Tutorial 3

CSC412: Probabilistic Machine Learning

Stephan Rabanser stephan@cs.toronto.edu

January 20, 2025

Topics of This Tutorial

- Hammersley-Clifford Theorem
- Gaussian Log-Likelihood
- Markov Random Fields as Exponential Families
- Variable Elimination
- Restricted Boltzmann Machines

The Intuition for How the Hammersley-Clifford Theorem Works

Consider a simple chain $X - Y - Z$. The corresponding graphical model is given by all distributions that factorize:

$$f(x, y, z) = \alpha(x, y)\beta(y, z).$$

We want to show that this is equivalent to $X \perp Z \mid Y$ as long as $\alpha(x, y) > 0$ and $\beta(y, z) > 0$ for all x, y, z.

$$f(x, y, z) = \alpha(x, y)\beta(y, z) \qquad \iff \qquad X \perp Z \mid Y$$

We will use the characterization that $X \perp Z \mid Y$ if and only if $f(x \mid y, z) = f(x \mid y)$.
$\Longleftarrow$: For the left implication note that

$$f(x, y, z) = f(y, z)f(x \mid y, z) = f(x \mid y)f(y, z).$$

So the statement works with $\alpha(x, y) = f(x \mid y)$ and $\beta(y, z) = f(y, z)$.

The Intuition for How the Hammersley-Clifford Theorem Works

$\Longrightarrow$: For the right implication note that

$$f(y, z) = \sum_x \alpha(x, y)\beta(y, z) = \left(\sum_x \alpha(x, y)\right) \beta(y, z),$$

and so

$$f(x \mid y, z) = \frac{f(x, y, z)}{f(y, z)} = \frac{\alpha(x, y)\beta(y, z)}{\left(\sum_x \alpha(x, y)\right) \beta(y, z)} = \frac{\alpha(x, y)}{\sum_x \alpha(x, y)},$$

which does not depend on z, proving the conditional independence.

Gaussian Log-Likelihood

Suppose we observe some data from the m-variate Gaussian distribution $\mathbf{x}_{1:n} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$. For this calculation, we will assume that the underlying mean is 0. This is something that can be assumed without loss of generality by centering the data. Denote $K = \Sigma^{-1}$. Then the corresponding density function is expressed as follows:

$$f(\mathbf{x}; K) = \frac{\sqrt{\det(K)}}{(2\pi)^{m/2}} \exp\left(-\tfrac{1}{2}\,\mathbf{x}^\top K\,\mathbf{x}\right).$$

The log density for a single data point is given by

$$\log f(\mathbf{x}; K) = -\frac{m}{2}\log(2\pi) + \frac{1}{2}\log\det K - \frac{1}{2}\mathbf{x}^\top K\mathbf{x}.$$

Up to the obvious constants that do not depend on K, the log-likelihood is

$$\ell_n(K; \mathbf{x}_{1:n}) = \frac{n}{2}\log\det(K) - \frac{1}{2}\sum_{i=1}^{n}\mathbf{x}_i^\top K\mathbf{x}_i.$$

Gaussian Log-Likelihood

Note that:

$$\sum_{i=1}^{n} \mathbf{x}_i^\top K \mathbf{x}_i = \sum_{i=1}^{n} \operatorname{tr}(\mathbf{x}_i^\top K \mathbf{x}_i) = \sum_{i=1}^{n} \operatorname{tr}(K \mathbf{x}_i \mathbf{x}_i^\top) = \operatorname{tr}(K n \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\top) = n \operatorname{tr}(K S_n),$$

where the empirical covariance is given by:

$$S_n = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\top.$$

With this new notation:

$$\ell_n(K; \mathbf{x}_{1:n}) = \frac{n}{2}(\log \det(K) - \operatorname{tr}(K S_n)).$$

Some useful facts:

- $\log \det(K)$ is a strictly concave function of K.
- $\operatorname{tr}(K S_n)$ is linear in K. Hence, $\operatorname{tr}(K S_n)$ is also strictly concave.
- The gradients are $\nabla_K \log \det(K) = K^{-1}$ and $\nabla_K \operatorname{tr}(K S_n) = S_n$.

Gaussian Log-Likelihood

Recall:

- The sum of a strictly concave function with another strictly concave function is strictly concave.
- Any first order maximizer is also a global maximizer in concave functions.
- We want to find the optimal estimator for K.

Using these insights we get:

$$\begin{aligned}\nabla_K \ell_n(K; \mathbf{x}_{1:n}) &= \frac{n}{2}(\nabla_K \log\det(K) - \nabla_K \operatorname{tr}(KS_n)) \stackrel{!}{=} 0 \\ &= \frac{n}{2}(K^{-1} - S_n) \stackrel{!}{=} 0\end{aligned}$$

This becomes 0 when $K^{-1} = S_n$, i.e. when we use the empirical covariance estimator.

Markov Random Fields as Exponential Families

Consider a simple undirected graph $X_1 - X_2 - X_3$ where each variable is binary. Consider the following graphical model:

$$p(x_1, x_2, x_3 \mid \theta) = \frac{1}{Z(\theta)} \psi_{1,2}(x_1, x_2 \mid \theta_{1,2}) \psi_{2,3}(x_2, x_3 \mid \theta_{2,3}),$$

or equivalently:

$$p(x_1, x_2, x_3 \mid \theta) = \exp\left\{\log \psi_{1,2}(x_1, x_2 \mid \theta_{1,2}) + \log \psi_{2,3}(x_2, x_3 \mid \theta_{2,3}) - \log Z(\theta)\right\}.$$

We wont worry about the normalization factor from here on onwards, i.e. $Z(\theta) = 1$.

Markov Random Fields as Exponential Families

The vector (x_1, x_2) takes four values $(0,0), (0,1), (1,0), (1,1)$. Take:

$$\theta_{1,2} := \begin{bmatrix} \log \psi_{1,2}(0,0) \\ \log \psi_{1,2}(0,1) \\ \log \psi_{1,2}(1,0) \\ \log \psi_{1,2}(1,1) \end{bmatrix} \in \mathbb{R}^4,$$

and let $\phi_{1,2}(x_1, x_2)$ be the function that satisfies:

$$\phi_{1,2}(0,0) = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad \phi_{1,2}(0,1) = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \quad \phi_{1,2}(1,0) = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}, \quad \phi_{1,2}(1,1) = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}.$$

Markov Random Fields as Exponential Families

With these definitions, $\log \psi_{1,2}(x_1, x_2 \mid \theta_{1,2}) = \theta_{1,2}^\top \phi_{1,2}(x_1, x_2)$. We define $\theta_{2,3}$ and $\phi_{2,3}(x_2, x_3)$ in a similar way, obtaining that:

$$p(x_1, x_2, x_3 \mid \theta) = \exp\left\{\theta_{1,2}^\top \phi_{1,2}(x_1, x_2) + \theta_{2,3}^\top \phi_{2,3}(x_2, x_3) - \log Z(\theta)\right\},$$

which forms an exponential family with sufficient statistics:

$$\phi_{1,2}(x_1, x_2) = \begin{bmatrix} (1 - x_1)(1 - x_2) \\ (1 - x_1)x_2 \\ x_1(1 - x_2) \\ x_1 x_2 \end{bmatrix}, \quad \phi_{2,3}(x_2, x_3) = \begin{bmatrix} (1 - x_2)(1 - x_3) \\ (1 - x_2)x_3 \\ x_2(1 - x_3) \\ x_2 x_3 \end{bmatrix},$$

and with $Z(\theta) = 1$.

Markov Random Fields as Exponential Families

As a side comment, we note that this exponential family is not minimal in the sense that the values of $\phi_{1,2}(x_1, x_2)$ and $\phi_{2,3}(x_2, x_3)$ lie in a hyperplane:

$$\phi_{1,2}(x_1, x_2)^\top \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} = 1 \quad \text{for all } (x_1, x_2) \in \{0, 1\}^2.$$

Non-minimal exponential families do not satisfy the gradient equation $\nabla A(\theta) = \mathbb{E}_\theta[T(X)]$ – indeed, here $A(\theta) = 0$. An easy solution is to get rid of the first coordinate in $\phi_{1,2}(x_1, x_2)$ and replace it with the corresponding functions of the remaining entries of $\phi_{1,2}(x_1, x_2)$. This defines new natural parameters:

$$\bar{\theta}_{1,2} = \begin{bmatrix} \log \psi_{1,2}(0,1) - \log \psi_{1,2}(0,0) \\ \log \psi_{1,2}(1,0) - \log \psi_{1,2}(0,0) \\ \log \psi_{1,2}(1,1) - \log \psi_{1,2}(0,0) \end{bmatrix}, \quad \bar{\theta}_{2,3} = \begin{bmatrix} \log \psi_{2,3}(0,1) - \log \psi_{2,3}(0,0) \\ \log \psi_{2,3}(1,0) - \log \psi_{2,3}(0,0) \\ \log \psi_{2,3}(1,1) - \log \psi_{2,3}(0,0) \end{bmatrix}.$$

Markov Random Fields as Exponential Families

And new sufficient statistics:

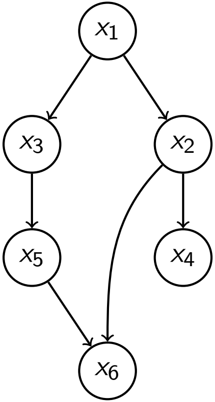
$$\bar{\phi}_{1,2}(x_1, x_2) = \begin{bmatrix} (1 - x_1)x_2 \\ x_1(1 - x_2) \\ x_1 x_2 \end{bmatrix}, \quad \bar{\phi}_{2,3}(x_2, x_3) = \begin{bmatrix} (1 - x_2)x_3 \\ x_2(1 - x_3) \\ x_2 x_3 \end{bmatrix}.$$

Moreover:

$$A(\bar{\theta}) = \log \psi_{1,2}(0, 0)\psi_{2,3}(0, 0),$$

which should now be explicitly expressed in terms of $\bar{\theta}_{1,2}$ and $\bar{\theta}_{2,3}$.

Simple variable elimination example



Suppose that we observe the variable $X_6 = \bar{x}_6$. What is $p(X_1 \mid \bar{x}_6)$? The corresponding DAG model implies the factorization:

$$p(x_1, \ldots, x_6) = p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_1)p(x_4 \mid x_2)p(x_5 \mid x_3)p(x_6 \mid x_2, x_5).$$

We have:

$$x_F = \{x_1\}, \quad x_E = \{x_6\}, \quad x_R = \{x_2, x_3, x_4, x_5\},$$

$$p(x_F \mid x_E) = \frac{p(x_F, x_E)}{p(x_E)} = \frac{\sum_{x_R} p(x_F, x_E, x_R)}{\sum_{x_F, x_R} p(x_F, x_E, x_R)},$$

$$\implies p(x_1 \mid \bar{x}_6) = \frac{p(x_1, \bar{x}_6)}{p(\bar{x}_6)} = \frac{p(x_1, \bar{x}_6)}{\sum_{x \in x_F, x_R} p(x, \bar{x}_6)}.$$

Simple Variable Elimination Example

To compute $p(x_1, \bar{x}_6)$, we use variable elimination in the order $2, 3, 4, 5$:

$$
\begin{aligned}
p(x_1, \bar{x}_6) &= p(x_1) \sum_{x_2} \sum_{x_3} \sum_{x_4} \sum_{x_5} p(x_2 \mid x_1) p(x_3 \mid x_1) p(x_4 \mid x_2) p(x_5 \mid x_3) p(\bar{x}_6 \mid x_2, x_5) \\
&= p(x_1) \sum_{x_2} p(x_2 \mid x_1) \sum_{x_3} p(x_3 \mid x_1) \sum_{x_4} p(x_4 \mid x_2) \sum_{x_5} p(x_5 \mid x_3) p(\bar{x}_6 \mid x_2, x_5) \\
&= p(x_1) \sum_{x_2} p(x_2 \mid x_1) \sum_{x_3} p(x_3 \mid x_1) \sum_{x_4} p(x_4 \mid x_2) p(\bar{x}_6 \mid x_2, x_3)
\end{aligned}
$$

Simple Variable Elimination Example

Note that $p(\bar{x}_6 \mid x_2, x_3)$ does not need to participate in $\sum_{x_4}$, so:

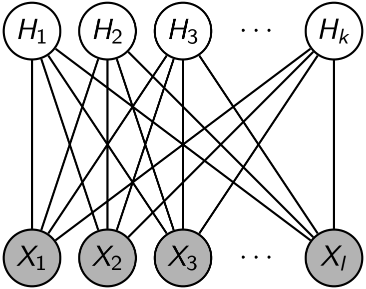
$$
\begin{aligned}
p(x_1, \bar{x}_6) &= p(x_1) \sum_{x_2} p(x_2 \mid x_1) \sum_{x_3} p(x_3 \mid x_1) p(\bar{x}_6 \mid x_2, x_3) \sum_{x_4} p(x_4 \mid x_2) \\
&= p(x_1) \sum_{x_2} p(x_2 \mid x_1) \sum_{x_3} p(x_3 \mid x_1) p(\bar{x}_6 \mid x_2, x_3) \\
&= p(x_1) \sum_{x_2} p(x_2 \mid x_1) p(\bar{x}_6 \mid x_1, x_2) \\
&= p(x_1) p(\bar{x}_6 \mid x_1)
\end{aligned}
$$

Finally:

$$
p(x_1 \mid \bar{x}_6) = \frac{p(x_1) p(\bar{x}_6 \mid x_1)}{\sum_{x_1} p(x_1) p(\bar{x}_6 \mid x_1)}.
$$

Restricted Boltzmann Machines

A restricted Boltzmann machine (RBM) is a simple generative stochastic artificial neural network model. In the language of today's lecture, it is obtained from a special form of the Ising model with variables $(X_1, \ldots, X_k, H_1, \ldots, H_l) \in \{-1, 1\}^{k+l}$. The underlying graph is the bipartite graph with all pairs $H_i - X_j$ connected but with no other edges.



Restricted Boltzmann Machines

The Ising model is then given by all distributions:

$$p(x,h) = p(x_1,\ldots,x_k,h_1,\ldots,h_l) \propto \exp\left\{\sum_i \alpha_i x_i + \sum_j \beta_j h_j + \sum_{i=1}^{k}\sum_{j=1}^{l} J_{ij} x_i h_j\right\}.$$

We can write it in terms of factors:

$$\psi_{X_i,H_j}(x_i,h_j) = \exp\left\{\frac{1}{l}\alpha_i x_i + \frac{1}{k}\beta_j h_j + J_{ij} x_i h_j\right\},$$

so that:

$$p(x,h) = p(x_1,\ldots,x_k,h_1,\ldots,h_l) = \frac{1}{Z}\prod_{i=1}^{k}\prod_{j=1}^{l}\psi_{X_i,H_j}(x_i,h_j).$$

Restricted Boltzmann Machines

Note that computing Z may be computationally expensive, but we will see that many quantities can be efficiently computed without knowing Z.

The corresponding RBM is given as the marginal distribution:

$$p(x) = \sum_{h \in \{-1,1\}^l} p(x, h).$$

Note that both:

$$\sum_{h \in \{-1,1\}^l} \prod_{i=1}^{k} \prod_{j=1}^{l} \psi_{X_i, H_j}(x_i, h_j) \quad \text{and} \quad \sum_{x \in \{-1,1\}^k} \prod_{i=1}^{k} \prod_{j=1}^{l} \psi_{X_i, H_j}(x_i, h_j)$$

can be computed very efficiently. This shows that both $p(x \mid h)$ and $p(h \mid x)$ are easy to obtain, and this computation does not even require any knowledge of the normalizing constant Z.

Restricted Boltzmann Machines

This computation confirms what we know from the Hammersley-Clifford theorem: that all H_i's are mutually independent given the vector X. The individual activation functions are given by:

$$p(h_j \mid x) = \frac{\prod_{i=1}^{k} \psi_{ij}(x_i, h_j)}{\prod_{i=1}^{k} \psi_{ij}(x_i, -1) + \prod_{i=1}^{k} \psi_{ij}(x_i, 1)} = \sigma\left(\beta_j + \sum_i J_{ij} x_i\right),$$

with:

$$\sigma(y) = \frac{e^y}{e^{-y} + e^y} = \frac{1}{1 + e^{-2y}},$$

called the sigmoid function.