PROBABILISTIC GRAPHICAL MODELS.

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 samples are "observed" or "visible". In lect 4 we will treat the case where some of the variables are "not accessible" or "hidden".

I. The Kullback-Leibler divergence.

Let $p(\underline{x})$ and $q(\underline{x})$ two probability distributions over a discrete alphabet $\underline{x} = (x_1 \ldots x_K) \in \mathcal{A}^K$ where $x_i \in \mathcal{A}$. By definition:

$$KL(p \| q) = \sum_{\underline{x}} p(\underline{x}) \log p(\underline{x}) - \sum_{\underline{x}} p(\underline{x}) \log q(\underline{x})$$

$$= \sum_{\underline{x}} p(\underline{x}) \log \left\{ \frac{p(\underline{x})}{q(\underline{x})} \right\}.$$

We also use the notation:

$$KL(p\|q) = \mathbb{E}_p[\log p(\underline{x})] - \mathbb{E}_p[\log q(\underline{x})]$$

$$= \mathbb{E}_p\left[\log\left\{\frac{p(\underline{x})}{q(\underline{x})}\right\}\right]$$

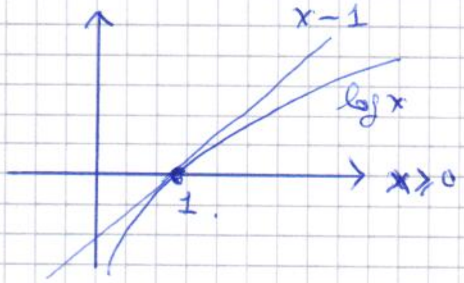
or sometimes $\left\langle \log \frac{p(\underline{x})}{q(\underline{x})} \right\rangle_p$.

<u>Main properties</u>:

$KL(p\|q) \geq 0$ & $= 0$ iff $p(\underline{x}) = q(\underline{x})$ $\forall \underline{x}$.

$KL(p\|q) \neq KL(q\|p)$ Not symmetric.

<u>Proof of positivity</u>:



$\log x \leq x - 1$ for $x \geq 0$

$$\Rightarrow \log \frac{q(\underline{x})}{p(\underline{x})} \leq \frac{q(\underline{x})}{p(\underline{x})} - 1$$

$$\Rightarrow 1 - \frac{q(\underline{x})}{p(\underline{x})} \leq -\log \frac{q(\underline{x})}{p(\underline{x})} = \log \frac{p(\underline{x})}{q(\underline{x})}$$

$$\Rightarrow p(\underline{x}) - q(\underline{x}) \leq p(\underline{x}) \log \frac{p(\underline{x})}{q(\underline{x})}$$

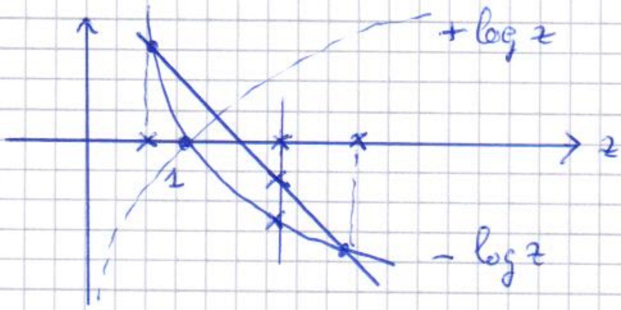
$$\Rightarrow \underbrace{\sum_{\underline{x}} p(\underline{x})}_{1} - \underbrace{\sum_{\underline{x}} q(\underline{x})}_{1} \leq \sum_{\underline{x}} p(\underline{x}) \log \frac{p(\underline{x})}{q(\underline{x})}$$

$$\Rightarrow 0 \leq \sum_{\underline{x}} p(\underline{x}) \log \frac{p(\underline{x})}{q(\underline{x})} = KL(p\|q)$$ ∎

Alternative proof by Jensen's inequality.

$$KL(p\|q) = \sum_{\underline{x}} p(\underline{x}) \log \frac{p(\underline{x})}{q(\underline{x})} = - \sum_{x} p(\underline{x}) \log \frac{q(\underline{x})}{p(\underline{x})}.$$

Note that $(-\log z)$ is a convex fct of $z \geq 0$:



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

$$\mathbb{E}_z(-\log z) \geq -\log \mathbb{E}_z(z).$$

i.e

$$\sum_{x} p(\underline{x}) \left(-\log \frac{q(\underline{x})}{p(\underline{x})}\right) \geq -\log \sum_{x} p(x) \frac{q(\underline{x})}{p(\underline{x})}$$

$$= -\log \sum_{x} q(\underline{x})$$

$$= -\log 1 = 0.$$

We have thus found $KL(p\|q) \geq 0$ ∎

II. Maximum Likelihood Method and KL minimization

Given a set of samples $\underline{x}^{(1)} \dots \underline{x}^{(N)}$ from distribution $p(\underline{x} \mid \vartheta)$ where ϑ denotes the set of parameters of p (say weights, biases in a Boltzmann machine ...) the log-likelihood of the data is by definition:

$$L(\vartheta) = \log \operatorname{Prob}(\underline{x}^{(1)} \dots \underline{x}^{(N)})$$

$$= \log \left\{ \prod_{m=1}^{N} p(\underline{x}^{(m)} \mid \vartheta) \right\}$$

under iid assumption for data samples

$$= \sum_{m=1}^{N} \log p(\underline{x}^{(m)} \mid \vartheta)$$

<u>Maximum likelihood principle</u>:

Set estimate $\hat{\vartheta} = \arg\max_{\vartheta} L(\vartheta)$.

<u>KL minimization</u>:

$$\text{Set } \hat{\vartheta} = \arg\min_{\vartheta} KL(q_{emp} \,\|\, p)$$

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

is the empirical distr of the data.

Claim : ML maximization $\Longleftrightarrow$ KL minimization.

Proof : $$KL(q_{emp} \| p) = \mathbb{E}_{q_{emp}} \log \left\{ \frac{q_{emp}(\underline{x})}{p(\underline{x} \mid \vartheta)} \right\}$$

$$= \underbrace{\mathbb{E}_{q_{emp}} \left(\log q_{emp}(\underline{x}) \right)}_{\text{independent of } \vartheta} - \mathbb{E}_{q_{emp}} \left(\log p(\underline{x} \mid \vartheta) \right)$$

Note that $$\mathbb{E}_{q_{emp}} \left(\log p(\underline{x}) \right) = \frac{1}{N} \sum_{m=1}^{N} \log p(\underline{x}^{(m)} \mid \vartheta)$$

$$= \frac{1}{N} L(\vartheta).$$

So :

$$KL(q_{emp} \| p) = \mathbb{E}_{q_{emp}} \left(\log q_{emp}(\underline{x}) \right) - \frac{1}{N} L(\vartheta).$$

$\Rightarrow$ $\max_{\vartheta} KL$ is equiv to $\min_{\vartheta} L(\vartheta)$. $\blacksquare$

III. ML Training of Belief Networks.

Recall for a BN we have a prob distr of the form

$$p(\underline{x}) = \prod_{i=1}^{K} p(x_i \mid (pa)_i)$$

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 $KL(q_{emp} \| p)$ where q_{emp} is the empirical distribution of data samples $\underline{x}^{(1)} \ldots \underline{x}^{(N)}$.

Lemma: $KL(q_{emp} \| p)$ 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 $\vartheta_{i|pa_i}$ from the equation of the Lemma.

Proof of Lemma.

$$KL(q_{emp} \| p) = \mathbb{E}_{q_{emp}}(\log q_{emp}) - \mathbb{E}_{q_{emp}}(\log p(\mathbf{x}))$$

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

$$= \mathbb{E}_{q_{emp}}(\log q_{emp}) - \sum_{i=1}^{K} \underbrace{\mathbb{E}_{q_{emp}}\left(\log p(x_i \mid (pa)_i)\right)}_{\mathbb{E}_{q_{emp}(x_i, (pa)_i)}\left(\log p(x_i \mid (pa)_i)\right)}$$

because $p(x_i \mid (pa)_i)$ depends only on $(x_i, (pa)_i)$.

Now we add and subtract an appropriate term:

$$KL(q_{emp} \| p) = \mathbb{E}_{q_{emp}}(\log q_{emp}) - \sum_{i=1}^{K} \mathbb{E}_{q_{emp}(x_i, (pa)_i)}\left[\log q_{emp}(x_i \mid pa_i)\right]$$

$$- \left\{ \sum_{i=1}^{K} \mathbb{E}_{q_{emp}(x_i, pa_i)}\left[\log p(x_i \mid pa_i)\right] - \mathbb{E}_{q_{emp}(x_i, pa_i)}\left[\log q_{emp}(x_i \mid pa_i)\right] \right\}$$

The first two terms are independent of the parameters $\vartheta_{i|pa_i}$ and play the role of a constant when we minimize.

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

$$\mathbb{E}_{q_{emp}(x_i,(pa)_i)} = \mathbb{E}_{q_{emp}(pa_i)} \mathbb{E}_{q_{emp}(x_i|pa_i)}$$

Thus

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

$$\sum_{i=1}^{K} \mathbb{E}_{q_{emp}(pa_i)} \left\{ \mathbb{E}_{q_{emp}(x_i|pa_i)} \log q_{emp}(x_i|pa_i) - \mathbb{E}_{q_{emp}(x_i|pa_i)} \log p(x_i|pa_i) \right\}$$

$$= \text{constant} + \sum_{i=1}^{K} \mathbb{E}_{q_{emp}(pa_i)} \left[KL\left(q_{emp}(x_i|pa_i) \| p(x_i|pa_i) \right) \right].$$

The $KL \geq 0$ is minimized (vanishes) for a set of parameters $\vartheta_{i|pa_i}$ such that

$$p(x_i|pa_i) = q_{emp}(x_i|pa_i)$$

IV ML Training of MRF or Factor graph models

Here $p(\underline{x}|\vartheta) = \frac{1}{Z(\vartheta)} \prod_c \psi_c(x_c|\vartheta_c)$.

For iid samples $(\underline{x}^{(1)} \dots \underline{x}^{(N)})$ we have

$$L(\vartheta) = \sum_{m=1}^{N} \log p(\underline{x}^{(m)})$$

$$= \sum_c \sum_{m=1}^{N} \log \psi_c(x_c^{(m)}|\vartheta_c) - N \log Z(\vartheta).$$

where

$$Z(\vartheta) = \sum_{\underline{x} \in \mathcal{A}^K} \prod_c \psi_c(x_c|\vartheta_c).$$

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

Computation of $\nabla_\vartheta L(\vartheta)$: For ϑ_c we have

$$\nabla_{\vartheta_c} L(\vartheta) = \underbrace{\sum_{m=1}^{N} \nabla_{\vartheta_c} \log \psi_c(x_c^{(m)}|\vartheta_c)}_{\text{easy and explicit}} - \underbrace{N \nabla_{\vartheta_c} \log Z(\vartheta)}_{?}$$

$$\nabla_{\vartheta_c} \log Z(\vartheta) = \frac{1}{Z(\vartheta)} \nabla_{\vartheta_c} Z(\vartheta)$$

$$= \frac{1}{Z(\vartheta)} \sum_{\underline{x}} \nabla_{\vartheta_c} \left\{ \prod_{c'} \psi_{c'}(x_{c'} | \vartheta_{c'}) \right\}$$

$$= \frac{1}{Z(\vartheta)} \sum_{\underline{x}} \nabla_{\vartheta_c} \psi_c(x_c | \vartheta_c) \cdot \underbrace{\prod_{c' \neq c} \psi_{c'}(x_{c'} | \vartheta_{c'})}_{\frac{\prod_{c'} \psi_{c'}(x_{c'} | \vartheta_{c'})}{\psi_c(x_c | \vartheta_c)}}$$

$$= \frac{1}{Z(\vartheta)} \sum_{\underline{x}} \left\{ \frac{\nabla_{\vartheta_c} \psi_c(x_c | \vartheta_c)}{\psi_c(x_c | \vartheta_c} \right\} \left\{ \prod_c \psi_c(x_c | \vartheta_c) \right\}.$$

$$= \left\langle \nabla_{\vartheta_c} \log \psi_c(x_c | \vartheta_c) \right\rangle$$

where $\langle A(\underline{x}) \rangle \equiv \frac{1}{Z} \sum_{\underline{x}} A(\underline{x}) \prod_c \psi_c(x_c)$ is the standard notation for Gibbs / MRF averages – Note that

$$\left\langle \nabla_{\vartheta_c} \log \psi_c(x_c | \vartheta_c) \right\rangle = \mathbb{E}_{\underbrace{p(x_c)}} \left[\nabla_{\vartheta_c} \psi_c(x_c | \vartheta_c) \right]$$

Marginal of $p(\underline{x} | \vartheta)$ over all variables $(x_1 \cdots x_K) \setminus x_c$.

Summarizing we have for all cliques C or factor nodes C :

$$\nabla_{\vartheta_c} L(\vartheta) = \underbrace{\sum_{m=1}^{N} \nabla_{\vartheta_c} \log \psi_c(x_c^{(m)} | \vartheta_c)}_{\text{easy}} - N \underbrace{\langle \nabla_{\vartheta_c} \log \psi_c(\underline{x}_c | \vartheta_c) \rangle}_{\text{requires marginalization}}$$

use Message passing,
or sampling, ...
(difficult in general).

<u>Example: Boltzman Machine or Ising Model.</u>

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

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

We have after application of above method (exercise):

$$\frac{\partial L}{\partial W_{ij}} = \sum_{m=1}^{N} \left(x_i^{(m)} x_j^{(m)} - \langle x_i x_j \rangle \right) . \quad \text{where}$$

$$\langle x_i x_j \rangle = \sum_{\underline{x}} x_i x_j \, p(\underline{x}) = \underbrace{\frac{1}{Z} \sum_{\underline{x}} x_i x_j \, e^{+\frac{1}{2} \underline{x}^T W \underline{x}}}_{\text{difficult to compute in general.}}$$