STA 414/2104:
Statistical Methods in Machine Learning II

Week 5: Markov Chain Monte Carlo (MCMC)
Tutorial: Appendix from the lecture

Piotr Zwiernik

University of Toronto

Maximum likelihood estimation for Markov chains

- We use MLE to estimate A from data $\mathcal{D} = \{x^{(1)}, ..., x^{(N)}\}$.
- Likelihood of any particular sentence $x^{(i)}$ of length T_i

$$p(x^{(i)}|\theta) = \prod_{j=1}^{K} \pi_j^{1[x_1^{(i)}=j]} \prod_{t=2}^{T_i} \prod_{j=1}^{K} \prod_{k=1}^{K} A_{jk}^{1[x_t^{(i)}=k, x_{t-1}^{(i)}=j]}$$

Maximum likelihood estimation for Markov chains

- We use MLE to estimate A from data $\mathcal{D} = \{x^{(1)}, ..., x^{(N)}\}$.
- Likelihood of any particular sentence $x^{(i)}$ of length T_i

$$p(x^{(i)}|\theta) = \prod_{j=1}^{K} \pi_j^{1[x_1^{(i)}=j]} \prod_{t=2}^{T_i} \prod_{j=1}^{K} \prod_{k=1}^{K} A_{jk}^{1[x_t^{(i)}=k, x_{t-1}^{(i)}=j]}$$

- Log-likelihood of $\mathcal{D}$ (all sentences treated as independent)

$$\log p(\mathcal{D}|\theta) = \sum_{i=1}^{N} \log p(x^{(i)}|\theta) = \sum_{j} N_j^1 \log \pi_j + \sum_{j} \sum_{k} N_{jk} \log A_{jk}$$

where we define the counts

$$N_j^1 = \sum_{i=1}^{N} 1[x_1^{(i)} = j], \qquad N_{jk} = \sum_{i=1}^{N} \sum_{t=1}^{T_i - 1} 1[x_t^{(i)} = j, x_{t+1}^{(i)} = k].$$

Maximum likelihood estimation for Markov chains

- We use MLE to estimate A from data $\mathcal{D} = \{x^{(1)}, ..., x^{(N)}\}$.
- Likelihood of any particular sentence $x^{(i)}$ of length T_i

$$p(x^{(i)}|\theta) = \prod_{j=1}^{K} \pi_j^{1[x_1^{(i)}=j]} \prod_{t=2}^{T_i} \prod_{j=1}^{K} \prod_{k=1}^{K} A_{jk}^{1[x_t^{(i)}=k, x_{t-1}^{(i)}=j]}$$

- Log-likelihood of $\mathcal{D}$ (all sentences treated as independent)

$$\log p(\mathcal{D}|\theta) = \sum_{i=1}^{N} \log p(x^{(i)}|\theta) = \sum_{j} N_j^1 \log \pi_j + \sum_{j} \sum_{k} N_{jk} \log A_{jk}$$

where we define the counts

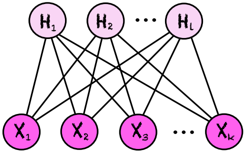
$$N_j^1 = \sum_{i=1}^{N} 1[x_1^{(i)} = j], \qquad N_{jk} = \sum_{i=1}^{N} \sum_{t=1}^{T_i - 1} 1[x_t^{(i)} = j, x_{t+1}^{(i)} = k].$$

- The MLE is given as $\hat{\pi}_j = \frac{N_j^1}{\sum_j N_j^1}$ $\quad \hat{A}_{jk} = \frac{N_{jk}}{\sum_k N_{jk}}$.

Gibbs example 2: Restricted Boltzmann machine

Model for $(X_1, \ldots, X_k, H_1, \ldots, H_l) \in \{-1, 1\}^{k+l}$ (c.f. **Tutorial 3**)

$$p(x_1, \ldots, x_k, h_1, \ldots, h_l) \propto \exp\{\sum_i \alpha_i x_i + \sum_i \beta_i h_i + \sum_{i=1}^{k} \sum_{j=1}^{l} J_{ij} x_i h_j\}.$$



We can easily generate new samples from the learned distribution.

$$p(x|h) = \prod_{i=1}^{l} p(x_i|h), \quad p(h|x) = \prod_{j=1}^{k} p(h_j|h)$$

$$p(x_i = 1|h) = \frac{\prod_{j=1}^{l} \psi_{ij}(x_i, h_j)}{\prod_{j=1}^{l} \psi_{ij}(-1, h_j) + \prod_{j=1}^{l} \psi_{ij}(1, h_j)} = \sigma(2(\alpha_i + \sum_{j=1}^{l} J_{ij} h_j))$$

$$p(h_j = 1|x) = \frac{\prod_{i=1}^{k} \psi_{ij}(x_i, h_j)}{\prod_{i=1}^{k} \psi_{ij}(x_i, -1) + \prod_{i=1}^{k} \psi_{ij}(x_i, 1)} = \sigma(2(\beta_j + \sum_{i=1}^{k} J_{ij} x_i))$$

with $\sigma(y) = 1/(1 + e^{-y})$ called the **sigmoid function**.