Conditional Random Fields

Natalie Parde
UIC CS 421

We’ve learned about a variety of text classification techniques….

- Hidden Markov Models
- Naïve Bayes
- Logistic Regression

Types of Classification Techniques

Label Type

- **Individual Labels**
  - Naïve Bayes
  - Logistic Regression
- **Sequences of Labels**
  - Hidden Markov Models
  - Conditional Random Fields

Model Type

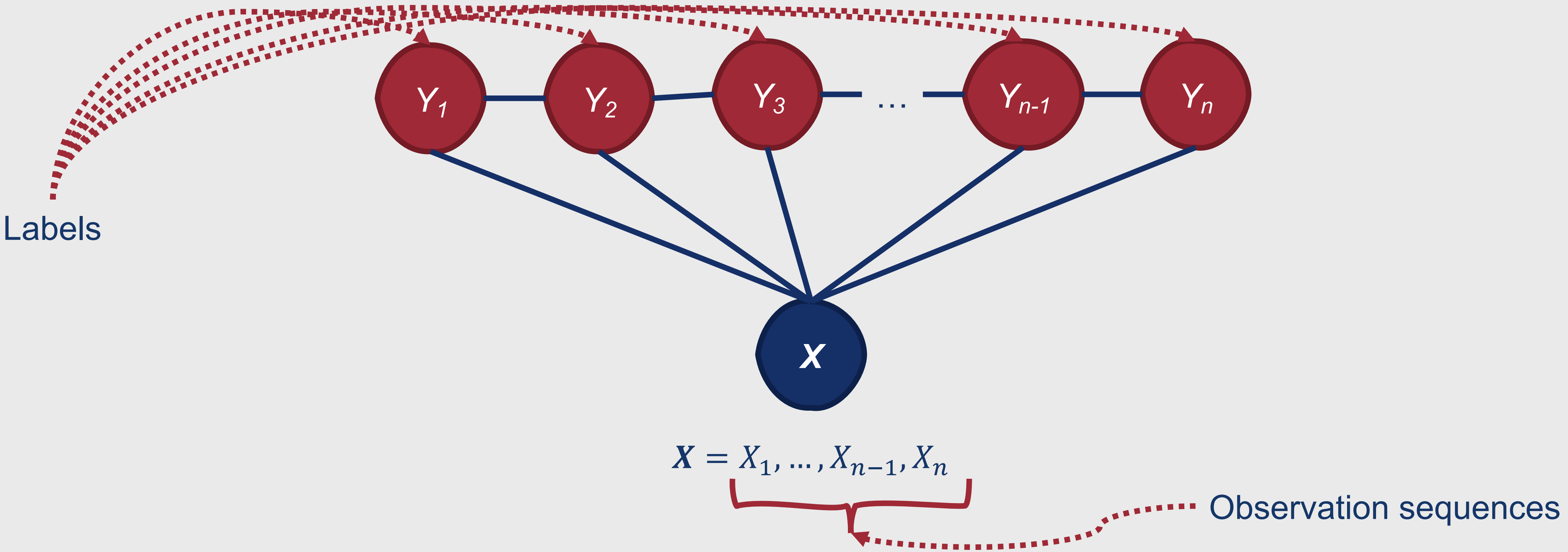
- **Generative**
  - Naïve Bayes
  - Hidden Markov Models
- **Discriminative**
  - Logistic Regression
  - Conditional Random Fields

Conditional Random Fields (CRFs)

- **Generalized multi-class logistic regression**
- Increased flexibility for sequence labeling
  - **HMMs: Joint probability** ranging over observations and corresponding labels
    - Can lead to rigid (and inaccurate) independence assumptions
  - **CRFs: Conditional probability** over label sequences given specific sequence of observations
    - Relaxes independence assumptions (model may more easily capture arbitrary or long-range dependencies)

Special Case of Markov Random Fields

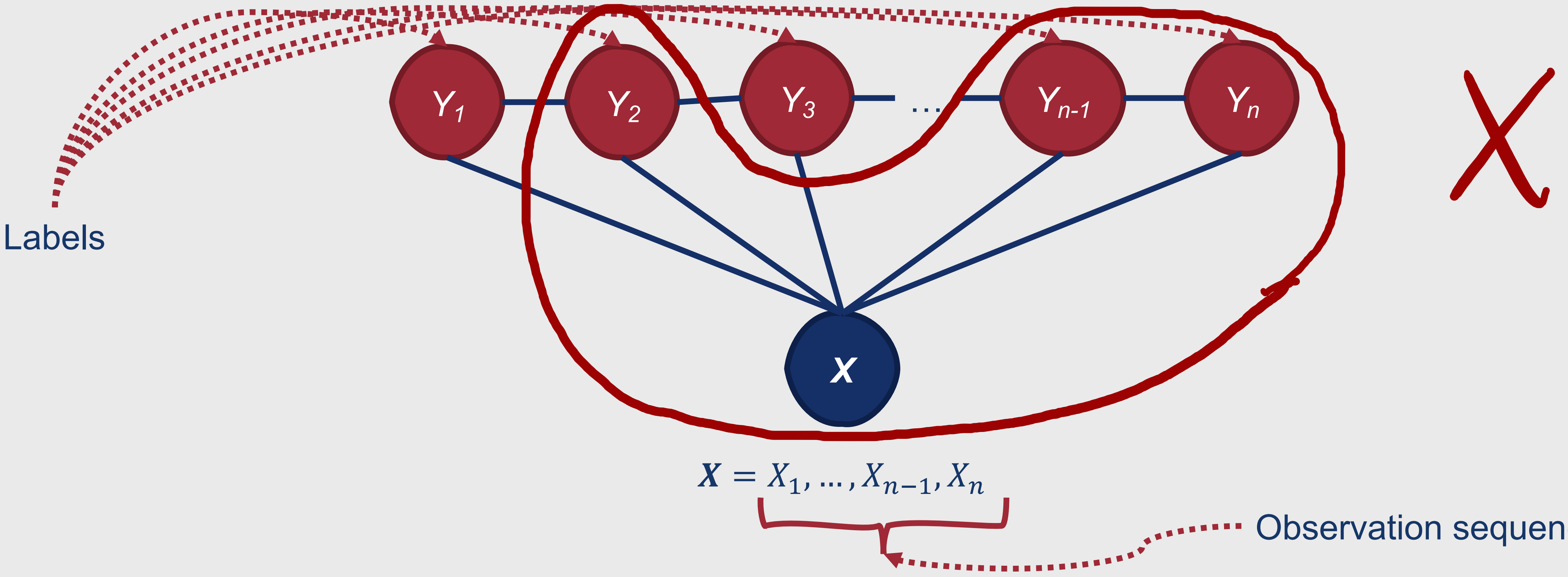
- Undirected graphical model



Special Case of Markov Random Fields

- Undirected graphical model

Conditionally independent labels cannot appear in the same potential function

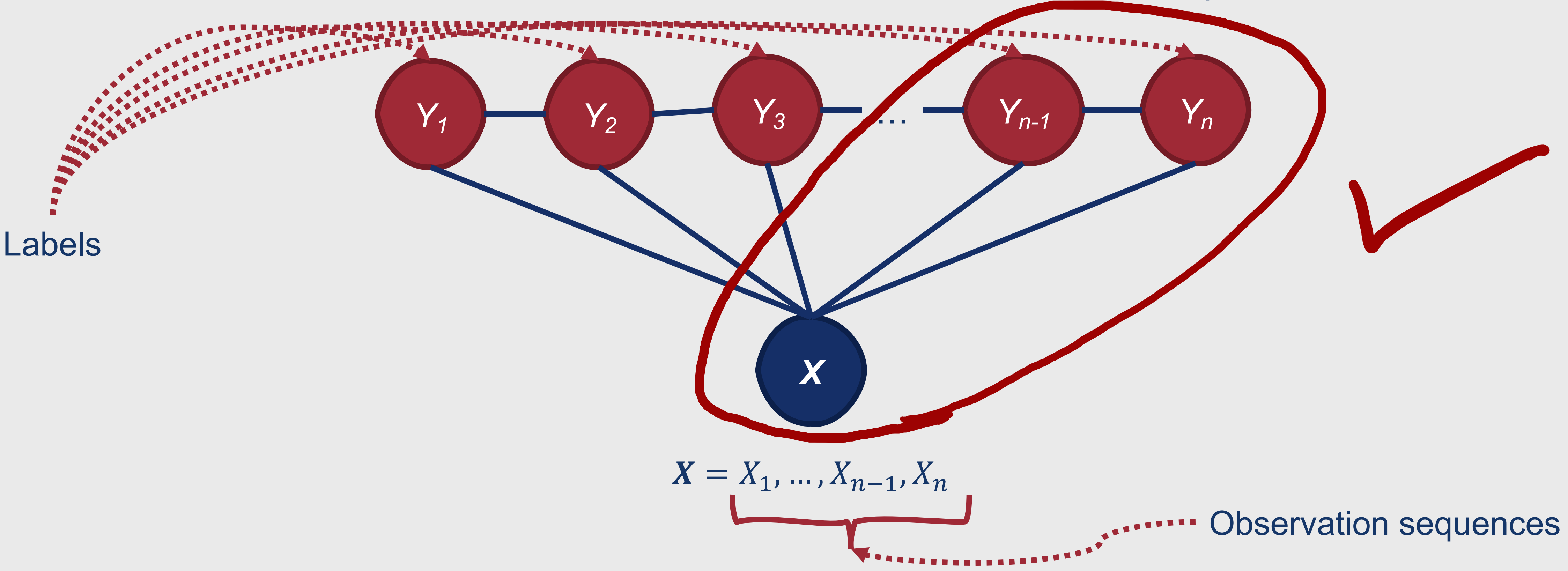


$$\boldsymbol{X} = \underbrace{X_1, \ldots, X_{n-1}, X_n}$$

Special Case of Markov Random Fields

- Undirected graphical model

Instead, require potential functions to operate only on random variables forming a maximal clique



Conditional Random Fields

- Probability of label sequence ***y*** given observation sequence ***x*** is then a normalized product of feature functions
  - $p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta}) = \underbrace{\frac{1}{Z(\boldsymbol{x})}}_{\text{Normalization factor}} e^{\sum_j \theta_j \underbrace{F_j(\boldsymbol{y},\boldsymbol{x})}_{\text{Feature function}}}$

Normalization factor

Feature function

$$F_j(\boldsymbol{y},\boldsymbol{x}) = \begin{cases} 1 & \text{if } x_1\text{=“COVID” and } y_1\text{=NOUN} \\ 0 & \text{otherwise} \end{cases}$$

Training CRFs

- Seek to find the model distribution with maximum entropy (distribution is as uniform as possible)
- Parameters can be optimized by minimizing cross-entropy loss
  - Log likelihood of a CRF:
    - $\mathcal{L}(\boldsymbol{\theta}) = \sum_k \left[\log \frac{1}{Z(\boldsymbol{x}^{(k)})} + \sum_j \theta_j F_j(\boldsymbol{y}^{(k)}, \boldsymbol{x}^{(k)})\right]$

Training CRFs

- Derivative of CRF log likelihood:
  - $\frac{\partial \mathcal{L}(\theta)}{\partial \theta_j} = E_{\tilde{p}(Y,X)}[F_j(Y,X)] - \sum_k E_{p(Y|x^{(k)},\theta)}[F_j(Y,x^{(k)})]$

Empirical distribution of training data

Expectation with respect to distribution *p*

How to efficiently compute expectation?

- Too many possible label sequences to compute naively
- Instead, we can turn to an old favorite …dynamic programming!
  - $E_{p(\boldsymbol{Y}|\boldsymbol{x}^{(k)},\boldsymbol{\theta})}\left[F_j(\boldsymbol{Y},\boldsymbol{x}^{(k)})\right] = \sum_{\boldsymbol{y}} p(\boldsymbol{Y}=\boldsymbol{y}|\boldsymbol{x}^{(k)},\boldsymbol{\theta})F_j(\boldsymbol{y},\boldsymbol{x}^{(k)})$
  - $p\left(Y_{i-1}=y', Y_i=y\middle|\boldsymbol{x}^{(k)},\boldsymbol{\theta}\right) = \frac{\alpha_{i-1}(y'|\boldsymbol{x})M_i(y',y|\boldsymbol{x})\beta_i(y|\boldsymbol{x})}{Z(\boldsymbol{x})}$

$$M_i(y',y|\boldsymbol{x}) = e^{\sum_j \theta_j f_j(y',y,x,i)}$$

$$Z(\boldsymbol{x}) = \left[\prod_{i=1}^{n+1} M_i(\boldsymbol{x})\right]$$

Check out Wallach (2004) for more details:
http://dirichlet.net/pdf/wallach04conditional.pdf