Lecture 3: Learning Parameters in Graphical Models.

There are two aspects in the subject of learning graphical models: parameter learning and structure (of graph) learning. We concentrate on parameter learning, (from samples) which is easier. In this lecture we suppose that all variables in the sample are "observed" or "visible". In later we will treat the case where some of the variables are "not accessible" or "hidden".

I. The Kullback-Leibler divergence.

Let $p(x)$ and $q(x)$ two probability distributions on a discrete alphabet $x = (x_1, ..., x_k) \in \mathbb{A}^k$ when $x_i \in \mathbb{A}$. By definition:

$$KL(p \parallel q) = \sum_x p(x) \log p(x) - \sum_x p(x) \log q(x)$$

$$= \sum_x p(x) \log \left( \frac{p(x)}{q(x)} \right)$$
We also use the notation:

\[ KL(p \parallel q) = \mathbb{E}_p \left[ \log \frac{p(x)}{q(x)} \right] - \mathbb{E}_q \left[ \log q(x) \right] \]

or sometimes \( \langle \log \frac{p(x)}{q(x)} \rangle_p \).

**Main properties:**

\[ KL(p \parallel q) \geq 0 \quad \text{&} \quad = 0 \iff p(x) = q(x) \quad \forall x. \]

\[ KL(p \parallel q) \neq KL(q \parallel p) \quad \text{Not symmetric} \]

**Proof of positivity:**

\[ \log x \leq x - 1 \quad \text{for} \quad x > 0 \]

\[ \Rightarrow \log \frac{q(x)}{p(x)} \leq q(x) - p(x) \]

\[ \Rightarrow 1 - \frac{q(x)}{p(x)} \leq - \log \frac{q(x)}{p(x)} = \log \frac{p(x)}{q(x)} \]

\[ \Rightarrow p(x) - q(x) \leq p(x) \log \frac{p(x)}{q(x)} \]

\[ \Rightarrow \sum_x p(x) - \sum_x q(x) \leq \sum_x p(x) \log \frac{p(x)}{q(x)} \]

\[ \leq \sum_x \frac{p(x)}{1} \log \frac{p(x)}{q(x)} = KL(p \parallel q) \]
Alternative proof by Jensen's inequality.

\[
KL(p \| q) = \sum_x p(x) \log \frac{p(x)}{q(x)} = - \sum_x p(x) \log \frac{q(x)}{p(x)}.
\]

Note that \((-\log z)\) is a convex function of \(z > 0\):

![Convex function graph]

By convexity, we have that the mean of \((-\log z)\) is bigger than the \((-\log)\) of the mean (see picture)

\[
\mathbb{E}_{\frac{1}{2}}(-\log z) > -\log \mathbb{E}_{\frac{1}{2}}(z).
\]

i.e.

\[
\sum_x p(x) \left( - \log \frac{q(x)}{p(x)} \right) > - \log \sum_x p(x) \frac{q(x)}{p(x)}
\]

\[
= - \log \sum_x q(x)
\]

\[
= - \log 1 = 0.
\]

We have thus found \(KL(p \| q) > 0\).
Given a set of samples $x^{(1)}, \ldots, x^{(N)}$ from distribution $p(x | \Theta)$ where $\Theta$ denotes the set of parameters of $p$ (say weights, bias in a Boltzmann machine ...). The log-likelihood of the data is by definition:

$$L(\Theta) = \log \text{Prob} (x^{(1)}, \ldots, x^{(N)})$$

$$= \log \left\{ \prod_{m=1}^{N} p(x^{(m)} | \Theta) \right\}$$

under iid assumption for data sample.

\[ \sum_{m=1}^{N} \log p(x^{(m)} | \Theta) \]

**Maximum likelihood principle:**

Set estimate $\hat{\Theta} = \arg \max \ L(\Theta)$.

**KL minimization:**

Set $\hat{\Theta} = \arg \min \ KL (q_{\text{emp}} || p)$

where $q_{\text{emp}} (x) = \frac{1}{N} \sum_{m=1}^{N} \delta_{x, x^{(m)}}$

is the empirical dist of the data.
Claim: \[ \text{ML maximization } \iff \text{KL minimization.} \]

Proof: 
\[
\text{KL}(q_{emp} \parallel p) = \mathbb{E}_{q_{emp}} \left[ \log \frac{q_{emp}(x)}{p(x|\theta)} \right]
\]
\[
= \mathbb{E}_{q_{emp}} \left[ \log q_{emp}(x) \right] - \mathbb{E}_{q_{emp}} \left[ \log p(x|\theta) \right]
\]
\[\text{independent of } \theta \]

Note that 
\[
\mathbb{E}_{q_{emp}} \left[ \log p(x) \right] = \frac{1}{N} \sum_{m=1}^{N} \log p(x(m)|\theta).
\]
\[= \frac{1}{N} \mathcal{L}(\theta).
\]

So,
\[
\text{KL}(q_{emp} \parallel p) = \mathbb{E}_{q_{emp}} \left[ \log q_{emp}(x) \right] - \frac{1}{N} \mathcal{L}(\theta).
\]

\[\implies \max_{\theta} \text{KL is equivalent to } \min_{\theta} \mathcal{L}(\theta).
\]
Recall for a BN we have a prob dist of the form

\[ p(x) = \prod_{i=1}^{k} p(x_i \mid (pa)_i) \]

where \((pa)_i\) are the variables which are parents of \(x_i\).

We assume that each \(p(x_i \mid (pa)_i)\) depends on a set of parameters that we call \(\theta_i \mid (pa)_i\).

We apply the ML principle or equivalently we want to minimize \(KL(\text{qemp} \parallel p)\) where \(q_{emp}\) is the empirical distribution of data sample \(X^{(1)}, \ldots, X^{(N)}\).

**Lemma:** \(KL(q_{emp} \parallel p)\) is minimized for \(p(x_i \mid (pa)_i) = q_{emp}(x_i \mid (pa)_i)\).

In practice we have

\[ q_{emp}(x_i = s \mid (pa)_i = t) = \frac{\sum_{m=1}^{N} \mathbb{1}(x_i^{(m)} = s, (pa)_i^{(m)} = t)}{\sum_{m=1}^{N} \mathbb{1}((pa)_i^{(m)} = t)} \]
which is therefore known. We solve for \( q_{\text{map}} \) from the equation of the Lemma.

Proof of Lemma.

\[
KL(q_{\text{map}} \parallel p) = \mathbb{E}_{q_{\text{map}}} \left( \log q_{\text{map}} \right) - \mathbb{E}_{q_{\text{map}}} \left( \log p(x) \right)
\]

\[
= \mathbb{E}_{q_{\text{map}}} \left( \log q_{\text{map}} \right) - \sum_{i=1}^{K} \mathbb{E}_{q_{\text{map}}} \left( \log p(x_i \mid \text{pa}_i) \right)
\]

because \( p(x_i \mid \text{pa}_i) \) depends only on \( (x_i, \text{pa}_i) \).

Now we add and subtract an appropriate term.

\[
KL(q_{\text{map}} \parallel p) = \mathbb{E}_{q_{\text{map}}} \left( \log q_{\text{map}} \right) - \sum_{i=1}^{K} \mathbb{E}_{q_{\text{map}}} \left( \log q_{\text{map}}(x_i \mid \text{pa}_i) \right)
\]

\[
- \sum_{i=1}^{K} \mathbb{E}_{q_{\text{map}}(x_i \mid \text{pa}_i)} \left[ \log p(x_i \mid \text{pa}_i) \right] + \mathbb{E}_{q_{\text{map}}(x_i \mid \text{pa}_i)} \left[ \log p(x_i \mid \text{pa}_i) \right]
\]
The first two terms are independent of the parameters $\Theta$ and play the role of a constant when we minimize.

For the last two terms we remark that by Bayes Law:

$$\int_{\Theta} \int_{\Omega} p(x_i, \Theta | y) \, d\Theta \, d\Omega = \int_{\Theta} \int_{\Omega} p(x_i, \Theta | y) \, d\Theta \, d\Omega$$

Thus,

$$KL (q_{exp} || p) = \text{constant} +$$

$$\sum_{i=1}^{K} \int_{\Theta} \int_{\Omega} q_{exp} (\Theta | y) \left\{ \int \int \log q_{exp} (x_i | \Theta, \Omega) \, d\Theta \, d\Omega \right\}$$

$$- \int \int \log q_{exp} (x_i | \Theta, \Omega) \, d\Theta \, d\Omega$$

$$= \text{constant} + \sum_{i=1}^{K} \int \int q_{exp} (\Theta | y) \left[ KL (q_{exp} (x_i | \Theta, \Omega) || p(x_i | \Theta, \Omega)) \right]$$

The $KL > 0$ is minimized (vanishes) for a set of parameters $\Theta$ such that

$$p(x_i | \Theta, \Omega) = q_{exp} (x_i | \Theta, \Omega)$$
Here \( p(x, \theta) = \frac{1}{Z(\theta)} \prod_c \psi_c(x_c, \theta_c) \).

For iid samples \( (x^{(1)}, \ldots, x^{(N)}) \) we have

\[
L(\theta) = \sum_{m=1}^{N} \log p(x^{(m)}) = \sum_c \sum_{m=1}^{N} \log \psi_c(x_c^{(m)}, \theta_c) = N \log Z(\theta).
\]

where

\[
Z(\theta) = \sum_{x \in \mathcal{X}} \prod_c \psi_c(x_c, \theta_c).
\]

Now \( \log Z(\theta) \) is intractable and all parameters are coupled. One can use gradient ascent in order to maximize \( L(\theta) \). This involves computing \( \nabla_{\theta} L(\theta) \).

**Computation of \( \nabla_{\theta} L(\theta) \):** For \( \theta_c \) we have

\[
\nabla_{\theta_c} L(\theta) = \sum_{m=1}^{N} \frac{\partial}{\partial \theta_c} \log \psi_c(x_c^{(m)}, \theta_c) = N \nabla_{\theta_c} \log Z(\theta).
\]

Easy and explicit?
\[ \nabla_{\theta_c} \log p(\theta) = \frac{1}{\pi(\theta)} \nabla_{\theta_c} \pi(\theta) \]

\[ = \frac{1}{\pi(\theta)} \sum_x \nabla_{\theta_c} \left\{ \prod_c \psi_c(x_c | \theta_c) \right\} \]

\[ = \frac{1}{\pi(\theta)} \sum_x \nabla_{\theta_c} \psi_c(x_c | \theta_c) \cdot \prod_{c' \neq c} \psi_c(x_c | \theta_c) \]

\[ = \frac{1}{\pi(\theta)} \sum_x \left\{ \frac{\nabla_{\theta_c} \psi_c(x_c | \theta_c)}{\psi_c(x_c | \theta_c)} \right\} \prod_{c} \psi_c(x_c | \theta_c) \]

\[ = \langle \nabla_{\theta_c} \log \psi_c(x_c | \theta_c) \rangle \]

where \( \langle A(x) \rangle = \frac{1}{\pi} \sum_x A(x) \prod_c \psi_c(x_c) \) is the standard notation for Gibbs / MRF energies. \( \langle \rangle \) not that

\[ \langle \nabla_{\theta_c} \log \psi_c(x_c | \theta_c) \rangle = \mathbb{E}_{P(x_c)} \left[ \nabla_{\theta_c} \psi_c(x_c | \theta_c) \right] \]

Marginal of \( P(x | \theta) \) over all variables \( (x, \ldots, x_c) \setminus x_c \).
Summarizing, we have for all classes $c$ or factor model $c$

$$\nabla \ell(c) = \sum_{m=1}^{n} \nabla \log \frac{y_{c}(x_{m}^{(m)}, 1^{c})}{y_{c}(x_{m}^{(m)}, 1^{c})} - N \left\langle \nabla \log y_{c}(x, 1^{c}) \right\rangle$$

---

easy

---

Requiring marginalization, use message passing, or sampling, ...

(difficult in general).

---

Example: Boltzmann Machine or Ising Model

$$p(x) = \frac{1}{Z(W)} e^{-\frac{1}{2} x^{T} W x}$$

where $(W)_{ij}$ is a weight matrix. (set $W_{ii} = 0$).

We have after application of above method (exercise):

$$\Theta_{ij} = \sum_{m=1}^{N} \left( x_{i}^{(m)} x_{j}^{(m)} - \left\langle x_{i} x_{j} \right\rangle \right),$$

where

$$\left\langle x_{i} x_{j} \right\rangle = \sum_{x} x_{i} x_{j} p(x) = \frac{1}{2} \sum_{x} x_{i} x_{j} e^{\frac{1}{2} x^{T} W x}$$

(difficult to compute in general).