STATISTICAL MACHINE LEARNING (WS2017)
SEMINAR 5

Assignment 1. Consider the following parameter estimation task. You are given i.i.d. training data $\mathcal{T}^m = \{x_i \in \mathbb{R} \mid i = 1, 2, \ldots, m\}$ generated from the normal distribution

$$p_\mu(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

and the task is to estimate its unknown mean μ.

a) Show that the maximum likelihood estimator is given by the arithmetic mean of the training data, i.e.

$$\mu^* = e_{ML}(\mathcal{T}^m) = \frac{1}{m} \sum_{i=1}^{m} x_i.$$

Prove that this estimator is unbiased.

b) Compute the variance of the maximum likelihood estimator, i.e.

$$\mathbb{E}_\mu\big[(\mu - e_{ML}(\mathcal{T}^m))^2\big].$$

How does it depend on σ, μ and m?

c) Someone (let us call him Mr. Y) proposes an even simpler estimator - he suggests to estimate the unknown mean μ of the distribution just by setting $\mu^* = x_1$ and dropping the rest of the training data. Moreover, he claims that his estimator is also unbiased. Show that this is true. Why would you nevertheless consider his estimator inferior to the maximum likelihood estimator?

Assignment 2. Consider the exponential family

$$p_{\boldsymbol{u}}(x, y) = \frac{1}{Z(\boldsymbol{u})} \exp \langle \boldsymbol{\phi}(x, y), \boldsymbol{u} \rangle$$

where $\boldsymbol{u} \in \mathbb{R}^k$ is a parameter vector, $\boldsymbol{\phi}(x, y) \in \mathbb{R}^k$ is a generalised feature map, $x \in \mathcal{X}$, $y \in \mathcal{Y}$ and $Z(\boldsymbol{u}) = \sum_{x,y} \exp \langle \boldsymbol{\phi}(x, y), \boldsymbol{u} \rangle$ is a normalising factor.

a) Prove that each model in this class is identifiable, provided that the affine hull of the set of vectors $\big\{ \boldsymbol{\phi}(x, y) \;\big|\; x \in \mathcal{X}, y \in \mathcal{Y} \big\}$ is the entire space $\mathbb{R}^k$ (or equivalently, there is no hyperplane containing all vectors).

b) Suppose that the parameter vector is bounded by $\|\boldsymbol{u}\| \leqslant R$ and assume that the components of the vectors $\boldsymbol{\phi}(x, y)$ are bounded in some interval $[a, b]$. Prove the Uniform Law of Large Numbers for the corresponding Maximum Likelihood Estimator by performing the following steps

(1) Denote the training data by $\mathcal{T}^m = \{(x^i, y^i) \mid i = 1, 2, \dots, m\}$, the log-likelihood of $\mathcal{T}^m$ by $L(\boldsymbol{u}, \mathcal{T}^m)$ and the expected log-likelihood by $L(\boldsymbol{u}) = \mathbb{E}_{\boldsymbol{v}} L(\boldsymbol{u}, \mathcal{T}^m)$, where $\boldsymbol{v} \in \mathbb{R}^k$ is the true but unknown model.
(2) Deduce that

$$L(\boldsymbol{u}, \mathcal{T}^m) - L(\boldsymbol{u}) = \langle \mathbb{E}_{\mathcal{T}^m} \boldsymbol{\phi} - \mathbb{E}_{\boldsymbol{v}}(\boldsymbol{\phi}), \boldsymbol{u} \rangle$$

holds, where $\mathbb{E}_{\mathcal{T}^m} \boldsymbol{\phi}$ denotes the arithmetic mean of the vectors $\boldsymbol{\phi}(x, y)$ on the training data and $\mathbb{E}_{\boldsymbol{v}}(\boldsymbol{\phi})$ denotes their expectation w.r.t. the true model.
(3) Prove that

$$\max_{\|\boldsymbol{u}\| \leqslant R} |L(\boldsymbol{u}, \mathcal{T}^m) - L(\boldsymbol{u})| = \|\mathbb{E}_{\mathcal{T}^m} \boldsymbol{\phi} - \mathbb{E}_{\boldsymbol{v}}(\boldsymbol{\phi})\| R$$

holds.
(4) Conclude the ULLN for MLE-s in this model class.

c) Prove that the logarithm of the probability

$$\log p_{\boldsymbol{u}}(x, y) = \langle \boldsymbol{\phi}(x, y), \boldsymbol{u} \rangle - \log Z(\boldsymbol{u})$$

is a concave function of $\boldsymbol{u}$ by verifying the following steps.

(1) prove the formulas $\nabla_{\boldsymbol{u}} \langle \boldsymbol{a}, \boldsymbol{u} \rangle = \boldsymbol{a}$ and $\nabla_{\boldsymbol{u}} f\big(g(\boldsymbol{u})\big) = f'\big(g(\boldsymbol{u})\big) \cdot \nabla_{\boldsymbol{u}} g(\boldsymbol{u})$ for the gradient by considering partial derivatives. Here, f is a function of a real variable, f' denotes its derivative and $\boldsymbol{a}$ denotes a constant vector.
(2) Prove that the gradient of $\log Z(\boldsymbol{u})$ is

$$\nabla_{\boldsymbol{u}} \log Z(\boldsymbol{u}) = \sum_{x,y} p_{\boldsymbol{u}}(x, y) \boldsymbol{\phi}(x, y) = \mathbb{E}_{\boldsymbol{u}}(\boldsymbol{\phi}).$$

(3) Prove that the second derivative of $\log Z(\boldsymbol{u})$ is

$$\nabla^2_{\boldsymbol{u}} \log Z(\boldsymbol{u}) = \sum_{x,y} p_{\boldsymbol{u}}(x, y) \boldsymbol{\phi}(x, y) \otimes \boldsymbol{\phi}(x, y) - \mathbb{E}_{\boldsymbol{u}}(\boldsymbol{\phi}) \otimes \mathbb{E}_{\boldsymbol{u}}(\boldsymbol{\phi}) =$$
$$= \mathbb{E}_{\boldsymbol{u}}\big[(\boldsymbol{\phi} - \mathbb{E}_{\boldsymbol{u}}(\boldsymbol{\phi})) \otimes (\boldsymbol{\phi} - \mathbb{E}_{\boldsymbol{u}}(\boldsymbol{\phi}))\big]$$

(4) Deduce that the second derivative is positive semi-definite matrix and conclude that $\log Z(\boldsymbol{u})$ is convex.

Assignment 3. Consider the block-coordinate ascent for Minka's 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 $\alpha^*(y) = p_\theta(y \mid x^i)$.