Conditional Random Fields

[Hanna M. Wallach, Conditional Random Fields: An Introduction, Technical Report MS-CIS-04-21, University of Pensylvania, 2004.]

CS 486/686

University of Waterloo

Lecture 19: November 13, 2012

Outline

- Conditional Random Fields

Conditional Random Fields

- CRF: special Markov network that represents a conditional distribution
- $\Pr(\mathbf{X}|\mathbf{E}) = 1/k(\mathbf{E})\ e^{\Sigma_j \lambda_j \phi_j(\mathbf{X},\mathbf{E})}$
  - NB: $k(\mathbf{E})$ is a normalization function (it is not a constant since it depends on $\mathbf{E}$ – see Slide 5)
- Useful in classification: Pr(class|input)
- Advantage: no need to model distribution over inputs

Conditional Random Fields

- Joint distribution:
  - $\Pr(\mathbf{X},\mathbf{E}) = 1/k\ e^{\Sigma_j \lambda_j \phi_j(\mathbf{X},\mathbf{E})}$
- Conditional distribution
  - $\Pr(\mathbf{X}|\mathbf{E}) = e^{\Sigma_j \lambda_j \phi_j(\mathbf{X},\mathbf{E})} / \Sigma_{\mathbf{X}}\ e^{\Sigma_j \lambda_j \phi_j(\mathbf{X},\mathbf{E})}$
- Partition features in two sets:
  - $\phi_{j1}(\mathbf{X},\mathbf{E})$: depend on at least one var in $\mathbf{X}$
  - $\phi_{j2}(\mathbf{E})$: depend only on evidence $\mathbf{E}$

Conditional Random Fields

- Simplified conditional distribution:
  - $\Pr(X|E) = \dfrac{e^{\Sigma_{j1} \lambda_{j1} \phi_{j1}(X,E) + \Sigma_{j2} \lambda_{j2} \phi_{j2}(E)}}{\Sigma_X e^{\Sigma_{j1} \lambda_{j1} \phi_{j1}(X,E) + \Sigma_{j2} \lambda_{j2} \phi_{j2}(E)}}$

$$= \frac{e^{\Sigma_{j1} \lambda_{j1} \phi_{j1}(X,E)}}{\Sigma_X e^{\Sigma_{j1} \lambda_{j1} \phi_{j1}(X,E)}} \cancel{\frac{e^{\Sigma_{j2} \lambda_{j2} \phi_{j2}(E)}}{e^{\Sigma_{j2} \lambda_{j2} \phi_{j2}(E)}}}$$

$$= 1/k(E)\, e^{\Sigma_{j1} \lambda_{j1} \phi_{j1}(X,E)}$$

- Evidence features can be ignored!

Parameter Learning

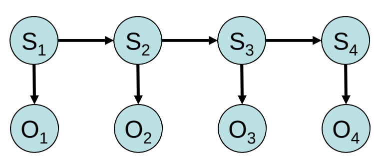
- Parameter learning is simplified since we don't need to model a distribution over the evidence
- Objective: maximum conditional likelihood
  - $\lambda^* = \text{argmax}_\lambda\ P(X=x|\lambda,E=e)$
  - Convex optimization, but no closed form
  - Use iterative technique (e.g., gradient descent)

Sequence Labeling

- Common task in
  - Entity recognition
  - Part of speech tagging
  - Robot localisation
  - Image segmentation

- $\mathbf{L^*} = \text{argmax}_{\mathbf{L}} \Pr(\mathbf{L}|\mathbf{O})$?
  $= \text{argmax}_{L_1,\ldots,L_n} \Pr(L_1,\ldots,L_n|O_1,\ldots,O_n)$?

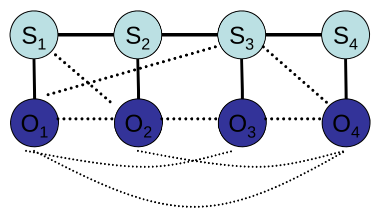
Hidden Markov Model



- Assumption: observations are independent given the hidden state

Conditional Random Fields

- Since the distribution over observations is not modeled, there is no independence assumption among observations



- Can also model long-range dependencies without significant computational cost

Entity Recognition

- Task: label each word with a predefined set of categories (e.g., person, organization, location, expression of time, etc.)
  - Ex: Jim bought 300 shares of Acme Corp. in 2006
    person nil nil nil nil org org nil time
- Possible features:
  - Is the word numeric or alphabetic?
  - Does the word contain capital letters?
  - Is the word followed by "Corp."?
  - Is the word preceded by "in"?
  - Is the preceding label an organization?

Next Class

- First-order logic