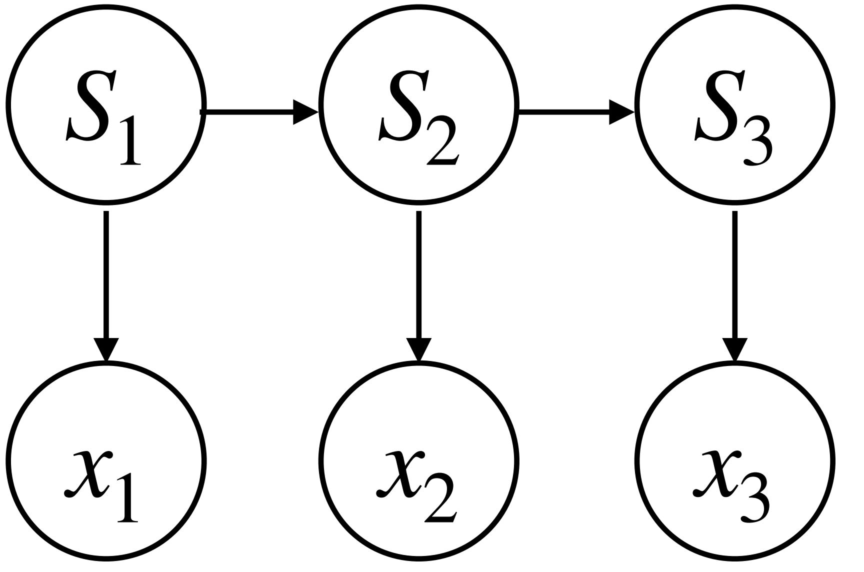
Markov Networks

Lili Mou
lmou@ualberta.ca
lili-mou.github.io

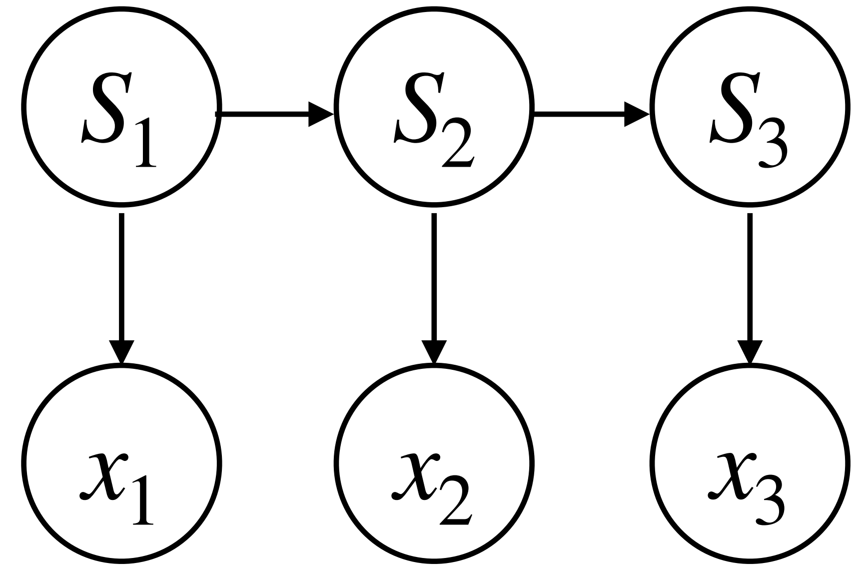
Pros & Cons of HMM



Pros

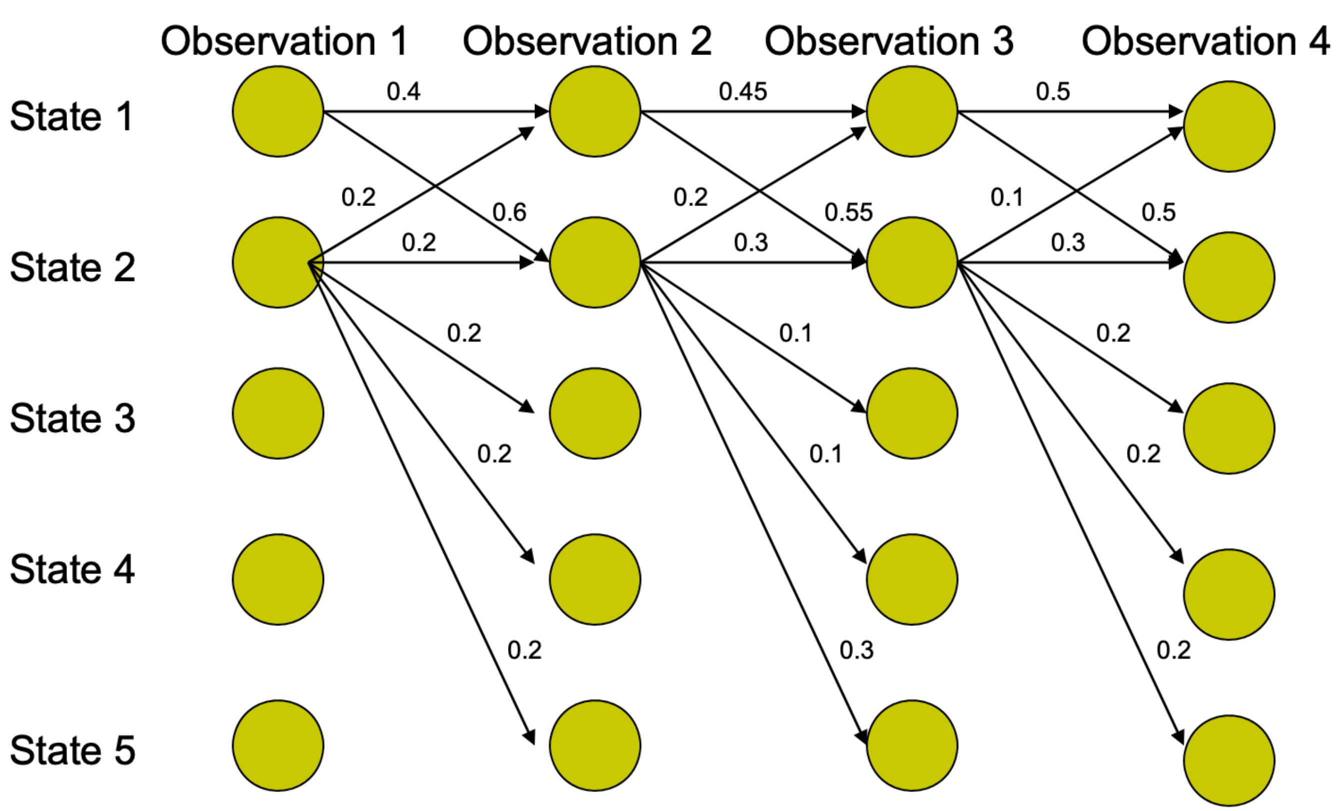
- Model the relationship among different time steps
- Implicit clustering
  - Not based on the similarity of observations themselves (cf. GMM)
  - But based on similarity of observations in state transition
- Support unsupervised training. E.g.,
  - States={rainy, snowy, sunny}
  - Observations={wet, icy, dry}

Pros & Cons of HMM



Cons

- The discriminative classification is over-simplified (can be addressed by reverse $s \rightarrow x$ to $x \rightarrow s$ and incorporate more features)
- Label bias problem



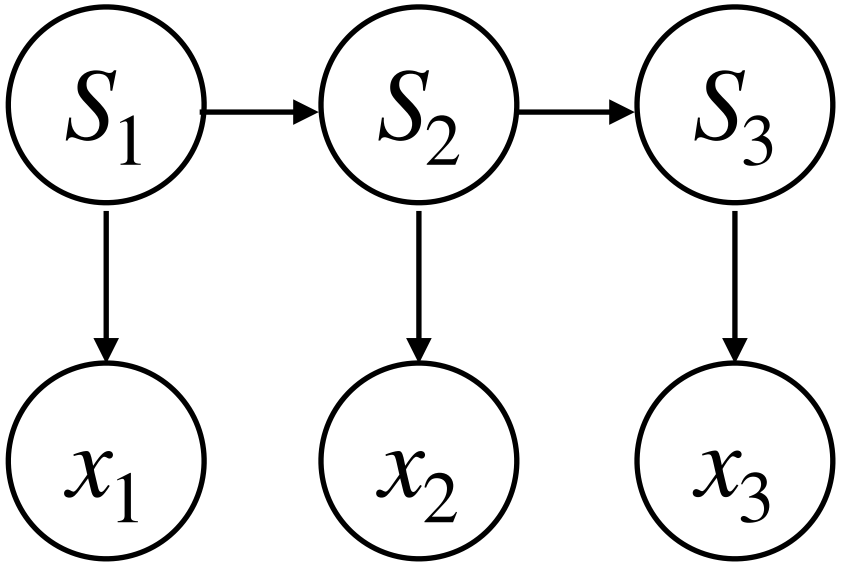
What the local transition probabilities say:

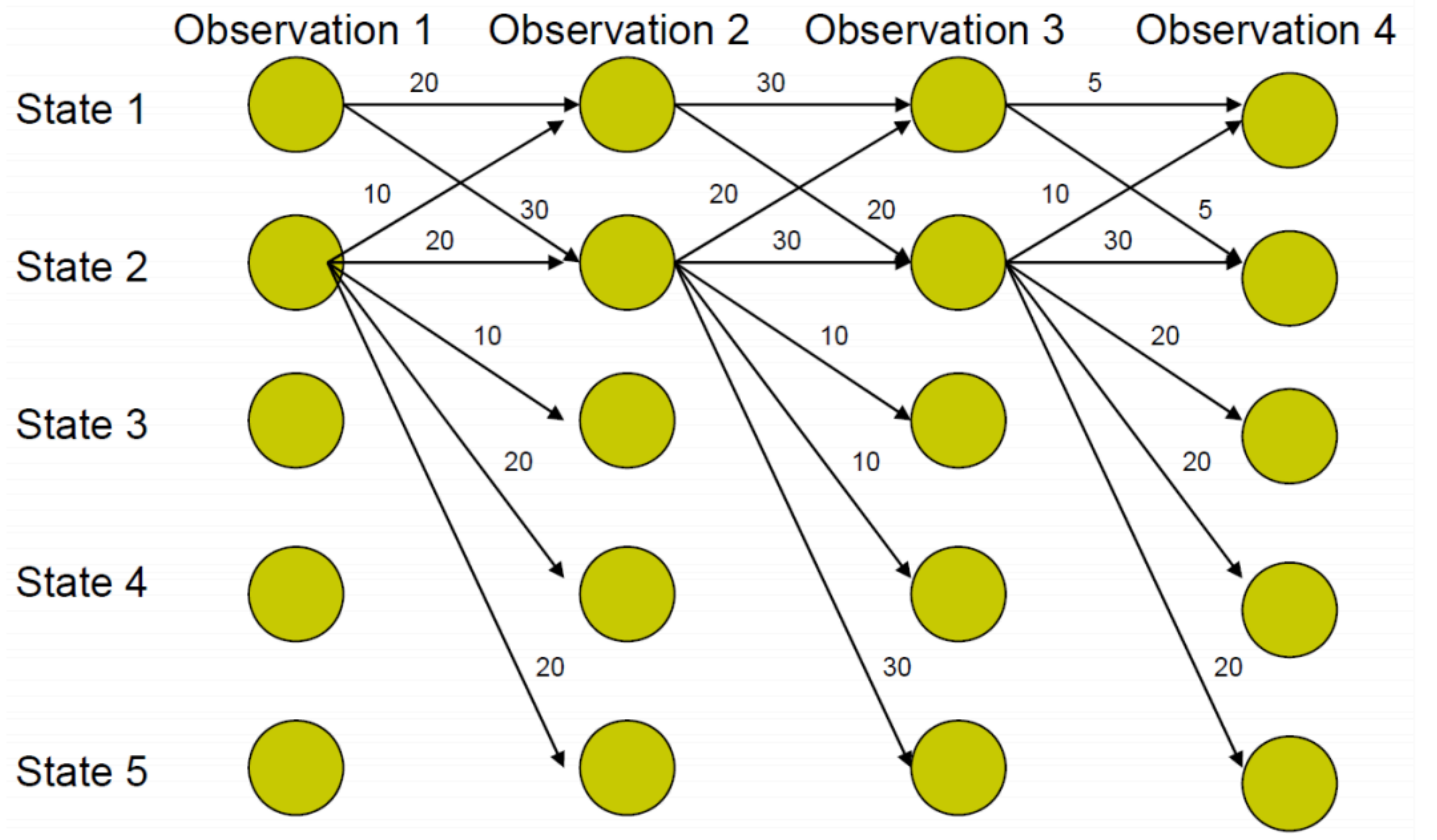
- State 1 almost always prefers to go to state 2
- State 2 almost always prefer to stay in state 2

Source: http://www.cs.cmu.edu/~epxing/Class/10708/lectures/lecture12-CRF.pdf

Undirected Graph

- Idea: Each local factor yields a scoring function, instead of a probability
- Normalizing the probability afterwards





Source: http://www.cs.cmu.edu/~epxing/Class/10708/lectures/lecture12-CRF.pdf

Markov Random Field

- Let $V = \{X_1, X_2, \cdots, X_N\}$ be the nodes
- The **scope** of a factor ϕ_i is a subset of V:

  $\{X_{i,1}, \cdots, X_{i,n_i}\}$, where $X_{i,j} \in V$
- A **factor** maps the values of a scope to a non-negative/ positive number

(Also model parameters)

$$\phi_i : X_{i,1}, X_{i,2}, \cdots, X_{i,n_i} \to \mathbb{R}^+$$

Suppose we have K factors in total

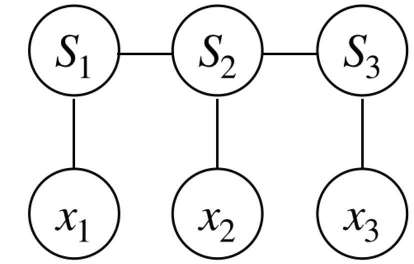
Def (unnormalized measure): $\widetilde{p}(x_1, \cdots, x_n) = \prod_{k=1}^{K} \phi_k(x_{k,1}, \cdots, x_{k,n_k})$

Def (partition function): $Z = \sum_{x_1, \cdots, x_n} p(x_1, \cdots, x_n)$

Def (Probability): $p(x_1, \cdots, x_n) = \frac{1}{Z} \widetilde{p}(x_1, \cdots, x_n)$

UNIVERSITY OF ALBERTA

Markov Network



- Let $V = \{X_1, X_2, \cdots, X_N\}$ be the nodes
- The **scope** of a factor ϕ_i is a subset of V:

  $\{X_{i,1}, \cdots, X_{i,n_i}\}$, where $X_{i,j} \in V$

- A **factor** maps the values of a scope to a non-negative/positive number

  $$\phi_i : X_{i,1}, X_{i,2}, \cdots, X_{i,n_i} \to \mathbb{R}^+$$

- A **Markov network** (induced by the MRF) is an undirected graph $G = \langle V, E \rangle$, where

  $$E = \{(i,j) : \exists k, \{x_i, x_j\} \subseteq scope(\phi_k)$$

Markov Random Field

Interpretation of the factors

- Local happiness for a certain assignment
- Not probability: $p(x_1, x_2) \neq \dfrac{\phi(x_1, x_2)}{\sum_{x_1, x_2} \phi(x_1, x_2)}$
- Not marginal probability: $p(x_1, x_2) \not\propto \sum_{x_2} \phi(x_1, x_2)$
- Posterior is local [**HW1**]

$$p(x_i \,|\, \boldsymbol{x}_{-i}) \propto \prod_{k: x_i \in scope(\phi_k)} \phi_k$$

Hint: $p(x_i \,|\, \boldsymbol{x}_{-i}) = \dfrac{p(x_i, \boldsymbol{x}_{-i})}{\sum_{x_i'} p(x_i', \boldsymbol{x}_{-i})}$*, where* $p(\,\cdot\,)$ *is a multiplication of many factors, which in turn can be grouped into two categories: those including* X_i *and those not including* X_i*. The latter is canceled out in both the numerator and the denominator.*

Application of MRF

- No explicit "cause and effect"
  - Entangled photons
  - Image pixels
  - Even in a sentence, a preceding word may not be a cause
  - Social network: everyone is influencing everyone else simultaneously
  - **HW2:** Give your own example. What else is more suitable to be modeled as an MRF than a BN? And why?

Log-Linear Model

- Another parametrization of the MRF

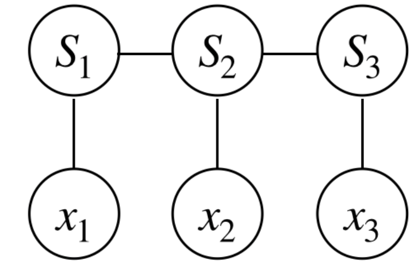
$$
\begin{aligned}
p(x_1, \cdots, x_n) &\propto \prod_{i=1}^{n} \phi(x_{i,1}, \cdots, x_{i,n_i}) \\
&= \exp\left\{ \sum_{i=1}^{n} \log \phi_i(x_{i,1}, \cdots, x_{i,n_i}) \right\} \\
&= \exp\left\{ \sum_{i=1}^{n} \sum_{x'_{i,1}, \cdots, x'_{i,n_i}} \underbrace{\log \phi_i(x'_{i,1}, \cdots, x'_{i,n_i})}_{\theta} \underbrace{\mathbb{I}\{x_{n_i}, \cdots x_{n_i} = x'_{n_i}, \cdots, x'_{n_i}\})}_{f} \right\}
\end{aligned}
$$

Parameters **Features**

$$
p(x_1, \cdots, x_n) = \frac{1}{Z} \exp\{ \sum_i \theta_i f_i(\boldsymbol{x}) \}
$$

The same as MN (suppose potentials >0)

Learning



- Unlike BN, MRF's weights can never be manually assigned
  - Humans are especially bad at expressing our vague intuition
- MRF's weights have to be learned in some principled way

Maximum likelihood estimation (MLE): $\frac{1}{N}\sum_j \log p(\boldsymbol{x}^{(j)}) = \frac{1}{N}\sum_j \log \frac{1}{Z}\exp\{\sum_i \theta_i f_i(\boldsymbol{x}^{(j)})\}$

$$\frac{\partial}{\partial \theta_i}\frac{1}{N}\sum_j \log \frac{1}{Z}\exp\{\sum_i \theta_i f_i(\boldsymbol{x}^{(j)})\}$$

$$= \frac{\partial}{\partial \theta_i}\frac{1}{N}\sum_j \log \exp\{\sum_{i'} \theta_{i'} f_i(x^{(j)})\} - \frac{\partial}{\partial \theta_i}\frac{1}{N}\sum_j \log \sum_{x'} \exp\{\sum_{i'} \theta_{i'} f_{i'}(x')\}$$

$$= \frac{1}{N}\sum_j f_i - \frac{1}{\sum_{x''}\exp\{\sum_i \theta_i f_i(x'')\}}\sum_{x'} \exp\{\sum_i \theta_i f_i(x')\} f_i(x')$$

$$= \frac{1}{N}\sum_j f_i - \sum_{x'} \frac{\exp\{\sum_i \theta_i f_i(x')\}}{\sum_{x''}\exp\{\sum_i \theta_i f_i(x'')\}} f_i(x')$$

$$= \mathbb{E}_{x\sim\mathscr{D}}[f_i] - \mathbb{E}_{x\sim p_{\boldsymbol{\theta}}(x)}[f_i(x)]$$

Expectation in data — Expectation in model

Conditional Random Fields

- Suppose the variables of a data sample can be separated into two parts:
  - The variables $\boldsymbol{x}$ are always given
  - The variables $\boldsymbol{y}$ are of particular interest

Suppose we have K factors in total. For a data sample

Def (unnormalized measure): $\widetilde{p}(\boldsymbol{x}, \boldsymbol{y}) = \prod_{k=1}^{K} \phi_k(\boldsymbol{x}, \boldsymbol{y})$

Def (partition function): $Z_{\boldsymbol{x}} = \sum_{\boldsymbol{y}} p(\boldsymbol{x}, \boldsymbol{y})$

Def (Probability): $p(\boldsymbol{y} \,|\, \boldsymbol{x}) = \frac{1}{Z_{\boldsymbol{x}}} \widetilde{p}(\boldsymbol{x}, \boldsymbol{y})$

HW: Proof that a CRF defined as such is equivalent to the conditional probability as defined in MRF.

Conditional Random Fields

- Suppose the variables of a data sample can be separated into two parts:
  - The variables $\boldsymbol{x}$ are always given
  - The variables $\boldsymbol{y}$ are of particular interest
  - MLE: maximizing $p(\boldsymbol{y}\,|\,\boldsymbol{x})$

MRF:

$$\frac{\partial}{\partial\theta_i}\log p(\boldsymbol{x}^{(i)})$$

$$= \mathbb{E}_{\boldsymbol{x}\sim\mathscr{D}}[f_i] - \mathbb{E}_{\boldsymbol{x}\sim p_{\boldsymbol{\theta}}(x)}[f_i(\boldsymbol{x})]$$

CRF:

Given each data sample $\boldsymbol{x}^{(i)}$

$$\frac{\partial}{\partial\theta_i}\log p(\boldsymbol{y}^{(i)})$$

$$= \mathbb{E}_{\boldsymbol{y}^{(i)}}[f_i] - \mathbb{E}_{\boldsymbol{y}\sim p_{\boldsymbol{\theta}}(\boldsymbol{y})}[f_i(\boldsymbol{y})]$$

This sample $\quad \boldsymbol{y}\sim p_{\boldsymbol{\theta}}(\boldsymbol{y}\,|\,\boldsymbol{x})$

Expectation in data — Expectation in model
(in CRF, given evidence of a particular data point)

Inference

- In general: Hard
- Chain MRF/CRF: DP as for HMM

Suggested Reading

- PGM course

  https://www.youtube.com/watch?v=q8vNcVmarcI&feature=youtu.be

- Chap 9, Bishop, *Pattern Recognition and Machine Learning*.

- Rabiner, L.R., 1989. A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE, 77(2), pp.257-286.

- Lafferty J, McCallum A, Pereira FC. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Thank you!

Q&A