Markov Networks
[KF] Chapter 4

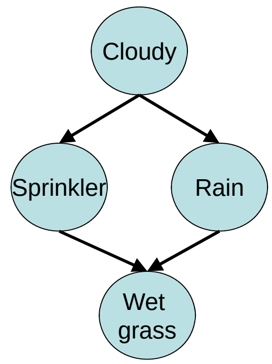
CS 786
University of Waterloo
Lecture 7: May 24, 2012

Outline

- Markov networks
  - a.k.a. Markov random fields
- Conditional random fields

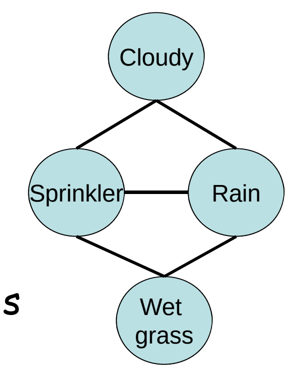
Recall Bayesian networks

- Directed acyclic graph
- Arcs often interpreted as causal relationships
- Joint distribution: product of conditional dist



Markov networks

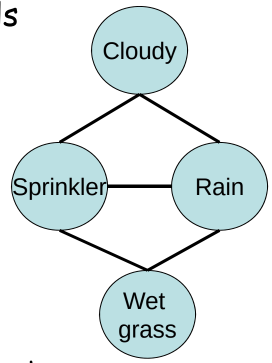
- Undirected graph
- Arcs simply indicate direct correlations
- Joint distribution: normalized product of potentials
- Popular in computer vision and natural language processing



Parameterization

- Joint: normalized product of potentials
  $Pr(\mathbf{X}) = 1/k \prod_j f_j(\mathbf{CLIQUE}_j)$
  $= 1/k\ f_1(C,S,R)\ f_2(S,R,W)$

  where k is a normalization constant
  $k = \Sigma_{X_i} \prod_j f_j(\mathbf{CLIQUE}_j)$
  $= \Sigma_{C,S,R,W}\ f_1(C,S,R)\ f_2(S,R,W)$



- Potential:
  - Non-negative factor
  - Potential for each maximal clique in the graph
  - Entries: "likelihood strength" of different configurations.

Potential Example

$f_1(C,S,R)$	
csr	3
cs~r	2.5
c~sr	5
c~s~r	5.5
~csr	0
~cs~r	2.5
~c~sr	0
~c~s~r	7

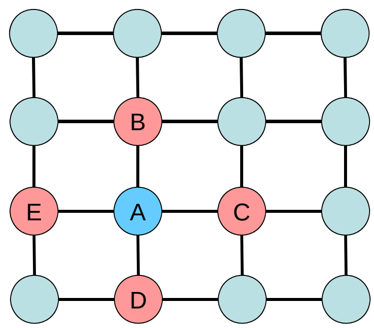
c~sr is more likely than cs~r

impossible configuration

Markov property

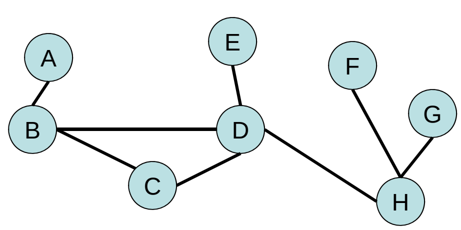
- Markov property: a variable is independent of all other variables given its immediate neighbours.
- Markov blanket: set of direct neighbours

MB(A) = {B,C,D,E}



Conditional Independence

- X and Y are independent given **Z** iff there doesn't exist any path between X and Y that doesn't contain any of the variables in **Z**
- Exercise:
  - A,E?
  - A,E|D?
  - A,E|C?
  - A,E|B,C?



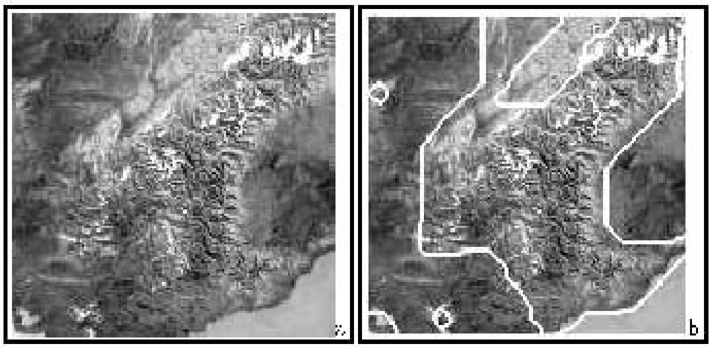
Interpretation

- Markov property has a price:
  - Numbers are not probabilities
- What are potentials?
  - They are indicative of local correlations
- What do the numbers mean?
  - They are indicative of the likelihood of each configuration
  - Numbers are usually learnt from data since it is hard to specify them by hand given their lack of a clear interpretation

Applications

- Natural language processing:
  - Part of speech tagging
- Computer vision
  - Image segmentation
- Any other application where there is no clear causal relationship

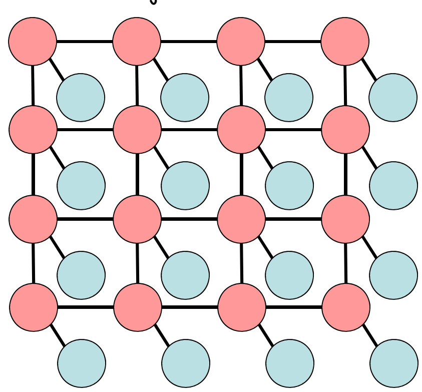
Image Segmentation



Segmentation of the Alps
Kervrann, Heitz (1995) A Markov Random Field model-based Approach to Unsupervised Texture Segmentation Using Local and Global Spatial Statistics, IEEE Transactions on Image Processing, vol 4, no 6, p 856-862

Image Segmentation

- Variables
  - Pixel features (e.g. intensities): X_{ij}
  - Pixel labels: Y_{ij}
- Correlations:
  - Neighbouring pixel labels are correlated
  - Label and features of a pixel are correlated
- Segmentation:
  - $\text{argmax}_Y \Pr(\mathbf{Y}|\mathbf{X})$?



Inference

- Markov nets: factored representation
  - Use variable elimination
- $P(\mathbf{X}|\mathbf{E}=\mathbf{e})$?
  - Restrict all factors that contain **E** to **e**
  - Sumout all variables that are not X or in **E**
  - Normalize the answer

Parameter Learning

- Maximum likelihood
  - $\theta^* = \text{argmax}_\theta P(\text{data}|\theta)$
- Complete data
  - Convex optimization, but no closed form solution
  - Iterative techniques such as gradient descent
- Incomplete data
  - Non-convex optimization
  - EM algorithm

Maximum likelihood

- Let θ be the set of parameters and $\mathbf{x}_i$ be the i^{th} instance in the dataset
- Optimization problem:
  - $\theta^* = \text{argmax}_\theta P(data|\theta)$
    $= \text{argmax}_\theta \Pi_i \Pr(\mathbf{x}_i|\theta)$
    $= \text{argmax}_\theta \Pi_i \frac{\Pi_j f(\mathbf{X}[j]=\mathbf{x}_i[j])}{\Sigma_{\mathbf{X}} \Pi_j f(\mathbf{X}[j]=\mathbf{x}_i[j])}$

    where $\mathbf{X}[j]$ is the clique of variables that potential j depends on and $\mathbf{x}[j]$ is a variable assignment for that clique

Maximum likelihood

- Let $\theta_{\mathbf{x}} = f(\mathbf{X}=\mathbf{x})$
- Optimization continued:
  - $\theta^* = \text{argmax}_\theta \Pi_i \frac{\Pi_j \theta_{\mathbf{x}_i[j]}}{\Sigma_{\mathbf{X}} \Pi_j \theta_{\mathbf{x}_i[j]}}$
    $= \text{argmax}_\theta \log \Pi_i \frac{\Pi_j \theta_{\mathbf{x}_i[j]}}{\Sigma_{\mathbf{X}} \Pi_j \theta_{\mathbf{x}_i[j]}}$
    $= \text{argmax}_\theta \Sigma_i \; \Sigma_j \log \theta_{\mathbf{x}_i[j]} - \log \Sigma_{\mathbf{X}} \Pi_j \theta_{\mathbf{x}_i[j]}$
- This is a non-concave optimization problem

Maximum likelihood

- Substitute $\lambda = \log \theta$ and the problem becomes **concave**:
  - $\lambda^* = \text{argmax}_\lambda \sum_i \sum_j \lambda_{x_i[j]} - \log \sum_x e^{\sum_j \lambda_{x_i[j]}}$
- Possible algorithms:
  - *Gradient ascent*
  - *Conjugate gradient*

CS786 Lecture Slides (c) 2012 P. Poupart

17

Feature-based Markov Networks

- Generalization of Markov networks
  - May not have a corresponding graph
  - Use features and weights instead of potentials
  - Use exponential representation
- $\Pr(\mathbf{X}=\mathbf{x}) = 1/k\ e^{\sum_j \lambda_j \phi_j(\mathbf{x}[j])}$
  where x[j] is a variable assignment for a subset of variables specific to ϕ_j
- Feature ϕ_j: Boolean function that maps partial variable assignments to 0 or 1
- Weight λ_j: real number

CS786 Lecture Slides (c) 2012 P. Poupart

18

Feature-based Markov Networks

- Potential-based Markov networks can always be converted to feature-based Markov networks

$$Pr(\mathbf{x}) = 1/k \, \Pi_j \, f_j(\mathbf{CLIQUE}_j = \mathbf{x}[j])$$
$$= 1/k \, e^{\Sigma_{j,\mathbf{clique}_j} \lambda_{j,\mathbf{clique}_j} \phi_{j,\mathbf{clique}_j}(\mathbf{x}[j])}$$

- $\lambda_{j,\mathbf{clique}_j} = \log f_j(\mathbf{CLIQUE}_j = \mathbf{x}[j])$
- $\phi_{j,\mathbf{clique}_j}(\mathbf{x}[j])$=1 if $\mathbf{clique}_j$=$\mathbf{x}[j]$, 0 otherwise

Example

$f_1(C,S,R)$	
csr	3
cs~r	2.5
c~sr	5
c~s~r	5.5
~csr	0
~cs~r	2.5
~c~sr	0
~c~s~r	7

weights	features	
$\lambda_{1,csr}$ = log 3	$\phi_{1,csr}$ (CSR) =	1 if CSR = csr
		0 otherwise
$\lambda_{1,*s\sim r}$ = log 2.5	$\phi_{1,*s\sim r}$(CSR) =	1 if CSR = *s~r
		0 otherwise
$\lambda_{1,c\sim sr}$ = log 5	$\phi_{c\sim sr}$(CSR) =	1 if CSR = c~sr
		0 otherwise
$\lambda_{1,c\sim s\sim r}$ = log 5.5	$\phi_{1,c\sim s\sim r}$ (CSR) =	1 if CSR = c~s~r
		0 otherwise
$\lambda_{1,\sim c*r}$ = log 0	$\phi_{1,\sim c*r}$(CSR) =	1 if CSR = ~c*r
		0 otherwise
$\lambda_{1,\sim c\sim s\sim r}$ = log 7	$\phi_{\sim c\sim s\sim r}$(CSR) =	1 if CSR = ~c~s~r
		0 otherwise

Features

- Features
  - Any Boolean function
  - Provide tremendous flexibility
- Example: text categorization
  - Simplest features: presence/absence of a word in a document
  - More complex features
    - Presence/absence of specific expressions
    - Presence/absence of two words within a certain window
    - Presence/absence of any combination of words
    - Presence/absence of a figure of style
    - Presence/absence of any linguistic feature

CS786 Lecture Slides (c) 2012 P. Poupart

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(**E**) is a normalization function (it is not a constant since it depends on **E** – see Slide 5)
- Useful in classification: Pr(class|input)
- Advantage: no need to model distribution over inputs

CS786 Lecture Slides (c) 2012 P. Poupart

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 **X**
  - $\phi_{j2}(\mathbf{E})$: depend only on evidence **E**

Conditional Random Fields

- Simplified conditional distribution:
  - $$Pr(X|E) = \frac{e^{\Sigma_{j1} \lambda_{j1} \phi_{j1}(\mathbf{X},\mathbf{E}) + \Sigma_{j2} \lambda_{j2} \phi_{j2}(\mathbf{E})}}{\Sigma_{\mathbf{X}} e^{\Sigma_{j1} \lambda_{j1} \phi_{j1}(\mathbf{X},\mathbf{E}) + \Sigma_{j2} \lambda_{j2} \phi_{j2}(\mathbf{E})}}$$
    $$= \frac{e^{\Sigma_{j1} \lambda_{j1} \phi_{j1}(\mathbf{X},\mathbf{E})}}{\Sigma_{\mathbf{X}} e^{\Sigma_{j1} \lambda_{j1} \phi_{j1}(\mathbf{X},\mathbf{E})}} \cdot \frac{\cancel{e^{\Sigma_{j2} \lambda_{j2} \phi_{j2}(\mathbf{E})}}}{\cancel{e^{\Sigma_{j2} \lambda_{j2} \phi_{j2}(\mathbf{E})}}}$$
    $$= 1/k(\mathbf{E}) \; e^{\Sigma_{j1} \lambda_{j1} \phi_{j1}(\mathbf{X},\mathbf{E})}$$
- Evidence features can be ignored!

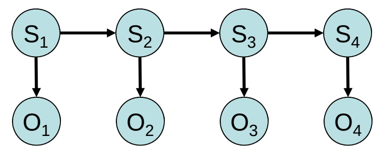
Parameter Learning

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

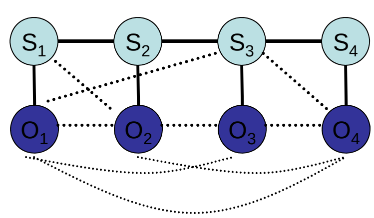
Hidden Markov Model



- 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

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?