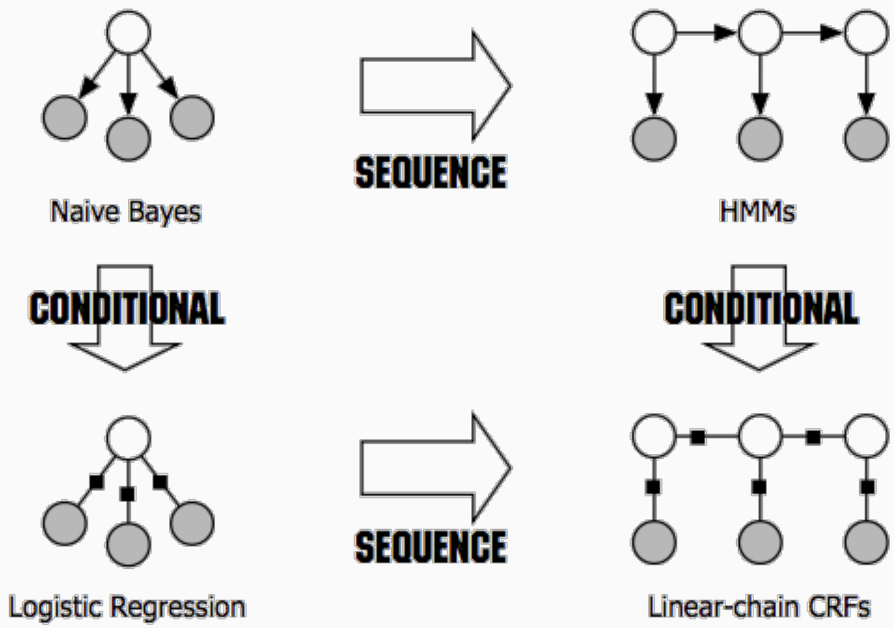
- A CRF is
  - A discriminative version of the MRF
    - Model P(**y** | **x**)
    - No factors that involve only the **x**'s
  - A structured extension of logistic regression

From generative models to CRF



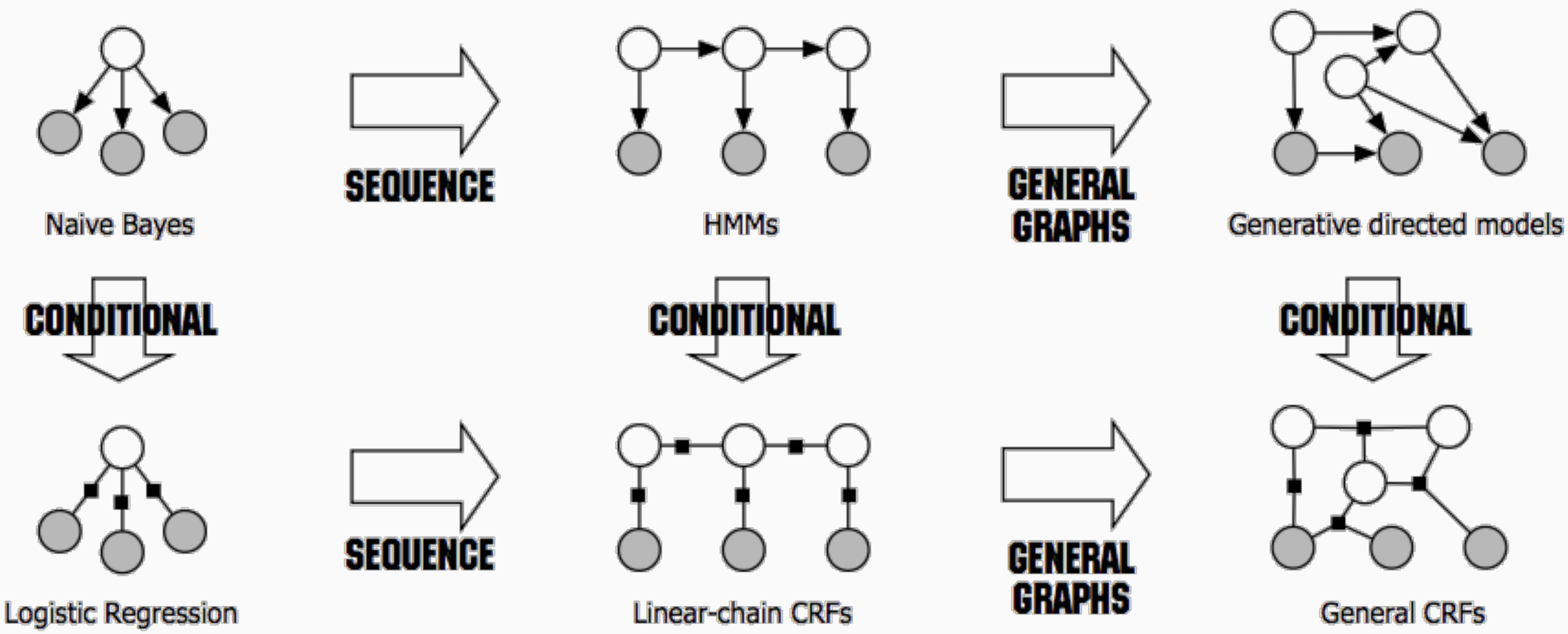
[Figure from Sutton and McCallum, '05]

From generative models to CRF



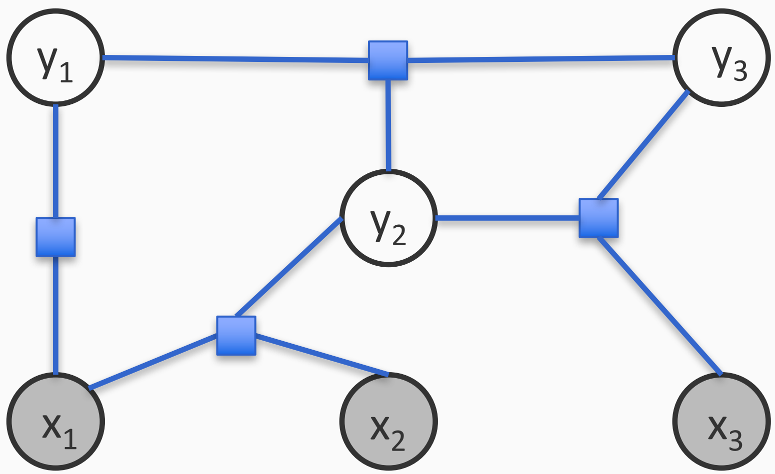
[Figure from Sutton and McCallum, ‘05]

From generative models to CRF

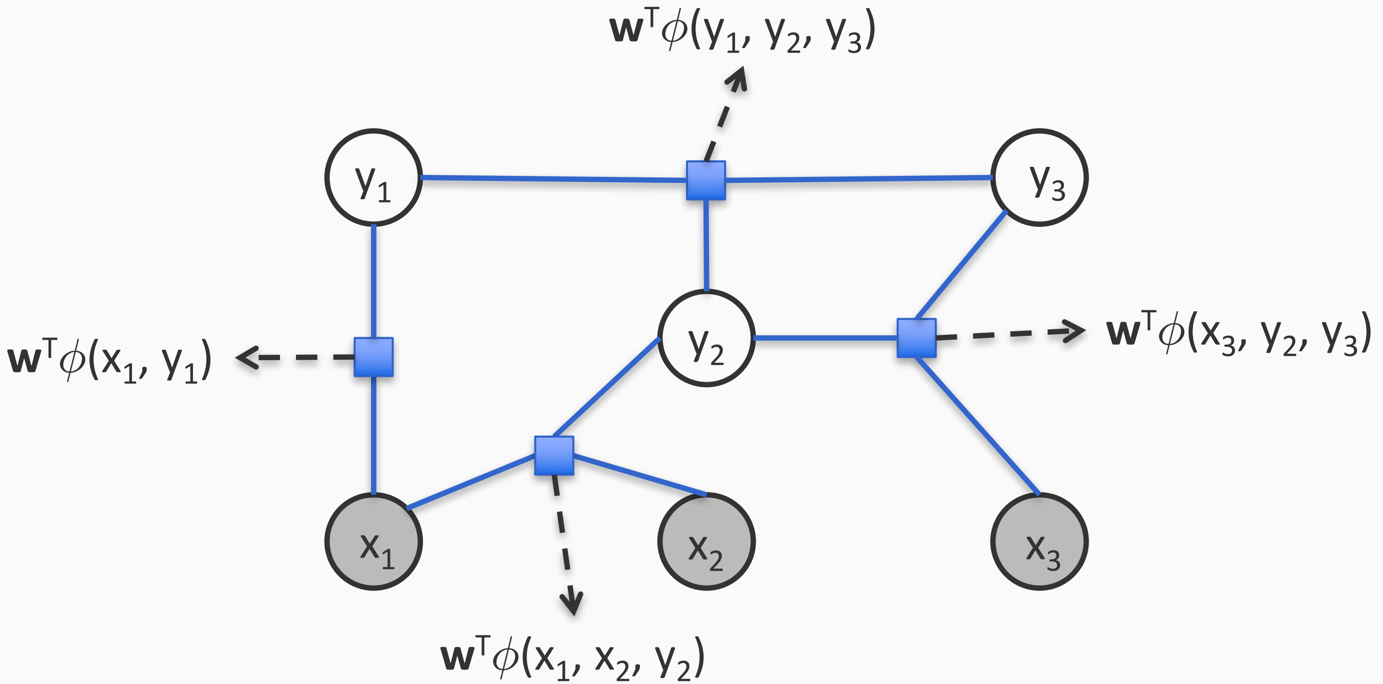


[Figure from Sutton and McCallum, ‘05]

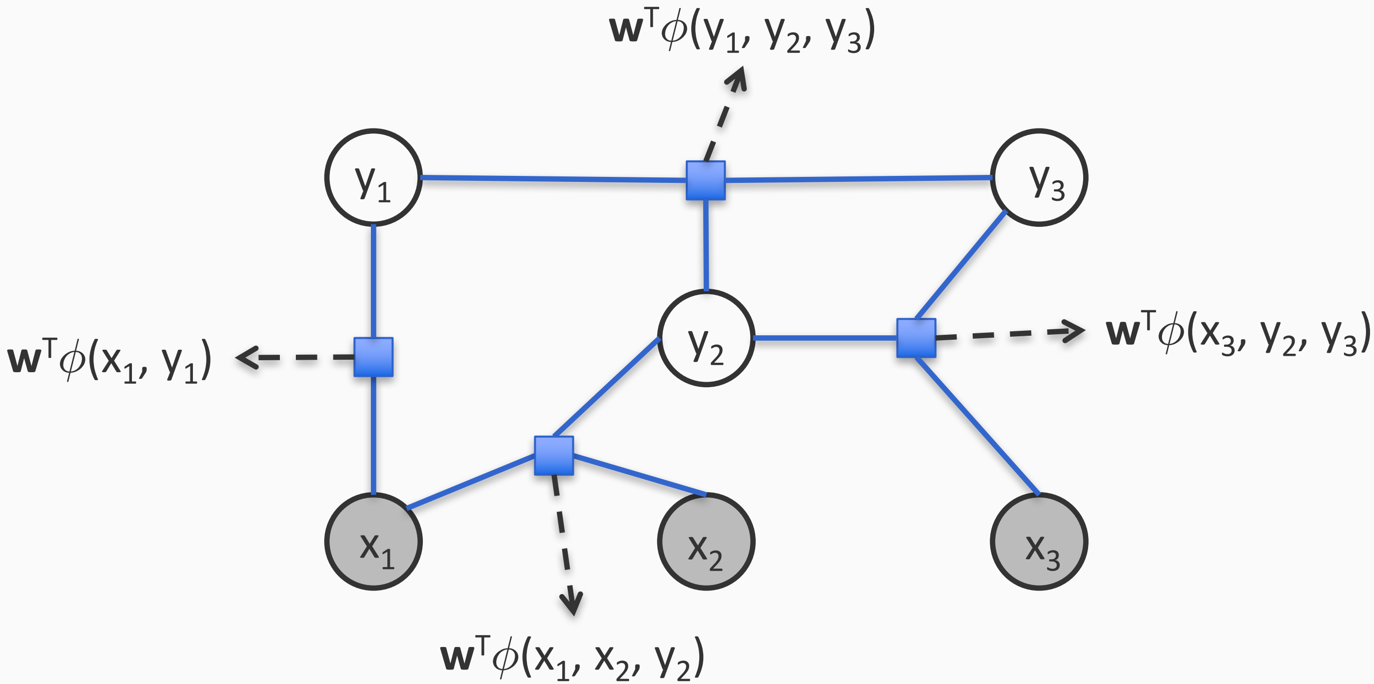
General CRFs



General CRFs



General CRFs



$$P(\mathbf{y}|\mathbf{x}) = \frac{\exp\left(\mathbf{w}^T\phi(\mathbf{x},\mathbf{y})\right)}{\sum_{\hat{\mathbf{y}}}\exp\left(\mathbf{w}^T\phi(\mathbf{x},\hat{\mathbf{y}})\right)}$$

$$\phi(\mathbf{x}, \mathbf{y}) = \phi(x_1, y_1) + \phi(y_1, y_2, y_3) + \phi(x_3, y_2, y_3) + \phi(x_1, x_2, y_2)$$

Computational questions

1. Learning: Given a training set {<$\mathbf{x}_i$, $\mathbf{y}_i$>}
   - Train via maximum likelihood (typically regularized)

$$\max_{\mathbf{w}} \sum_i \log P(\mathbf{y}_i|\mathbf{x}_i, \mathbf{w}) = \max_{\mathbf{w}} \sum_i \mathbf{w}^T \phi(\mathbf{x}_i, \mathbf{y}_i) - \log Z_{\mathbf{w}}(\mathbf{x}_i)$$

   - Need to compute partition function during training

$$Z_{\mathbf{w}}(\mathbf{x}_i) = \sum_{\mathbf{y}} \exp(\mathbf{w}^T \phi(\mathbf{x}_i, \mathbf{y}))$$

Computational questions

1. Learning: Given a training set $\{<\mathbf{x}_i, \mathbf{y}_i>\}$
   - Train via maximum likelihood (typically regularized)

$$\max_{\mathbf{w}} \sum_i \log P(\mathbf{y}_i | \mathbf{x}_i, \mathbf{w}) = \max_{\mathbf{w}} \sum_i \mathbf{w}^T \phi(\mathbf{x}_i, \mathbf{y}_i) - \log Z_{\mathbf{w}}(\mathbf{x}_i)$$

   - Need to compute partition function during training

$$Z_{\mathbf{w}}(\mathbf{x}_i) = \sum_{\mathbf{y}} \exp(\mathbf{w}^T \phi(\mathbf{x}_i, \mathbf{y}))$$

Computational questions

1. Learning: Given a training set $\{<\mathbf{x}_i, \mathbf{y}_i>\}$
   – Train via maximum likelihood (typically regularized)

$$\max_{\mathbf{w}} \sum_i \log P(\mathbf{y}_i|\mathbf{x}_i, \mathbf{w}) = \max_{\mathbf{w}} \sum_i \mathbf{w}^T \phi(\mathbf{x}_i, \mathbf{y}_i) - \log Z_{\mathbf{w}}(\mathbf{x}_i)$$

   – Need to compute partition function during training

$$Z_{\mathbf{w}}(\mathbf{x}_i) = \sum_{\mathbf{y}} \exp(\mathbf{w}^T \phi(\mathbf{x}_i, \mathbf{y}))$$

2. Prediction $\max_{\mathbf{y}} \mathbf{w}^T \phi(\mathbf{x}, \mathbf{y})$
   – Go over all possible assignments to the **y**'s
   – Find the one with the highest probability/score