vak niet aanwezig!

Hertentamen (1) April 26, 2006

Neural Networks

The problems are to be solved within 3 hrs. **The use of supporting material (books, notes, calculators) is not allowed.** In each of the four problems you can achieve up to 2.5 points, with a total maximum of 10 points.

1. Perceptron storage problem

Consider a set of data $\mathbb{D} = (\boldsymbol{\xi}^\mu, S^\mu)_{\mu=1}^P$ where $\boldsymbol{\xi}^\mu \in \mathbb{R}^N$ and $S^\mu \in \{+1, -1\}$. In this problem, we assume that $\mathbb{D}$ is homogeneously linearly separable.

a) Formulate the perceptron storage problem as the search for a vector $\mathbf{w} \in \mathbb{R}^N$ which satisfies a set of equations. Re-write the problem using a set of inequalities.

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

c) Assume we have found two different solutions $\mathbf{w}^{(1)}$ and $\mathbf{w}^{(2)}$ of the perceptron storage problem for $\mathbb{D}$. Assume furthermore that $\mathbf{w}^{(1)}$ can be written as a linear combination

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

whereas the difference $(\mathbf{w}^{(2)} - \mathbf{w}^{(1)})$ is orthogonal to all the $\boldsymbol{\xi}^\mu$ in $\mathbb{D}$, i.e. $(\mathbf{w}^{(2)} - \mathbf{w}^{(1)}) \cdot \boldsymbol{\xi}^\mu = 0$ for $\mu = 1, 2, \dots P$.

Show that $\kappa(\mathbf{w}^{(1)}) > \kappa(\mathbf{w}^{(2)})$. What does the result imply for the perceptron of optimal stability $\mathbf{w}_{max}$?

2. Learning a linearly separable rule

Here we consider perceptron training from linearly separable data $\mathbb{D} = \{\boldsymbol{\xi}^\mu, S_R^\mu\}_{\mu=1}^P$ where noise-free labels $S_R^\mu = \text{sign}[\mathbf{w}^* \cdot \boldsymbol{\xi}^\mu]$ are provided by a teacher vector $\mathbf{w}^* \in \mathbb{R}^N$ with $|\mathbf{w}^*| = 1$. Assume that by some training process we have obtained a perceptron vector $\mathbf{w} \in \mathbb{R}^N$ from the data $\mathbb{D}$.

a) Define the terms *training error* and *generalization error* in the context of this situation.

b) Assume that random input vectors $\boldsymbol{\xi} \in \mathbb{R}^N$ are generated with equal probability anywhere on the *hypersphere* with squared radius $\boldsymbol{\xi}^2 = 1$. Given $\mathbf{w}^*$ and a vector $\mathbf{w} \in \mathbb{R}^N$, what is the probability for *disagreement*, $\text{sign}[\mathbf{w} \cdot \boldsymbol{\xi}] \neq \text{sign}[\mathbf{w}^* \cdot \boldsymbol{\xi}]$? You can "derive" the result from a sketch of the situation in $N = 2$ dimensions.

c) Explain Rosenblatt's perceptron algorithm for a given set of examples $\mathbb{D}$ in terms of a few lines of *pseudocode*.

3. Classification with multilayer networks

a) Explain the so-called *committee machine* with inputs $\boldsymbol{\xi} \in I\!R^N$, K hidden units $\left(\{\sigma_k = \pm 1\}_{k=1}^K\right)$, and corresponding weight vectors $\mathbf{w}_k \in I\!R^N$. Define the output $S(\boldsymbol{\xi}) \in \{-1, +1\}$ as a function of the input.

b) Now consider the so-called *parity machine* with N-dim. input and K hidden units. Define the output $S(\boldsymbol{\xi}) \in \{-1, +1\}$ as a function of the input.

c) Illustrate the case $K = 3$ for *parity* and *committee machine* in terms of a geometric interpretation. Why would you expect that the parity machine should have a greater *storage capacity* in terms of implementing random sets $I\!D = \{\boldsymbol{\xi}^\mu, S(\boldsymbol{\xi}^\mu)\}$?

4. Regression with neural networks

a) Explain the term *overfitting* in the context of regression problem. Use - as an example - regression with a continuous neural network with K hidden units and discuss the role of K. What is the meaning of *bias* and *variance* in this context?

(Bonus) b) Explain the method of *n-fold cross validation* and how it could be used for choosing an appropriate student complexity. You may discuss it in terms of the same example as in (a). Does cross validation provide good estimates of bias and variance?

c) Describe and explain *weight decay* as a method of *regularization*. Why does weight decay *smoothen* the network output? Explain also *early stopping* and why it can give similar results as *weight decay* .

4b) will be treated as a bonus problem (max. ½ extra point)