Machine Learning 2 **Due: 13:00, April 16, 2018**

Homework 2

Instructors: Stephan Bongers, Stephan Alaniz Kupsch
Email: s.r.bongers@uva.nl, s.a.alanizkupsch@uva.nl

You are allowed to discuss with your colleagues but you should write the answers in *your own words*. If you discuss with others, write down the name of your collaborators on top of the first page. No points will be deducted for collaborations. If we find similarities in solutions beyond the listed collaborations we will consider it as cheating.
We will not accept any late submissions under any circumstances. The solutions to the previous homework will be handed out in the class at the beginning of the next homework session. After this point, late submissions will be automatically graded zero.
$\star$ denotes bonus exercise. You earn extra points for solving each bonus exercise. All bonus points earned will be added to your total homework points.

Problem 1. (0.5 + 0.5 + 1 + 1 = 3 pts)

1. Given three discrete random variables X, Y and Z. Give the definition of the mutual information $I(X;Y)$ and the conditional mutual information $I(X;Y|Z)$. Explain what the (conditional) mutual information measures.

Consider three variables x, y, $z \in \{0, 1\}$ having the joint distribution $p(x, y, z)$ given in Table 1.

2. Evaluate the quantity $I(X;Y)$ and show that it is greater than zero. Hint: Compute the tables for $p(x, y)$, $p(x)$ and for $p(y)$. Moreover, remember that we use the convention that $0 \cdot ln(0) := 0$. Interpret this result, i.e. what does it mean that $I(X;Y) > 0$?

3. Evaluate $I(X;Y|Z)$ and show that it is equal to zero. Hint: Compute the tables for $p(x, y|z)$, $p(x|z)$ and for $p(y|z)$. Interpret this result, i.e. what does it mean that $I(X;Y|Z) = 0$?

4. Show that $p(x, y, z) = p(x)p(z|x)p(y|z)$, and draw the corresponding directed graph.

Table 1: The joint distribution over three binary variables.

x	y	z	$p(x, y, z)$
0	0	0	0.192
0	0	1	0.144
0	1	0	0.048
0	1	1	0.216
1	0	0	0.192
1	0	1	0.064
1	1	0	0.048
1	1	1	0.096

Problem 2. (2 pts)
Consider all the Bayesian networks consisting of three vertices X, Y and Z. Group them into clusters such that all the graphs in each cluster encode the same set of independence relations. Draw those clusters and write down the set of independence relations for each cluster.

Problem 3. (1 + 1 = 2 pts)

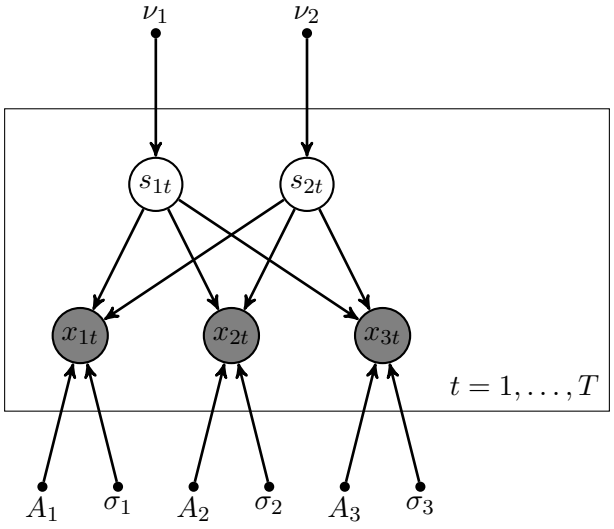
1. Given distributions p and q of a continuous random variable, Kullback-Leibler divergence of q from p is defined as
$$\mathcal{KL}(p||q) = -\int p(\mathbf{x}) \ln \left\{ \frac{q(\mathbf{x})}{p(\mathbf{x})} \right\} d\mathbf{x}$$
Evaluate the Kullback-Leibler divergence when $p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $q(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\mathbf{m}, \mathbf{L})$

2. Entropy of a distribution p is given by
$$\mathcal{H}(p) = -\int p(\mathbf{x}) \ln p(\mathbf{x}) d\mathbf{x}$$
Derive the entropy of the multivariate Gaussian $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$

Problem 4. (0.25 + 0.25 + 0.25 + 0.25 + 0.5 + 0.25$\star$ + 0.25$\star$ + 0.25 + 0.25 = 2 pts + 0.5$\star$ pts)
Consider a time sequence of T samples from K_s *statistically independent* sound sources: $\{s_{it}\} = (s_{i1}, ..s_{iT})$, where i labels the source and t the time it was emitted. We record K_x *noisy* linear mixtures of these sound sources:
$$x_{kt} = \sum_{i=1}^{K_s} A_{ki} s_{it} + \epsilon_{kt} \quad t = 1..T, \quad k = 1..K_x$$
$$s_{it} \sim \mathcal{T}(0, \nu_i), \qquad \epsilon_{kt} \sim \mathcal{N}(0, \sigma_k^2).$$
where s_{it} is distributed as a zero mean Student's T distribution with ν_i degrees of freedom and ϵ_{kt} a noise random variable drawn from a zero mean normal (Gaussian) distribution with standard deviation σ_k. We assume that the sources were generated independently and we also assume that there are no statistical dependencies between samples generated at different points in time.

1. Explain why this is an ICA model.

For $K_s = 2$ sound sources and $K_x = 3$ recordings we have the following graphical model:



Where $A_k = (A_{k1}, A_{k2})$ and we used the plate notation to indicate the replication over T.

2. Write a general (Bayesian network) expression for the joint probability distribution

$$p(\{s_{1t}\}, \{s_{2t}\}, \{x_{1t}\}, \{x_{2t}\}, \{x_{3t}\}), \quad t = 1..T$$

   Factorize the distribution into smaller conditional and marginal distributions as much as possible. Use explicit (conditional) distributions such as Normal and Student's T distributions instead of a generic form "p" as much as possible.

3. Explain what the term "explaining away" means and indicate if this explaining away phenomenon is present in the ICA model under discussion.

4. Since samples across time t are independent, we will ignore the index t in the following two questions (you may imagine $t = 1$). For all of the (conditional) independence expressions below, state if they are true or (typically) false:

   (a) $x_1 \perp\!\!\!\perp x_2 | \varnothing$
   (b) $s_1 \perp\!\!\!\perp s_2 | \varnothing$
   (c) $x_1 \perp\!\!\!\perp s_1 | \varnothing$
   (d) $x_1 \perp\!\!\!\perp x_2 | \{s_1, s_2\}$
   (e) $x_1 \perp\!\!\!\perp x_2 | s_1$
   (f) $s_1 \perp\!\!\!\perp s_2 | \{x_1, x_2, x_3\}$
   (g) $s_1 \perp\!\!\!\perp s_2 | x_1$
   (h) $x_1 \perp\!\!\!\perp s_1 | \{s_2, x_2, x_3\}$

5. What is the Markov blanket of s_1? What is the Markov blanket of x_1?

From now on we will assume that the number of sources and the number of recordings are the same (complete ICA), i.e. $K_x = K_s = K$. We will also assume that the relation between sources and recordings is deterministic, i.e.

$$x_{kt} = \sum_{i=1}^{K} A_{ki} s_{it} \quad t = 1..T, \quad k = 1..K$$
$$s_{it} \sim \mathcal{T}(0, \nu_i).$$

We call $W = A^{-1}$ the inverse of the mixing matrix A (aka the "unmixing matrix"), such that

$$s_{it} = \sum_{k=1}^{K} W_{ik} x_{kt} \quad t = 1, ..T, \quad i = 1..K$$

In the following question, you may use the general expression:

$$p_X(x) = p_S(s(x)) \; |\det Jac(s \to x)|$$

where $Jac(s \to x)$ is the Jacobian for a transformation from the random variable s to x.

6. $\star$ Write an explicit expression in terms of W and the sources' student's T distributions $\mathcal{T}(s_i|0, \nu_i)$ of the probability:

$$p(\{x_{kt}\}|W, \{\nu_i\}) \quad t = 1..T, \quad k = 1..K$$

7. $\star$ Write down the *log-likelihood* of the complete deterministic ICA model above.

8. Explain in detail the "stochastic gradient ascent" optimization algorithm to maximize the log-likelihood of the previous question. Note: you do not have to derive or provide the expression of the gradient; instead you can provide a general description of the algorithm.

9. In which limit do you expect overfitting: $K \gg T$ or $T \gg K$? Explain your answer.