Exam in
Neural Networks and Learning Systems
- TBMI26

Time: 2009-08-21 at 14-18
Teacher: Magnus Borga, Phone: 0708 387534
Allowed additional material: Calculator, Tefyma, Beta, Physics handbook

The exam consist of three parts:

Part 1 Consists of ten questions. The questions test general knowledge and understanding of central concepts in the course. The answers should be short and given on the blank space after each question. Any calculations does not have to be presented. Maximum one point per question.

Part 2 Consists of five questions. These questions can require a more detailed knowledge. Also here, the answers should be short and given on the blank space after each question. Only requested calculations have to be presented. Maximum two points per question.

Part 3 Consists of four questions. All assumptions and calculations made should be presented. Reasonable simplifications may be done in the calculations. All calculations and answers should be on separate papers (not in the exam). Each question gives maximum five points.

The maximum sum of points is 40 and to pass the exam (grade 3) normally 18 points are required. There is no requirement of a certain number of points in the different parts of the exam. The answers may be given in English or Swedish.

The result will be reported at 090904 on the latest. The exams will then be available at IMT.

GOOD LUCK!
1. One class of learning methods tries to find a good representation of the training data, without any other information than the input signals. What is this class of methods called?

2. In most ANNs, a neuron is a very simple computational unit. Give a mathematical expression for the computation in such a unit. Explain the notations you use!

3. What does the term “stochastic gradient search” mean?

4. Mention two ways of speeding up gradient search!
5. $\mathbf{Ae} = \lambda \mathbf{e}$ is an eigenvalue problem. What is the interpretation of the largest eigenvalue ($\lambda_1$) in the following two cases:
   a) $\mathbf{A} = \mathbf{C}_{xx}$
   b) $\mathbf{A} = \mathbf{C}_{xx}^{-1} \mathbf{C}_{xy} \mathbf{C}_{yy}^{-1} \mathbf{C}_{yx}$, 
   
   where $\mathbf{C}_{xx} = E[(\mathbf{x} - \bar{\mathbf{x}})(\mathbf{x} - \bar{\mathbf{x}})^T]$, $\mathbf{C}_{yy} = E[(\mathbf{y} - \bar{\mathbf{y}})(\mathbf{y} - \bar{\mathbf{y}})^T]$ and $\mathbf{C}_{xy} = \mathbf{C}_{yx}^T = E[(\mathbf{x} - \bar{\mathbf{x}})(\mathbf{y} - \bar{\mathbf{y}})^T]$?

6. Explain why classification (in most cases) can be seen as a function approximation problem.

7. It can be shown that if $\mathbf{e}_i$ is an eigenvector to $\mathbf{A}^T \mathbf{A}$, then $\mathbf{Ae}_i$ is an eigenvector to $\mathbf{AA}^T$. When is this relationship useful? Give an example.
8. Write the cost function being optimized with Fishers linear discriminant function. Explain the notations you use!

9. What is the assumption that motivates the use of kurtosis (forth order moment) to separate mixed signals?

10. Why is the mutation operation important in genetic algorithms?
11. Draw (approximately) the decision boundary after convergence if a one layer back-propagation network with a linear activation function is trained on the data in the figure below. Also draw the decision boundary if a sigmoidal activation function is used.

12. A pair of two-dimensional variables $x$ and $y$ are analysed with CCA by presenting four pairs $x_i$ and $y_i$ as shown in the figure below. Draw the vectors corresponding to the largest canonical correlation in $x$ and $y$ respectively in the figure below!
13. Consider the two signals

\[ s_1(t) = a(t) + 0.1 \ c(t) \]

and

\[ s_2(t) = b(t) - 0.1 \ c(t) \]

where \( a(t) \), \( b(t) \) and \( c(t) \) are uncorrelated signals. These two signals are then mixed together by a two-by-two full-rank mixing matrix:

\[ x(t) = Ms(t). \]

What happens if the mixture \( x(t) \) is analyzed with CCA-based blind source separation? Why?

14. Calculate the stationary distribution, if possible, for the following two Markov models. Arrows without an associated probability number indicates the probability 1.

15. \((0 \ 1 \ 1 \ 0)\) and \((0 \ 1 \ 0 \ 1)\) are two individuals in a population in a GA. Which new individuals can be created by crossover between these two individuals? (1p)

How many different individuals can be created by crossover between two strings of length 4? (1p)
Part 3

16. Assume we have a signal

\[ x = \begin{pmatrix} x_1(t) \\ x_2(t) \end{pmatrix} \]

as given in the table below.

<table>
<thead>
<tr>
<th>t</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>(x_1(t))</td>
<td>8</td>
<td>5</td>
<td>3</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>(x_2(t))</td>
<td>1</td>
<td>0</td>
<td>-1</td>
<td>1</td>
<td>-1</td>
</tr>
</tbody>
</table>

a) We think the data volume is too large. Show how the data volume can be reduced to half the size by using principal component analysis. Also, state the resulting data volume after the reduction. (3p)

b) Show how the original signal can be reconstructed given the result from a). (2p)

17. The value of a state \(x\) given the policy \(f\) can be written as

\[ V_f(x(t)) = \sum_{i=0}^{\infty} \gamma^i r(t + i). \]

a) Write \(V_f\) as a recursive expression. (3p.)

b) Derive Bellman’s optimality equation. (2p.)
18. Assume we would like to use a two layer feed forward network according to figure 1 (i.e. one output layer and one hidden layer with two units each) in order to classify a two dimensional signal \( \mathbf{x} = (x_1 \ x_2)^T \) into two classes. The activation function
\[
\sigma(x) = \tanh(x), \text{i.e.} \quad \sigma(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}
\]
is used in both the layers. We want to use standard back propagation in order to train the network to minimize the square mean error over all the training examples, i.e.
\[
\mathcal{E} = \sum_{\mu} \| \mathbf{d}_\mu - \mathbf{y}_\mu \|^2.
\]
The network have to be trained “on line”.
State how the weights should be updated for the:

a) output layer. (2p)

b) input layer. (3p)

19. a) The training of a self organizing network (Kohonen/SOM) has just started its second iteration. The network has 4 units and a 1D topology. The training set consists of 500 two dimensional examples. It is given that the learning rate is decreasing linearly with the training examples with an initial (iteration 0) value of 0.9 and a final (iteration 500) value of 0. The neighborhood function is
\[
h(i, j) = \exp(-\frac{d(i,j)^2}{2\sigma^2}), \text{ where } \sigma \text{ decreases linearly with the training examples with an initial (iteration 0) value of 2 and a final (iteration 500) value of 0.001.}
\]
The weights before the second update are:
\[
\begin{align*}
\mathbf{w}_1 &= \begin{bmatrix} 7 \\ 1 \end{bmatrix}, \quad \mathbf{w}_2 = \begin{bmatrix} 8 \\ -2 \end{bmatrix}, \quad \mathbf{w}_3 = \begin{bmatrix} 9 \\ -2 \end{bmatrix} \text{ and } \mathbf{w}_4 = \begin{bmatrix} 3 \\ 3 \end{bmatrix}
\end{align*}
\]
Your task is to perform an update of the network for training example
\[
\mathbf{x}_2 = \begin{bmatrix} 5 \\ 9 \end{bmatrix}.
\]
State the values of the weights after this update. (3p.)

b) Describe a cost function that are minimized by the updates in a). Assume that \(h(i, j)\) is approaching 0. (2p.)