Introduction to Intelligent Systems
Final exam

1) (1 pt) **Statistical decision theory**
Assume that you are given a set of 100 000 binary feature vectors, each of which is a binary code of the iris pattern of a person. The set contains 100 iris codes of each of 1000 persons. Describe how you would use this data to design an authentication system based on statistical decision theory.

2) (1 pt) **Dendrogram**
Construct a dendrogram for the one-dimensional data $D = \{1, 2, 3, 6, 7, 11, 14\}$. As a distance between two clusters D_i and D_j use the minimum distance between a point from D_i and a point from D_j, for all possible pairs of such points. The distance between two data points is defined as the absolute value of the difference of their values.

3) (1pt) **A two-category problem**
Consider the following multiset (bag) S of labeled single-feature patterns in a two-category problem. Each pattern has the form $(f|w)$ where f is a feature that can take one of the following values: A, B, C, D, while w is a category label that can take the values w_1 or w_2:

$$\begin{aligned} S &= \{(A|w_1), (A|w_1), (B|w_1), (A|w_1), (B|w_1), (C|w_1), (B|w_1), \\ &\quad (A|w_2), (C|w_2), (C|w_2), (D|w_2), (D|w_2), (C|w_2), (B|w_2), (C|w_2)\} \end{aligned}$$

a) Compute the misclassification impurity of S.

b) Split S in two subsets L and R using the following query: Q1: Put a pattern $(f|w)$ in L if $f = A$, otherwise put it in R. Compute the impurity reduction achieved by this split.

c) Split S in two subsets L and R using the following query: Q2: Put a pattern $(f|w)$ in L if $f = B$, otherwise put it in R. Compute the impurity reduction achieved by this split.

d) Which of the two queries Q1 and Q2 would you use for buildung a decision tree? Why?

4) (1.5 pt) **Edge Detection.**
Consider a grey-value image f.

a. (0.5pt) Sobel gradients in the image in horizontal (easterly) direction can be detected by linear filtering using the filter kernel (or mask) in Fig. 1 (left). Give the Sobel kernel to detect gradients in northerly direction.

b. (1 pt) A discrete second derivative filter in the x-direction $\frac{\partial^2}{\partial x^2}$ is defined by convolution with the kernel in Fig. 1(right). If image f is constant, the result of this filter will be zero in every pixel. Show by calculation that the result for an image $f(x, y) = ax^2 + bx + c$, is $-2a$ for each pixel with a, b, c constants. (Hint: just fill in the equation for discrete convolution for **one** arbitrary point (x, y))

-1	0	1
-2	0	2
-1	0	1

0	0	0
-1	2	-1
0	0	0

Figure 1: *Convolution masks for the Sobel x-gradient filter (left) and the second-order x-derivative filter (right).*

5) (1.5 pt) **Thresholding.**
Consider the problem of thresholding a grey-level image f in which background and objects might vary in intensity.

a. (0.5pt) Describe the difference between global and local thresholding. Which approach would you use in the above case and why?

b. (0.5pt) Describe the principle of histogram-based thresholding in the case of a bi-modal grey-level histogram. Give advantages and disadvantages of this approach.

b. (0.5pt) RATS uses the following statistic to compute a threshold in a region:

$$T = \frac{\sum_{(x,y)\in W} w(x, y) f(x, y)}{\sum_{(x,y)\in W} w(x, y)} \tag{1}$$

which is a weighted average of grey levels in the window W. What kind of operator is used as weight function w, and how do they deal with the presence of noise?

6) (1 pt) **Distance measures**
Consider an LVQ classifier with two prototypes in two dimensions, i.e. inputs $\boldsymbol{\xi} \in I\!R^2$. Assume the prototypes representing class 1 and 2 are located in $\boldsymbol{w}^{(1)} = (0, 1)^\top$ and $\boldsymbol{w}^{(2)} = (1, 0)^\top$, respectively. Give an analytical expression for the decision boundary and sketch the classification scheme graphically for the following distance measures:

a) squared Euclidean distance, i.e.

$$d(\boldsymbol{w}^{(j)}, \boldsymbol{\xi}) = \left(\boldsymbol{w}^{(j)} - \boldsymbol{\xi}\right)^2$$

b) modified Euclidean distance with global relevances, i.e.

$$d(\boldsymbol{w}^{(j)}, \boldsymbol{\xi}) = \sum_{i=1}^{2} \lambda_i \left(w_i^{(j)} - \xi_i\right)^2 \quad \text{for } \lambda_1 = 0.8, \lambda_2 = 0.2$$

c) modified Euclidean distance with local relevances, i.e.

$$d(\boldsymbol{w}^j, \boldsymbol{\xi}) = \sum_{i=1}^{2} \lambda_i^{(j)} \left(w_i^{(j)} - \xi_i \right)^2 \quad \text{for } \lambda_1^{(1)} = 0.8, \lambda_2^{(1)} = 0.2 \quad \text{and } \lambda_1^{(2)} = \lambda_2^{(2)} = 0.5$$

7) (1 pt) **VQ and LVQ**
Here you can restrict the discussion to the use of Euclidean distances and consider systems with two N-dim. prototypes $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$.

a) Consider a data set $D = \{\boldsymbol{\xi}^\mu\}_{\mu=1}^{P}$ containing N-dim. feature vectors $\boldsymbol{\xi}^\mu$. Explain the unsupervised competitive learning algorithm (Vector Quantization) discussed in class, for example in a few lines of *pseudo-code*. Which is the main difference in comparison with the K-means algorithm?

b) Now consider a set $D = \{\boldsymbol{\xi}^\mu, S^\mu\}_{\mu=1}^{P}$ containing N-dim. feature vectors $\boldsymbol{\xi}^\mu$ and labels $S^\mu \in \{1, 2\}$. Explain the supervised Learning Vector Quantization algorithm (LVQ1) discussed in class, for example in a few lines of *pseudo-code*.

8) (1 pt) **Overfitting**
Explain in your own words the problem of *overfitting*. Refer to the simple example of curve-fitting, where a polynomial is fitted to noisy data points. Explain also the meaning of the terms bias and variance in this context. In Learning Vector Quantization, which parameter would play a similar role as the degree of the polynomial fit?