Exam: Introduction to Intelligent Systems 2012-11-09

NO OPEN BOOK! GEEN OPEN BOEK! - It is not allowed to use the course book(s) or slides or any other (printed, written or electronic) material during the exam. You may only use a simple electronic calculator. Give sufficient explanations to demonstrate how you come to a given solution or answer! The 'weight' of each problem is specified by a number of points, e.g. (1 p). You may give answers in English, Dutch or German language.
Be precise and write down equations where appropriate. Do not answer questions with just "Yes" or "No", always provide reasons/arguments for your answers!

---

1) **Hopfield Model (1 point)**
Consider a Hopfield neural network with N fully connected neurons of the McCulloch-Pitts type: $S_i(t) \in \{-1,+1\}$, $(i = 1, 2, \dots, N)$. These neurons display either maximal activity (+1) or minimal activity (-1).
Given the synaptic weights $w_{ij} \in \Re$ $(i = 1, 2, \dots, N; j = 1, 2, \dots, N)$ with $w_{ii} = 0$ and activities $S_i(t) = \pm 1$ $(i = 1, 2, \dots, N)$ at discrete time t, write down the update equation which defines the activity $S_j(t+1)$ in the next time step.
Explain (in words and math) why connections for which holds $w_{ij} > 0$ can be interpreted as *excitatory synapses* in this model.

2) **Learning Vector Quantization (1.5 points)**
Learning Vector Quantization was discussed in class as an example system for classification. Assume that we deploy standard Euclidean distance for *N*-dimensional feature vectors $\mathbf{x} \in \Re^N$ which have to be assigned to three different classes.
   a) Explain how the classification scheme is implemented by an LVQ system with one given prototype per class: $\mathbf{w}^{(j)} \in \Re^N$ $(j = 1, 2, 3)$. What are the main differences in comparison with the simple nearest neighbor classifier?
   b) For the above system, explain LVQ1, the LVQ training procedure discussed in class, in terms of a few lines of pseudo-code. Consider a given set of examples containing *N*-dimensional feature vectors $\mathbf{x}$ and the corresponding class labels. Be precise and provide equations which define the distance measure, the actual update steps, etc. If the update contains control parameters, explain their role.

3) **Unsupervised Learning (1 point)**
   a) Name and briefly explain three possible aims of unsupervised learning from high-dimensional data.
   b) Name and explain one algorithm that can be used for (one of) these aims. Explain your example algorithm in words. You do not have to specify mathematical update equations or provide pseudo-code, but it should become clear which aim the algorithm achieves and how it works.

4) **Learning Vector Quantization and Overfitting (1 point)**
   Assume somebody claims that using more prototypes in Learning Vector Quantization will always result in better classification.
   a) Is this statement definitely right or wrong, and – if so - in what sense? Give precise arguments which support or contradict the claim.
   b) Suggest a possible strategy to determine a *good* choice of prototypes, given a set of labeled example data.
   c) What kind of classifier could you obtain if the number of prototypes equals the number of labeled examples.

5) **Bayesian Decision Theory. Normal distributions. Maximum likelihood estimation. (3.5 points)**

   Let us consider a two-category classification problem, with categories (classes) A and B with prior probabilities $P_A = 1/3$ and $P_B = 2/3$. The class-conditional probability densities $p(x|A)$ and $p(x|B)$ are one-dimensional normal distributions:

$$p(x|A) \sim N(\mu_A, \sigma^2_A), \qquad p(x|B) \sim N(\mu_B, \sigma^2_B)$$

   Let us consider the sets of observations $S_A = \{-2, -0.5, 0, 0.5, 2\}$ for category A and $S_B = \{1, 2.5, 3, 3.5, 5\}$ for category B.

   Problems:

   a) Express analytically the position(s) of the optimal Bayesian decision boundary or boundaries as a function of P_A, μ_A, σ_A, P_B, μ_B, σ_B.
   b) Find the analytical conditions for having 0, 1, 2, or 3 decision boundaries. For each possible case, draw qualitative graphs of the posterior probability functions $P_A p(x|A)$ and $P_B p(x|B)$, which illustrate why the number of decision boundaries depends on the parameters P_A, μ_A, σ_A, P_B, μ_B, σ_B.
   c) Compute maximum likelihood estimations of μ_A, σ_A, μ_B, σ_B.
   d) Plot sketches of the two posterior probability density functions, together with the data sets and the estimated means and standard deviations.
   e) Find the equation of the optimal decision criterion between the two classes. Find the value(s) of the decision criterion and indicate it/them in the plot of the posterior probability distributions and data mentioned above.

f) Classify the following points: -3, -1, 1, 2, 4, 6.
g) How can you estimate the classification error of this classifier? (If possible, write a mathematical expression.)

Math reminder:

$$\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right)$$

$$x_{1,2} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

x	0.5	1	1.5	2	2.5	3	3.5	4
$\exp(-\frac{1}{2}x^2)$	0.8825	0.6065	0.3247	0.1353	0.0439	0.0111	0.0022	0.0003
$\ln(x)$	-0.6931	0	0.4055	0.6931	0.9163	1.0986	1.2528	1.3863

6. Hierarchical clustering (1 point).

	O_1	O_2	O_3	O_4	O_5	O_6
O_1		5	1	5	5	5
O_2			5	3	2	3
O_3				5	5	5
O_4					3	1
O_5						3
O_6						

The following upper triangular matrix describes the dissimilarities between six objects. Use the algorithm presented in the lectures to derive a dendrogram for these objects. Assume that the dissimilarity between two clusters of points is defined by the dissimilarity of their least dissimilar elements.