Neural Networks (2013/14)
Example exam, December 2013

In the final exam, four problems are to be solved within 3 hours. **The use of supporting material (books, notes, calculators) is not allowed**. You can achieve up 9 points, in total. The exam grade will be "1.0 + your number of points".
Important hints: never just answer a question with "Yes" or "No", always give arguments for your conclusion. Be as precise as possible and use math where it makes sense.

---

1) Model neurons and networks

a) Consider a single neuron of the McCulloch Pitts type. Define precisely how its state of activity is determined from the neurons it is connected to (its *neighbors*). Explain why a positive weight can be interpreted as representing an *excitatory synapse*.

b) Consider a Hopfield model consisting of N McCulloch Pitts type of neurons with activities $S_j(t) \in \{-1, +1\}$ $(j = 1, 2, \ldots, N)$. Write down an update equation that specifies $S_i(t+1)$ as a function of the neural activities $S_j(t)$ in the previous time step.

c) How is a set of *patterns* $\{\boldsymbol{\xi}^\mu \in I\!R^N\}$ $(\mu = 1, 2, \ldots, P)$ stored in the simple Hopfield model? Explain in words, how it can work as an associative memory.

2) Perceptron storage problem

Consider a set of data $I\!D = \{\boldsymbol{\xi}^\mu, S^\mu\}_{\mu=1}^P$ where $\boldsymbol{\xi}^\mu \in I\!R^N$ and $S^\mu \in \{+1, -1\}$. You can assume that the data is homogeneously linearly separable.

a) Define the stability $\kappa(\mathbf{w})$ of a perceptron solution $\mathbf{w}$ with respect to the given set of data $I\!D$. Give a geometric interpretation and provide a sketch of an illustration. Explain in words why $\kappa(\mathbf{w})$ quantifies the stability of the perceptron output with respect to noise.

b) Assume you have found two different solutions $\mathbf{w}^{(1)}$ and $\mathbf{w}^{(2)}$ of the perceptron storage problem for data set $I\!D$. Assume furthermore that $\mathbf{w}^{(1)}$ can be written as a linear combination

$$\mathbf{w}^{(1)} = \sum_{\mu=1}^{P} x^\mu \, \boldsymbol{\xi}^\mu \, S^\mu \quad \text{with} \quad x^\mu \in I\!R,$$

whereas the difference vector $\mathbf{w}^{(2)} - \mathbf{w}^{(1)}$ is orthogonal to all the vectors $\boldsymbol{\xi}^\mu \in I\!D$.

Consider the stabilities of the competing solutions and prove (give precise mathematical arguments) that $\kappa(\mathbf{w}^{(1)}) \geq \kappa(\mathbf{w}^{(2)})$ holds true. What does this result imply for the perceptron of optimal stability and potential training algorithms?

3) Learning a linearly separable rule

Here we consider linearly separable data $I\!D = \{\boldsymbol{\xi}^\mu, S_R^\mu\}_{\mu=1}^P$ where noise free labels $S_R^\mu = \text{sign}[\mathbf{w}^* \cdot \boldsymbol{\xi}^\mu]$ are provided by a teacher vector $\mathbf{w}^* \in I\!R^N$ with $|\mathbf{w}^*| = 1$.

a) Define and explain the term *version space* precisely in this context, provide a mathematical definition as a set of vectors and also a simplifying graphical illustration. Give a brief argument why one can expect the perceptron of maximum stability to display good generalization behavior.

b) Define and explain the *(Rosenblatt) Perceptron* algorithm for a given set of examples $I\!D$. Be precise, for instance by writing it in a few lines of *pseudocode*. Also include a *stopping criterion*.

c) While experimenting with the Rosenblatt perceptron (with initial $\mathbf{w}(0) = 0$) in the practicals, your partner has a brilliant idea: the use of a larger learning rate. His/her argument: updating $\mathbf{w}$ by Hebbian terms of the form $\eta\, \boldsymbol{\xi}^\mu\, S^\mu$ with a large $\eta > 1$ should give (I) faster convergence and (II) a better perceptron vector. Are you convinced? Give precise arguments for yor answer!

Note: The following will be treated in January, consider it as an outlook ...

4) Learning by gradient descent

Consider a feed-forward continuous neural network with an N-dim. input layer and one very simple, <u>linear</u> unit with continuous output

$$\sigma(\boldsymbol{\xi}) = \mathbf{w} \cdot \boldsymbol{\xi} \;\in I\!R$$

Here, $\boldsymbol{\xi}$ denotes an N-dim. input vector and $\mathbf{w}$ is adaptive weight vector.

a) Given a set of training examples, i.e. inputs $\boldsymbol{\xi}^\mu$ with continuous labels $\tau^\mu \in I\!R$, consider the quadratic error measure

$$E(\mathbf{w}) = \frac{1}{2} \sum_{\mu=1}^{P} \left(\sigma(\boldsymbol{\xi}^\mu) - \tau^\mu\right)^2 .$$

as a cost function for training Derive a gradient descent learning step for the adaptive weights with respect to the cost function E.

b) What are the <u>necessary</u> conditions for a weight vector $\mathbf{w}^*$ to be a local minimum of E? You don't have to discuss sufficient conditions here. Assume some $\mathbf{w}^*$ does indeed satisfy the necessary conditions, but it is <u>not</u> a local minimum. What else could $\mathbf{w}^*$ correspond to?

c) Discuss qualitatively (in words) the role of the step size or learning rate η in the gradient descent algorithm. What can happen if η is (too) small or (too) large, respectively?