# STATISTICAL MACHINE LEARNING (WS2016)
# EXAM (90 MIN / 25P)

**Assignment 1. (6p)** We are given a set $\mathcal{H} = \{h_i \colon \mathcal{X} \to \{1,\ldots,100\} \mid i = 1,\ldots,1000\}$ containing $1000$ strategies each predicting the human age $y \in \{1,\ldots,100\}$ from a facial image $x \in \mathcal{X}$. The quality of a single strategy is measured by the expected absolute deviation between the predicted age and the true age

$$R^{\mathrm{MAE}}(h) = \mathbb{E}_{(x,y)\sim p}(|y - h(x)|)\,,$$

where the expectation is computed w.r.t. an unknown distribution $p(x,y)$. The empirical estimate of $R^{\mathrm{MAE}}(h)$ reads

$$R_{\mathcal{T}^m}(h) = \frac{1}{m}\sum_{j=1}^{m} |y^j - h(x^j)|$$

where $\mathcal{T}^m = \{(x^j, y^j) \in (\mathcal{X}\times\mathcal{Y}) \mid j = 1,\ldots,m\}$ is a set of examples drawn from i.i.d. random variables with the distribution $p(x,y)$. What is the minimal number of the training examples $m$ which guarantees that $R^{\mathrm{MAE}}(h)$ is in the interval $[R_{\mathcal{T}^m}(h) - 1, R_{\mathcal{T}^m}(h) + 1]$ for every $h \in \mathcal{H}$ with probability at least $95\%$?

**Assignment 2. (5p)** A generic linear classifier $h\colon \mathcal{X} \to \mathcal{Y}$ reads

$$h(\boldsymbol{x};\boldsymbol{w}) = \operatorname*{arg\,max}_{y\in\mathcal{Y}} \langle \boldsymbol{w}, \boldsymbol{\phi}(\boldsymbol{x}, y)\rangle \tag{1}$$

where $\boldsymbol{w} \in \mathbb{R}^n$ are parameters and $\boldsymbol{\phi}\colon \mathcal{X}\times\mathcal{Y} \to \mathbb{R}^n$ is a joint feature map. Given a training set $\mathcal{T}^m = \{(\boldsymbol{x}^i, y^i) \in \mathcal{X}\times\mathcal{Y} \mid i = 1,\ldots,m\}$, the SO-SVM algorithm learns the parameters of the classifier (1) by solving a convex problem

$$\boldsymbol{w}^* = \operatorname*{arg\,min}_{\boldsymbol{w}\in\mathbb{R}^n} \left(\frac{\lambda}{2}\|\boldsymbol{w}\|^2 + F(\boldsymbol{w})\right) \tag{2}$$

where $\lambda > 0$ is a regularization constant, the empirical risk proxy $F(\boldsymbol{w})$ is defined by

$$F(\boldsymbol{w}) = \frac{1}{m}\sum_{i=1}^{m} \max\{0, \max_{y\in\mathcal{Y}\setminus\{y^i\}} (\ell(y^i, y) + \langle \boldsymbol{w}, \boldsymbol{\phi}(\boldsymbol{x}^i, y)\rangle - \langle \boldsymbol{w}, \boldsymbol{\phi}(\boldsymbol{x}^i, y^i)\rangle\}$$

and $\ell\colon \mathcal{Y}\times\mathcal{Y} \to \mathbb{R}_+$ is a target loss function.

Assume we want to use the SO-SVM for learning a classifier predicting age from a face. In particular, let $\boldsymbol{x} \in \mathcal{X} = \mathbb{R}^n$ be a feature descriptor of an input facial image, $\mathcal{Y} = \{1,\ldots,100\}$ be a set of age categories and $\ell(y, y') = |y - y'|$ a loss penalizing the predictions. The age is to be predicted by a linear classifier

$$h'(\boldsymbol{x}; \boldsymbol{v}, b_1, \ldots, b_Y) = \operatorname*{arg\,max}_{y\in\mathcal{Y}} (y\langle \boldsymbol{x}, \boldsymbol{v}\rangle + b_y) \tag{3}$$

where $\boldsymbol{v} \in \mathbb{R}^n$ and $b_y \in \mathbb{R}$, $y \in \mathcal{Y}$, are unknown parameters.

**a)** Define the joint feature map $\boldsymbol{\phi}(\boldsymbol{x}, y)$ so that the classifier (3) becomes an instance of the generic classifier (1).

**b)** Reformulate the problem (2) for learning the parameters ($\boldsymbol{v} \in \mathbb{R}^n$, $b_y \in \mathbb{R}$, $y \in \mathcal{Y}$) of the age classifier (3) as a convex quadratic program.
*Hint: Replace the non-linear terms in (2) by slack variables.*

**c)** What is the number of variables and the number of linear constraints of the quadratic program?

**Assignment 3. (8p)** The probability density of the real valued random variable $X > 0$ is a mixture of two exponential distributions

$$p(x) = c\lambda_1 e^{-\lambda_1 x} + (1-c)\lambda_2 e^{-\lambda_2 x},$$

where $c$ and $(1-c)$ are the mixture weights of the two components. Given and i.i.d. training set $\mathcal{T}_m = \{x_i \in \mathbb{R}_+ \mid i = 1, 2, \ldots, m\}$, the task is to estimate the parameters $c, \lambda_1, \lambda_2$ by the maximum likelihood estimator. Derive an EM-algorithm for this task.

**a)** Formulate the algorithm. Solve the E-step and give a formula for the auxiliary variables $\alpha(x_i)$ describing the posterior probability for $x_i$ to belong to first component of the distribution. (The probability to belong to second component is given by $1 - \alpha(x_i)$.)

**b)** Show that the optimisation w.r.t. the unknown parameters $c, \lambda_1, \lambda_2$ in the M-step decomposes into independent tasks.

**c)** Derive a solution for each of them. Are these tasks concave?

**Assignment 4. (4p)** Define a module (layer) joining a linear layer and a squared error layer. Give the forward, backward and parameter messages.

**Assignment 5. (2p)** Give a formal definition of a *greedy* policy $\pi(a|s)$ by means of the corresponding state-value function $v_\pi(s)$.