Name:

Matrikelnummer: Studienkennzahl (optional):

Please write your Name and Matrikelnummer on each additional submitted page !

Number of pages: **9**

Number of additional pages (sheets):

Duration of the exam: 90 min

Exam - Vorlesungsprüfung
Machine Learning for Visual Computing
(maximum score achievable: 100 points; answers may be written in German or English)
Sample

1. Linear Basis Function Models and Training (overall 25 points)

   Consider a training set $\mathcal{T} = \{\mathbf{x}_i, t_i\}$, $i = 1, \dots N$ which has been generated by

$$t_i = f(\mathbf{x}_i) + \boldsymbol{\epsilon}_i \tag{1}$$

   where $f : \mathbb{R}^D \to \mathbb{R}$ is a linear or nonlinear function and $\boldsymbol{\epsilon}_i$ are independently drawn from $\mathcal{N}(0, \beta^{-1})$. Remember that the probability density function of the univariate and multivariate Normal distribution is

$$\mathcal{N}(z|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(z-\mu)^2\right\}, \tag{2}$$

   respectively

$$\mathcal{N}(\mathbf{z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{D}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\mathbf{z}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{z}-\boldsymbol{\mu})\right\}. \tag{3}$$

   Assume the model $\hat{f}$ of the function is a linear basis function model (LBFM), i.e. $\hat{f}(\mathbf{x}) = \mathbf{w}^T\boldsymbol{\phi}(\mathbf{x})$.

   (a) Use the probability density function of the normal distribution to show that the $\mathbf{w}$ that minimizes the *sum-of-squares* training error corresponds to the $\mathbf{w}$ with maximal *likelihood* under the above assumptions.
   (5 points)

(b) Using the prior $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \alpha^{-1}\mathbf{I})$ show that the minimization of the *regularized sum-of-squares* training error corresponds to the *maximum a-posteriori* (MAP) learning method.
(5 points)

(c) Describe how the predictive distribution $p(t|\mathbf{x}, \mathcal{T}, \alpha, \beta)$ can be derived.
(5 points)

(d) What is the mean of the predictive distribution at input position $\mathbf{x}$?
(5 points)

(e) What is the effect of regularization and what is it used for?
(5 points)

2. Binary Classification (overall 9 points)

Let the scalar values x_i and corresponding t_i be the training data (input/output pairs) of a linear basis function model:

$$x_1 = -3, x_2 = -1, x_3 = 1, x_4 = 3$$
$$t_1 = -1, t_2 = 1, t_3 = 1, t_4 = -1$$

Write down the respective weight vector $\mathbf{w}$ of each of the two discrimination functions $d(x) = \mathbf{w}^T \Phi(x)$ given the following two basis transformations, such that the training set is correctly classified in each case:

(a)

$$\Phi(x) = \begin{pmatrix} 1 \\ x^2 \end{pmatrix}$$

(b)

$$\Phi(x) = \begin{pmatrix} e^{-\frac{1}{2}(x-x_1)^2} \\ e^{-\frac{1}{2}(x-x_2)^2} \\ e^{-\frac{1}{2}(x-x_3)^2} \\ e^{-\frac{1}{2}(x-x_4)^2} \end{pmatrix}$$

Hint: You can find a solution weight vector by simple considerations. There is no need for elaborate calculations.

3. Linear Regression/Linear Units (8 points)

   Consider the following training set consisting of (scalar inputs) x_i and (scalar targets) t_i.

$$\begin{aligned} x_1 = -3, x_2 = -1, x_3 = 0, x_4 = 4 \\ t_1 = 6, t_2 = 2, t_3 = -2, t_4 = -8 \end{aligned}$$

   Calculate the estimated w_0 (*bias*) and w_1 of a linear unit, which correspond to minimal training error. *Hint*: Use homogeneous coordinates. Calculate with the help of the pseudo-inverse (can be calculated easily in this example).

4. Principal Component Analysis (PCA) (overall 17 points)

Consider the following data set $\mathcal{T} = \{\mathbf{x}_i\}$:

$$\mathbf{x}_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \mathbf{x}_2 = \begin{pmatrix} 6 \\ 0 \end{pmatrix}, \mathbf{x}_3 = \begin{pmatrix} 0 \\ 6 \end{pmatrix} \tag{4}$$

(a) Calculate the empirical covariance matrix? (5 points)

(b) Which one of the following matrices is the eigenvector matrix with eigenvectors of the empirical covariance matrix in its columns (with descending eigenvalue from left to right)? (single choice question) (3 points)

☐ $\frac{1}{\sqrt{2}}\begin{pmatrix} -1 & -1 \\ 1 & -1 \end{pmatrix}$ ☐ $\begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}$ ☐ $\frac{1}{\sqrt{5}}\begin{pmatrix} -2 & 1 \\ 1 & 2 \end{pmatrix}$

☐ $\frac{1}{\sqrt{2}}\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$ ☐ $\frac{1}{\sqrt{5}}\begin{pmatrix} -1 & 2 \\ 2 & 1 \end{pmatrix}$ ☐ $\frac{1}{\sqrt{17}}\begin{pmatrix} 1 & -4 \\ 4 & 1 \end{pmatrix}$

(c) Calculate the coefficients of the (new) point $\mathbf{x}_{new} = \begin{pmatrix} 2 \\ 2 \end{pmatrix}$ in the space spanned by the *first* principal component. How large is the reconstruction error of this point? (5 points)

(d) Which criterium is maximized and which is minimized by PCA? (4 points)

5. Neural Networks (overall 10 points):

   Use Figure 1 to explain how the gradient of a square loss function is evaluated for a single observation. Explain how each of the updates $\Delta w_1, \Delta w_2, \Delta w_3$ corresponding to w_1, w_2, w_3 is calculated, given all units have the same activation function $g(a)$. (Use $g(a)'$ to denote its derivative.)

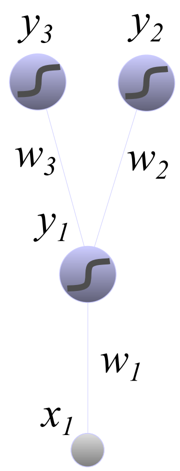

Figure 1: A MLP with 2 output units.

6. What is the *mean squared error* (MSE) and the *bias-variance dilemma* in the context of the MSE? (6 points)

7. Support Vector Machines (SVM) (overall 15 points)

Consider the following 4 linear separable points:

$$\begin{array}{llll} \omega_1: & \mathbf{x}_1 = (-1,0)^T & \mathbf{x}_2 = (0,0)^T & \mathbf{x}_3 = (0,-1)^T \\ \omega_2: & \mathbf{x}_4 = (1,1)^T & \mathbf{x}_5 = (2,1)^T & \mathbf{x}_6 = (3,0)^T, \end{array} \tag{5}$$

with corresponding class labels $t_1 = 1, t_2 = 1, t_3 = 1, t_4 = -1, t_5 = -1, t_6 = -1$. Assume in (a) and (b), that no *kernel*-functions are used, i.e. we use only inner products $k_{ij} = \mathbf{x}_i^T \mathbf{x}_j$ (k_{ij} being the elements of the kernel matrix).

(a) Sketch the data points. Which are the support vectors? Add at least two more data points with corresponding class labels, such that the set of support vectors remains unchanged (2 pts.)

(b) What is the *dual optimization problem*? (3 pts.)

(c) Demonstrate the "kernelization" of a linear model, i.e. how to use a dual formulation and a kernel function to transform $f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0$ into a nonlinear function of $\mathbf{x}$, assuming that $\mathbf{w}$ can be expressed by the training data. (5 pts.)

(d) What is the difference of the learning principle implemented by a perceptron and the one implemented by a SVM. (5 pts.)

8. Expectation Maximization (EM) (10 points)

   (a) Decribe the two steps (E-step und M-step) of the EM algorithm in the context of

      - *k-means* Clustering,
      - *maximum likelihood* fitting of *Gaussian mixture models*.

   (b) What are *latent variables* in this context?