Introduction to Intelligent Systems

Written Exam, 2010-11-11

Duration: **3 hours. No additional (printed, electronic or other) material is allowed – this is a closed book exam!**
In total, 9 points can be achieved, the grade for the exam will be "1.0 + number of points".

---

1) (1pt) **McCulloch Pitts Neurons and Hopfield Model**
Consider a fully connected Hopfield neural network with N neurons $S_j \in \{-1,+1\}$ $(j = 1,2,\ldots,N)$ of the McCulloch Pitts type. These neurons display either maximal activity (+1) or no activity at all (-1).
Assume that the synaptic weights $w_{ij} \in I\!R$ with $w_{ii} = 0$ $(i,j = 1,2,\ldots,N)$, and the activities at a (discrete) time t are known: $S_j(t)$ $(j = 1,2,\ldots,N)$. Write down the update equation which defines the activity $S_i(t+1)$ in the next time step.
Explain (use words and math) why the connections $w_{ij} > 0$ can be interpreted as a representation of *excitatory synapses*.

2) (1pt) **Learning Vector Quantization**
Consider a set $\{\vec{\xi}^{\mu}, \sigma^{\mu}\}_{\mu=1}^{P}$ containing N-dimensional feature vectors $\vec{\xi}^{\mu}$ and the corresponding class labels σ^{μ}. Explain supervised learning by LVQ1, the original Learning Vector Quantization algorithm which we discussed in class.
Here, you can restrict the discussion to a two class problem, i.e. $\sigma^{\mu} = 1$ or $\sigma^{\mu} = 2$. Assume that the standard Euclidean distance measure is used and discuss only the case of two N-dimensional prototypes $\vec{w}_1$ and $\vec{w}_2$.
Explain the LVQ1 algorithm in terms of a few lines of pseudocode. Be precise and provide equations which define the distance measure, the identification of the *winning prototype*, and the actual update step. Also specify the order in which the data is presented and suggest at least one possible initialization of the system. If your algorithm contains one or several parameters, explain their role in the training process.

3) (1 pt) **Vector Quantization and Clustering**
Vector Quantization (VQ), i.e. the minimization of the quantization error, can be used for (but should not be confused with) clustering.
Consider two-dimensional feature vectors $\vec{\xi}^{\mu} \in I\!R^2$ $(\mu = 1,2,\ldots P)$. Argue in terms of simple example situations (which you should illustrate graphically and explain in words) that

a) The result of VQ depends strongly on the number of prototypes used in the system.
b) Rescaling one of the features, for instance all first components ξ_1^{μ} in the data, with a constant scaling factor can drastically change the outcome of VQ. Discuss this effect in terms of two clusters and a VQ system with two prototypes.

4) (0.5 pt) **Definitions of some pattern recognition concepts**

a) Give a definition of the concept 'pattern'. Give an example of a pattern. Specify the four important constituent steps of the process of pattern recognition. Give an example.

b) What does the word 'statistical' refer to in 'statistical pattern recognition'? How does it manifest itself in practical applications?

c) What are the differences between the concepts of 'detection', 'authentication', verification', 'identification' and 'classification'. Give examples.

5) (1 pt) **Iris pattern recognition**
Assume that you are given a set of 1 000 000 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 10000 persons.

a) Describe how you would use this data to design an authentication system based on statistical decision theory.

b) How will you estimate the probability of admitting an impostor by your system (i.e. wrongly classifying the iris patterns of two different persons as stemming from one and the same person).

6) (1.5 pt) **Hierarchical clustering**
Consider the following set of points in a two-dimensional feature space: $a = (1,1), b = (15,7), c = (5,12), d = (9,9), e = (1,2), f = (6,10), g = (11,9), h = (3,3), i = (11,8), j = (4,11)$.

a) Construct a dendrogram for this set of points. 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.

b) Using the dendrogram, cluster the points in (b_1) two, (b_2) three and (b_3) four clusters.

Hint: Creating and using the full distance matrix will take you a lot of time but you do not need to do that. You can reduce the amount of computations by plotting the data in a two-dimensional space and deciding visually on the order in which points should be hierarchically clustered. The level (distance) at which two points or two clusters are merged into a (bigger) cluster must however be explicitly specified in the dendrogram.

7) (1.5 pt) **Edge Detection.**
Consider a gray-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) shown in Fig. 1 (left). Give the Sobel kernel to detect gradients in vertical (northerly) direction.

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

0	0	0
$-\frac{1}{2}$	0	$\frac{1}{2}$
0	0	0

Figure 1: (Left) Convolution masks for the Sobel x-gradient filter and (right) the first x-derivative filter.

b) (1 pt) A discrete first derivative filter in the x-direction $\frac{\partial}{\partial x}$ is defined by convolution with the kernel shown in Fig. 1 (right). If an image f is constant, the result of this filter will be zero in every pixel. Show by calculation that the result for an image $f(x, y) = ax^2 + bx + c$ (with a, b, c constants), is $2ax + b$ for each pixel. (Hint: fill in the equation for discrete convolution for point (x, y), i.e. do not use an example value such as $(10, 8)$).

8) (1.5 pt) **Thresholding.**
Consider the problem of thresholding a gray-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) Niblack thresholding uses a moving window to compute local thresholds from local statistics of the mean and standard deviation in a local window. Describe the principle of the method, which parameter choices are needed, and explain drawbacks and advantages.

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

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

which is a weighted average of gray levels in the window W. What kind of operator is used as a weight function w?