STATISTICAL MACHINE LEARNING (WS2018)
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

$$p_{\mu_0}(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\mu_0)^2}{2}}$$

and the task is to estimate its unknown mean μ_0.

a) Prove that the expected log-likelihood

$$L(\mu) = \int_{-\infty}^{\infty} p_{\mu_0}(x) \log p_\mu(x)\, dx$$

has a unique global maximum at μ_0.

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

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

where

$$e_{ML}(\mathcal{T}^m) = \arg\max_{\mu} \frac{1}{m} \sum_{i=1}^{m} \log p_\mu(x).$$

Prove that this estimator is unbiased.

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

$$\mathbb{E}_{\mu_0}\big[(\mu_0 - e_{ML}(\mathcal{T}^m))^2\big].$$

How does it depend on μ_0 and m?

Assignment 2. Consider the exponential family,

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

where $\boldsymbol{u} \in \mathbb{R}^k$ is a parameter vector, $\boldsymbol{\phi}(x) \in \mathbb{R}^k$ is a feature map, $x \in \mathcal{X}$ and $Z(\boldsymbol{u}) = \sum_x \exp \langle \boldsymbol{\phi}(x), \boldsymbol{u} \rangle$ is a normalising factor. This defines a parametrised class of probability distributions. (Show that the class of univariate normal distributions is an particular example of such a family.)

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

b) Derive the formula for the log-likelihood of given training data $\mathcal{T}^m = \{(x^i) \mid i = 1, 2, \dots, m\}$. Prove that the logarithm of the probability

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

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

(1) Prove that the gradient of $\log Z(\boldsymbol{u})$ is

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

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

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

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

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

(1) Denote the log-likelihood of the training data $\mathcal{T}^m$ 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 $\boldsymbol{\phi}(x)$ 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.