Course on probabilistic graphical models
Master MVA 2017-2018
Practice exercises

October 19, 2017

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 is not covered in the homeworks. Also, all 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. Some exercises are easy, and a few can be much harder so don't be discouraged if you find some of them difficult. They are primarily designed to help you review and consolidate your understanding of the course.

1. Let $p(x, y)$ be a joint probability distribution on two dependent random variables X and Y taking both values in $\{0, 1\}$
   (a) Write the joint distribution on (x, y) in exponential form
   (b) What is the vector of sufficient statistics
   (c) Show that this distribution is in fact just a multinomial distribution
   (d) Generalize the to joint distribution of n dependent Bernoulli random variables

2. Consider a directed graphical model on the random variables $(X_1, \ldots, X_d)$ in which the likelihood takes the form
$$p_\theta(x) = \prod_{j=1}^{d} p_{\theta_j}(x_j|x_{\pi_j}) \quad \text{with} \quad \theta = (\theta_1, \ldots, \theta_d) \in \Theta_1 \times \ldots \times \Theta_d.$$
   Assume that we have an i.i.d. sample of observations $x^{(1)}, \ldots, x^{(n)}$ with $x^{(i)} = \left(x_1^{(i)}, \ldots, x_d^{(i)}\right)$.
   (a) Explain why the maximum likelihood estimator can a priori be estimated easily.
   (b) Why is it key to make the assumption that the set of possible values of the parameters is of the form $\Theta_1 \times \ldots \times \Theta_d$?
   (c) Consider now the case in which the DAG is a tree and where X_j takes values in $\{0, 1\}$. Let $\theta_1 := \mathbb{P}(X_1 = 1)$ and, for $j > 1$, let $\theta_{j1} := \mathbb{P}(X_j = 1|X_{\pi_j} = 1)$ and $\theta_{j0} := \mathbb{P}(X_j = 1|X_{\pi_j} = 0)$; let $\theta = (\theta_1, \theta_{21}, \theta_{20}, \ldots, \theta_{d1}, \theta_{d0})$ be parameterizing the model. What is the maximum likelihood estimator for θ based on the sample of size n introduced in the previous question?
   (d) If we assume now that $\theta_{jk} = \theta_{j'k}$ for all $2 \leq j, j' \leq d$ and $k \in \{0, 1\}$, is it still possible to compute easily the maximum likelihood estimator? If so compute it.

3. Consider the joint Gaussian mixture distribution on $(X, Y) \in \mathbb{R}^2$ of the form:
$$p(x, y) = \frac{1}{4\sqrt{\pi}}\left(e^{-[x^2+y^2]} + e^{-[(x-1)^2+y^2]} + e^{-[(x-1)^2+(y-1)^2]} + e^{-[x^2+(y-1)^2]}\right)$$

(a) Are X and Y correlated?

(b) Are X and Y independent?

(c) Propose a random variable Z such that $X \perp\!\!\!\perp Y | Z$

4. Consider X and Y two independent Bernoulli random variable. Let $Z = \max(X, Y) - XY$.

   (a) Which of the following statement are true: $Y \perp\!\!\!\perp Z$, $X \perp\!\!\!\perp Y \mid Z$, $X \perp\!\!\!\perp Z \mid Y$, $Z \perp\!\!\!\perp Y \mid X$?

   (b) What is the smallest undirected graphical model containing the joint distribution over (X, Y, Z)? Is this satisfactory?

5. Assume that $(X_1, Y_1) \perp\!\!\!\perp (X_2, Y_2)$. Is it true that $X_1 \perp\!\!\!\perp X_2 \mid (Y_1, Y_2)$? Prove or disprove.

6. Let $X_1, \ldots, X_d$ be independent discrete random variables. Let $H(X_1, \ldots, X_d)$ and $H(X_i)$ denote respectively the entropy of the joint distribution of $(X_1, \ldots, X_d)$ and the entropy of the marginal distribution of X_i. Show that

$$H(X_1, \ldots, X_d) = \sum_{j=1}^{d} H(X_i)$$

7. **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.

8. 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.

9. Let $p(x_1, \ldots, x_d)$ be a distribution that factorizes according to an undirected tree with node set V and edge set E. Let $p(x_i, x_j)$ denote the marginal distribution over the pair of variables (X_i, X_j) and $p(x_i)$ be the marginal distribution over the variable X_i.

   (a) Show that

$$p(x_1, \ldots, x_d) = \prod_{i \in V} p(x_i) \prod_{\{i,j\} \in E} \frac{p(x_i, x_j)}{p(x_i)p(x_j)}.$$

   (b) What is the exponential family form of the distribution if all variables X_i take values in $\{0, 1\}$?