# STATISTICAL MACHINE LEARNING (WS2022)
# EXAM T2 (90MIN / 24P)

**Assignment 1.** (6p) Let $\mathcal{X}$ be a set of images. Let $s\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a function which measures a dissimilarity between a pair of images and is defined as

$$s(x, x') = (\boldsymbol{\phi}(x) - \boldsymbol{\phi}(x'))^T \mathbf{W}(\boldsymbol{\phi}(x) - \boldsymbol{\phi}(x'))\,,$$

where $\boldsymbol{\phi}\colon \mathcal{X} \to \mathbb{R}^n$ is a function extracting $n$ features from image $x \in \mathcal{X}$, and $\boldsymbol{W} \in \mathbb{R}^{n\times n}$ is a square matrix. Consider a classifier $h\colon \mathcal{X} \times \mathcal{X} \to \{-1, +1\}$ assigning a pair of images $(x, x') \in \mathcal{X} \times \mathcal{X}$ into the positive class if their dissimilarity $s(x, x')$ is not higher than a threshold $b \in \mathbb{R}$ and to the negative class otherwise, i.e.

$$h(x, x'; \boldsymbol{W}, b) = \operatorname{sign}(b - s(x, x'))\,, \tag{1}$$

where $\boldsymbol{W} \in \mathbb{R}^{n\times n}$ and $b \in \mathbb{R}$ are parameters of the classifier.

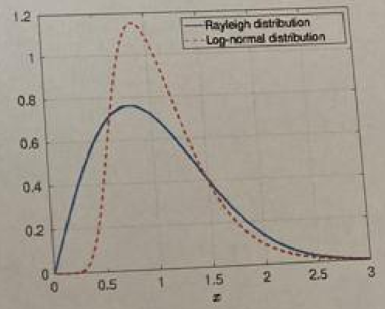
**a)** Show that (1) is an instance of a linear classifier.

**b)** Let $\mathcal{T}^m = \{(x_A^j, x_b^j, y^j) \in (\mathcal{X} \times \mathcal{X} \times \{+1, -1\}) \mid j = 1, \dots, m\}$ be a set of training examples composed of a pair of images $(x_A, x_b)$ and their label $y$. Describe a variant of the Perceptron algorithm which finds the parameters $\boldsymbol{W} \in \mathbb{R}^{n\times n}$ and $b \in \mathbb{R}$ such that the classifier (1) predicts all examples from $\mathcal{T}^m$ correctly provided such parameters exist.

**c)** Let $\mathcal{H} = \{h(x, x'; \boldsymbol{W}, b) = \operatorname{sign}(b - s(x, x')) \mid \boldsymbol{w} \in \mathbb{R}^{n\times n}, b \in \mathbb{R}\}$ be hypothesis space containing all classifiers (1). What is the Vapnik-Chervonenkis dimension of $\mathcal{H}$?

**d)** Does the uniform law of large numbers applies for the hypothesis space $\mathcal{H}$ defined in the assignment c) ?

**Assignment 2** (5p). Consider the following two probability density distributions.



The first distribution is a Rayleigh distribution with unit mean:

$$f(x) = \frac{\pi x}{2} \exp\left(\frac{-x^2 \pi}{4}\right),$$

where $x \in [0, +\infty)$, $\mathbb{E}[x] = 1$, and $\mathbb{E}[x^2] = \frac{4}{\pi}$. The second distribution is a log-normal distribution:

$$g(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(\frac{-(\log(x))^2}{2\sigma^2}\right),$$

where $x \in (0, +\infty)$.

Find its parameter value $\sigma$ that minimises the Kullback-Leiber divergence $D_{KL}(f \parallel g) = \int_0^{+\infty} f(x) \log\left(\frac{f(x)}{g(x)}\right) \mathrm{d}x$.

**a)** First calculate $\log\left(\frac{f(x)}{g(x)}\right)$.

**b)** Approximate $\log(x)$ with its first order Taylor approximation: $\log(x) \approx x - 1$.

**c)** Optimize $\sigma$ according to $D_{KL}(f \parallel g)$.

**Assignment 3** (5p). Consider a hidden Markov model for sequences of features $x = (x_1, x_2, \dots, x_n)$ and sequences of hidden states $s = (s_1, s_2, \dots, s_n)$, $s_i \in K$, where $K$ is a finite set. The probability is given by

$$p(x, s) = p(s_1) \prod_{i=2}^{n} p(s_i \mid s_{i-1}) \prod_{i=1}^{n} p(x_i \mid s_i).$$

Given a sequence of features $x = (x_1, x_2, \dots, x_n)$, we want to compute the average number of occurrences of a specific symbol $k_* \in K$ in the sequence $s = (s_1, s_2, \dots, s_n)$, i.e.

$$\mathbb{E}_{s \sim p(s \mid x)} \sum_{i=1}^{n} [\![ s_i = k_* ]\!].$$

Give an algorithm for computing this expectation. What is its run-time complexity?

**Assignment 4** (4p). Define a neural module (layer) joining a linear layer and a *softsign* layer. Give the forward, backward and parameter messages. Consider $n$ inputs, $K$ units of the linear layer and $K$ units of the *softsign* layer, each processing the output of the corresponding unit of the preceding linear layer. Each Softsign unit applies the non-linearity:

$$f(x) = \frac{x}{|x| + 1}$$

- The forward message is defined as a function of layer outputs w.r.t. to its inputs.
- The backward message is defined as the set of derivatives of all layer outputs w.r.t. to all layer inputs.
- Finally, the parameter message is defined as the set of derivatives of all layer outputs w.r.t. to all layer parameters.

**Assignment 5** (4p). Consider a regression problem with multiple training datasets $\mathcal{T}^m = \{(x_i, y_i) \mid i = 1, \dots, m\}$ of size $m$ generated by using

$$y = f(x) + \epsilon,$$

where the $x_i$ are drawn i.i.d. from some distribution $p(x)$ and $\epsilon$ is noise with $\mathbb{E}(\epsilon) = 0$ and $\mathrm{Var}(\epsilon) = \sigma^2$. Derive the bias-variance decomposition for the regression model which averages the targets of $\mathcal{T}^m$ and approximates $f(x)$ by this constant:

$$h_m(x) = \frac{1}{m} \sum_{i=1}^{m} y_i.$$

Give the squared bias:

$$\mathbb{E}_x\left[\left(g_m(x) - f(x)\right)^2\right] = \mathbb{E}_x\left[\left(\mathbb{E}_{\mathcal{T}^m}\left(h_m(x)\right) - f(x)\right)^2\right]$$

and the variance $\mathbb{V}_{x,\mathcal{T}^m}\left(h_m(x)\right)$. Simplify both results if possible.