STATISTICAL MACHINE LEARNING (WS2020)
SEMINAR 6

Assignment 1.
a) Let X be a discrete random variable taking values $x \in \{1, \dots, K\}$. Compute the Kullback-Leibler divergence $D_{KL}(q(x) \,\|\, p(x))$ for the following distributions: $p(x)$ is uniform and $q_k(x) = \delta(x, k)$, where $\delta(.,.)$ denotes the Kronecker symbol.

b) Let $X \in \mathbb{R}$ be a real valued random variable. Consider the following two distributions: $p(x)$ is a normal distribution with zero mean and unit variance and $q_\alpha(x)$ is uniform on the interval $|x - \alpha| < \epsilon$. Compute their Kullback-Leibler divergence in the limit $\epsilon \to 0$.

Assignment 2. Let $X \in \mathbb{R}^n$ be a random vector with discrete valued components.

a) Assume that the two distributions $p(x)$ and $q(x)$ factorise w.r.t. the components of x, i.e. $p(x) = \prod_{i=1}^n p_i(x_i)$ and similarly $q(x) = \prod_{i=1}^n q_i(x_i)$. Prove that their KL-divergence is the sum of the KL-divergences of the marginal distributions, i.e.

$$D_{KL}(q(x) \,\|\, p(x)) = \sum_{i=1}^{n} D_{KL}(q_i(x_i) \,\|\, p_i(x_i)).$$

b) Let us now assume that $p(x)$ is an arbitrary distribution, and we want to approximate it by the factorising distribution q that has minimal KL-divergence $D_{KL}(p(x) \,\|\, q(x))$. Prove that the optimum is attained if the factors $q_i(x_i)$ coincide with the marginal probabilities of $p(x)$.

Assignment 3. Consider the block-coordinate ascent for the lower bound of the log-likelihood (in the EM algorithm)

$$L_B(\theta, \alpha, \mathcal{T}^m) = \frac{1}{m} \sum_{i=1}^{m} \sum_{y \in \mathcal{Y}} \alpha_i(y) \log p_\theta(x^i, y) - \frac{1}{m} \sum_{i=1}^{m} \sum_{y \in \mathcal{Y}} \alpha_i(y) \log \alpha_i(y).$$

The E-step requires to maximise it w.r.t. α-s for fixed θ.
a) Show that the maximisation decomposes into independent maximisation tasks of the type

$$\sum_{y \in \mathcal{Y}} \alpha(y) \log p_\theta(x^i, y) - \sum_{y \in \mathcal{Y}} \alpha(y) \log \alpha(y) \to \max_{\alpha}$$
$$\text{s.t.} \sum_{y \in \mathcal{Y}} \alpha(y) = 1 \;\text{ and }\; \alpha(y) \geqslant 0 \;\; \forall y \in \mathcal{Y}$$

b) Prove that the function

$$g(x) = \begin{cases} x \log x & \text{if } x > 0, \\ 0 & \text{if } x = 0 \\ \infty & \text{otherwise} \end{cases}$$

is convex for $x > 0$ and implicitly accounts for the constraint $x \geqslant 0$. Conclude that the optimisation task in a) is convex.
c) Analyse the Lagrange dual of this task and deduce the solution $\alpha^*(y) = p_\theta(y \mid x^i)$.

Assignment 4. Consider the following probabilistic model for real valued sequences $\boldsymbol{x} = (x_1, \ldots, x_n)$, $x_i \in \mathbb{R}$ of fixed length n. Each sequence is a combination of a leading part $i \leqslant k$ and a trailing part $i > k$. The boundary $k = 1, \ldots, n$ is random with some categorical distribution $\boldsymbol{\pi} \in \mathbb{R}^n_+$, $\sum_k \pi_k = 1$. The values x_i, in the leading and trailing part are statistically independent and distributed with some probability density function $p_1(x)$ and $p_2(x)$ respectively. Altogether the distribution for pairs $(\boldsymbol{x}, k)$ reads

$$p(\boldsymbol{x}, k) = \pi_k \prod_{i=1}^{k} p_1(x_i) \prod_{j=k+1}^{n} p_2(x_j). \tag{1}$$

The densities p_1 and p_2 are known. Given an i.i.d. sample of sequences $\mathcal{T}^m = \{\boldsymbol{x}^\ell \in \mathbb{R}^n \mid \ell = 1, \ldots, m\}$, the task is to estimate the unknown boundary distribution $\boldsymbol{\pi}$ by the EM-algorithm.
a) The E-step of the algorithm requires to compute the values of auxiliary variables $\alpha_\ell^{(t)}(k) = p(k \mid \boldsymbol{x}^\ell)$ for each example $\boldsymbol{x}^\ell$ given the current estimate $\boldsymbol{\pi}^{(t)}$ of the boundary distribution. Give a formula for computing these values from model (1).
b) The M-step requires to solve the optimisation problem

$$\frac{1}{m} \sum_{\ell=1}^{m} \sum_{k=1}^{n} \alpha_\ell^{(t)}(k) \log p(\boldsymbol{x}^\ell, k) \to \max_{\boldsymbol{\pi}}.$$

Substitute the model (1) and solve the optimisation task.