Schrijfopdracht 4

Semester 1 2014/15

Deadline: 26 November 2014
MAG IN TEAMS VAN 2 GEMAAKT WORDEN

1. Consider Bias-Variance decomposition (1 punt)
   (a) Define the error of an estimator and its decomposition into bias and variance.
   (b) Given a dataset and a neural network implementation, can we estimate the bias and variance components of the error? If not, why not? If yes, how?
   (c) What is the effect of increasing the number of nodes on the hidden layer on training-set and test-set error and on bias and variance?
   (d) Normally, modifications to the learning algorithm that reduce the bias, increase the variance and vice versa. Is it possible that a change to neural network learning reduces both variance and bias? If not, why not? If yes, give an example.

2. (1 punt)

   Beschouw datasets:

x1	1	1	2	3	4	4	4	7	8	8	8
x2	3	6	6	5	1	3	6	7	6	7	3
y	0	0	1	1	0	0	1	1	1	0	0

   (a) teken deze in 2 D en teken daarin de class boundaries die je vindt met: (1) beslisbomen, (2) 1-nearest neighbour (3) logistische regressie en (4) logistische regressie met kwadratische termen toegevoegd. (Mooi als je dit plot of met Latex maakt, maar getekende grafieken, gescand en in de file gevoegd is ook OK.)
   (b) (Moeiilijke vraag) Welke boundary lijkt je het beste? Waarom?

3. Cost functions (1 punt)
   (a) An interesting problem is to evaluate the *predictive power* of a single variable for classification. This can be useful for several reasons. We may want to discard this variable, to reduce the complexity of the learning problem. In decision tree learning we use it to select the next variable for extending the tree. We have seen to measures for this: the cost of a single variable logistic regression hypothesis, information gain and accuracy. Suppose that we want to predict y from x and our data are the following:

x	1	3	7	10	11	15
y	0	0	1	1	0	1

   (b) Calculate the *information gain*, the *cost* using the cost function of Logistic Regression and the *accuracy* of setting the boundary at $x = 8.5$.
   (c) Is this the best possible class boundary for these measures? Explain your answer.