Course on probabilistic graphical models
Master MVA 2015-2016
Review exercises

December 9, 2015

These exercises are not meant to provide an exhaustive coverage of the material to review for the final exam. To some extend they focus more specifically on material that was not covered in the homeworks and that corresponds to the four last lectures. Also, these exercises should not be taken as representative of the difficulty of the questions posed at the exam, although several questions of the exam are likely to have a similar style. So don't be discouraged if you find some of them difficult, in particular the problem, entitled trees, entropies and polytope, which is slightly harder. They are primarily designed to help you review and consolidate your understanding of the course.

Various exercises

1. Consider a real-valued Gaussian homogeneous Markov chain specified by the recurrence $X_{t+1} = \rho X_t + \epsilon$ with $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Assuming that it is a Gaussian distribution, find the stationary distribution of this Markov chain, that is the distribution P such that if $\mathbb{P}(X_1 \in A) = P(A)$ then $\mathbb{P}(X_t \in A) = P(A)$.

2. Let $X_1, \dots, X_d$ be independent discrete random variables. Let $H(X_1, \dots, X_d)$ and $H(X_i)$ denote respectively the entropy of the joint distribution of $(X_1, \dots, X_d)$ and the entropy of the marginal distribution of X_i. Show that
$$H(X_1, \dots, X_d) = \sum_{j=1}^{d} H(X_i)$$

3. **Online Bayesian computations.** Let $(x_1, \dots, x_n)$ be a sequence of data observations. Let $\mathcal{P}_\Theta = \{p(x|\theta),\, \theta \in \Theta\}$ be a model for each of these observations and $\pi_0(\theta)$ a prior density on $\Theta \subset \mathbb{R}^d$. Define recursively
$$\pi_t(\theta) = \frac{p(x_t|\theta)\, \pi_{t-1}(\theta)}{p_t(x_t)} \qquad \text{with} \qquad p_t(x_t) := \int p(x_t|\theta)\, \pi_{t-1}(\theta)\, d\theta.$$
Prove that $\pi_n(\theta)$ is the posterior distribution given all the data $\pi_n(\theta) = p(\theta|x_1, \dots, x_n)$.

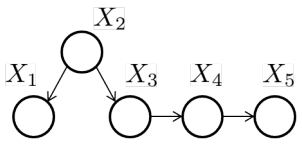
4. **EM *vs* gradient descent in latent variable models.** Consider a latent variable model, with two variables (X, H), where X is observed and H is an unobserved latent variable. Let $\mathcal{P}_\Theta = \{p_\theta(x, h),\, \theta \in \Theta\}$ be a model for the pair (X, H), and consider the problem of learning the parameters in the model from observations of X alone, with marginal likelihood $p_\theta(x) = \int p_\theta(x, h)\, dh$. We have seen in class the EM algorithm, whose principle is to iteratively maximize a complete expected log-likelihood of the form $\mathbb{E}_{q_t}[\log p_\theta(x, H)]$ for $q_t(h) = p_{\theta_t}(h|x)$. The motivation was that the marginal log-likelihood $\log p_\theta(x)$ is non-convex and therefore not so simple to optimize. But the fact that a function is non-convex does not make it impossible to use methods such as gradient descent as soon as the log-likelihood is differentiable. In this exercise, we consider that option.

(a) Show that under reasonable assumptions, if $g(\theta) = \nabla_\theta \log p_\theta(x)$ then the gradient of the marginal log-likelihood at a current value θ_t, $g(\theta_t)$, is in fact equal to the expected value under some distribution q_t that you will specify of the gradient of the complete log-likelihood.

(b) Explain how you would use this to compute $g(\theta_t)$ and why the previous result suggests that a step of probabilistic inference can in general not be avoided in the algorithm to compute that gradient.

5. Consider a distribution p that factorizes according to a directed graph G. Write p as an explicit product of potentials that show that p factorizes also with respected to the moralized graph associated to G.

6. Consider running the sum-product algorithm on the following directed graphical model (DGM), where we suppose that each random variable takes value in $\{1, \ldots, K\}$:



Express all answers as functions of the conditional probabilities $p(x_i|x_\pi)$ for all i.

(a) What is the message $m_{1\to 2}(x_2)$ sent from node X_1 to node X_2 during sum-product? Gives its simplest form.

(b) Suppose that we observe the value of $X_3 = \bar{x}_3$ and that we want to compute $p(x_2|\bar{x}_3)$. Gives the message $m_{3\to 2}(x_2)$ sent from node X_3 to node X_2.

(c) Give the expression in terms of messages during sum-product for $p(x_2|\bar{x}_3)$, as well as its simplified form.

Trees, entropies and polytopes

In the entire problem, $[K]$ denotes the set $\{1, \ldots, K\}$.

Part A

Consider a distribution $p(x_1, \ldots, x_d)$ that factorizes according to an undirected tree $T = (V, E)$. Without loss of generality, we will assume that X_i is a binary indicator vector of dimension K, i.e. that X_i takes values in $\{0,1\}^K \cap \left\{x \mid \sum_{k=1}^K x_k = 1\right\}$.

1. Show that the joint distribution on $(X_1, \ldots, X_d)$, i.e. $p(x_1, \ldots, x_d)$, can be always be expressed as an explicit function of the marginal distributions $\big(p_i(x_i)\big)_{i\in V}$ and of the pairwise distributions $\big(p_{ij}(x_i, x_j)\big)_{\{i,j\}\in E}$ only.

2. Give an expression of $p(x_1, \ldots, x_d)$ as a function of $(p_i)_{i\in V}$ and $(p_{ij})_{\{i,j\}\in E}$ that does not rely on some arbitrary orientation of the tree. (You might have to guess what the general form is and prove that it is correct by induction).

3. Deduce from the previous formula that the entropy $H(X_1, \ldots, X_d)$ for a distribution that factorizes according to a tree can be expressed as a function of $H(X_i)$ for all $i \in V$ and of all $I(X_i, X_j)$ for all $\{i, j\} \in E$.

4. Show that it can alternatively be expressed as a function of $H(X_i)$ for $i \in V$, $H(X_i, X_j)$ for $\{i, j\} \in E$ and of the degrees d_i of all nodes.

5. If the distribution $p(x_1, \ldots, x_d) > 0$, show that it can be written in exponential family form. What are the elements of the vector $\phi(x)$ of sufficient statistics? What are the components of the vector η of natural parameters?

6. What are the components of the vector μ of moment parameters?

7. Show that given a vector of moment parameters μ, the corresponding vector of natural parameters η can be expressed in closed form as a simple function of μ. Explicit the function.

8. Show that the entropy $H(X_1, \ldots, X_d)$ can be expressed as a function of μ only. Explicit that function.

Part B

We consider the graphical model $\mathcal{P}_G$ consisting of all distributions over $\mathcal{X} := \mathcal{X}_1 \times \ldots \times \mathcal{X}_d$ with $|\mathcal{X}_i| = K_i < \infty$ that factorize according to some undirected graph $G = (V, E)$. For convenience, we assume that $K_i = K$ for all i and identify $\mathcal{X}_i$ with a copy of $\{0,1\}^K \cap \{x \mid \sum_{k=1}^K x_k = 1\}$.

Consider the set $\mathcal{H}_G$ defined by

$$\mathcal{H}_G := \{\mu = \big((\mu_i)_{i\in V}, (\mu_{ij})_{\{i,j\}\in E}\big)^\top \mid \mu_i \in [0,1]^K, \mu_{ij} \in [0,1]^{K\times K}\}.$$

The *local polytope* of $\mathcal{P}_G$ is the polytope $\mathcal{L}_G$ defined by

$$\mathcal{L}_G := \Big\{\mu \in \mathcal{H}_G \mid \forall \{i,j\} \in E,\ \sum_{l=1}^K [\mu_{ij}]_{kl} = [\mu_i]_k,\ \sum_{k=1}^K [\mu_{ij}]_{kl} = [\mu_j]_l\Big\}.$$

The *moment polytope* (or *marginal polytope*) is the set $\mathcal{M}_G$ defined by

$$\mathcal{M}_G = \{\mu \in \mathbb{R}^{Kd+K^2(d-1)} \mid \exists p \in \triangle_{\mathcal{X}}, \quad \mathbb{E}_p[\phi(X)] = \mu\},$$

where $\triangle_{\mathcal{X}}$ is the set of all probability distributions over $\mathcal{X}$, and

$$\phi(x) = \big((x_{ik})_{i\in V, k\in[K]}, (x_{ik}x_{jl})_{\{i,j\}\in E,\, k,l\in[K]}\big)^\top.$$

1. Show that for any graph G, we have $\mathcal{M}_G \subset \mathcal{L}_G$.

2. If $G = T$ is a tree, then use the results of part A to show that for any element $\tau \in \mathcal{L}_T$ such that all components of τ are strictly positive, it is possible to construct a distribution in $\mathcal{P}_T$ whose moments are equal to τ.

3. Use the result of the previous question to show that for a tree T, we must have $\mathcal{M}_T = \mathcal{L}_T$.

Bayesian regression

Consider the Gaussian probabilistic conditional model seen in class for linear regression in which given a pair of variables (X, Y) with X taking values in $\mathbb{R}^d$ and Y in $\mathbb{R}$, we model the conditional distribution of Y given $X = x$ by a Gaussian distribution $\mathcal{N}(w^\top x, \sigma^2)$ parametrized by w and σ^2. Assume that σ^2 is fixed and w unknown and that the problem of learning the linear regression is approached from a Bayesian point of view, by placing a Gaussian prior distribution on w of the form $\mathcal{N}(0, \tau^2 I_d)$.

1. Compute the parameters of the joint distribution of (w, Y).

2. Compute de posterior distribution on w.

3. Compute the predictive distribution over a new output variable y' given a new input x'.

Chromatic Gibbs sampler

It is interesting to consider variants of Gibbs sampling in which several variables are resampled at the same time and that could be executed in parallel. A natural naive attempt is to try at time t to resample in parallel each variable X_i from a Bernoulli distribution with probability $\sigma\big(\eta_i + \sum_{j \in N_i} \eta_{ij} x_j^{(t-1)}\big)$, where σ is the logistic function.

- Find a counterexample with two nodes that shows that this parallel update does not converge to the Gibbs distribution.

We will focus on the case where the graph is a two-dimensional grid of size $n \times m$ with each node not on the boundary connected to four neighbors. We partition the nodes according to a checkerboard pattern: let $A = \{(i,j) \in [n] \times [m] \mid i + j \text{ is even } \}$ and $B = ([n] \times [m]) \backslash A$ with the notation $[k] = \{1, \dots, k\}$. Consider a reduced graph composed of two nodes associated with the variables X_A and X_B, with $X_A = (X_{i,j})_{(i,j) \in A}$ and likewise for X_B. Consider, on this reduced graph, the standard Gibbs sampling scheme that samples X_A conditionally on X_B and then X_B conditionally on X_A.

- Characterize the conditional distribution of $X_A | X_B$ and show how to sample from $X_A | X_B$ easily. Propose an efficient algorithm to partially parallelize Gibbs sampling.
- How would you generalize this idea to a general graph?