Name and Student ID: Signature:

Machine Learning BLG527E, Jan 10, 2013, Final Exam.

1	2	3	4	Total

Duration: 120 minutes.
Write your answers neatly in the space provided for them. Write your name on each sheet. Books and notes are closed. Good Luck!

QUESTIONS

QUESTION1) [30 points, 5 points each] (use at most 3 sentences per question, use of formulas, drawings etc. to better express yourself is encouraged.)

- **a)** Compare Adaboost and bagging.
- **b)** What is ROC and AUC?
- **c)** How do you use momentum and adaptive learning rate for training a multilayer perceptron?
- **d)** What is the difference between k-means clustering and Gaussian Mixture Model (GMM) clustering?
- **e)** Compare backward and forward feature selection.
- **f)** Compare Mahalabonis and Euclidean distance.

QUESTION2) [40 points, 10 points each] (use at most 10 sentences per question, use of formulas, drawings etc. to better express yourself is encouraged.)

- **a)** Assume that you have K classifiers which are independent and ith classifier has generalization error of E_i. Show that the average of these classifiers has less generalization error. Why, in practice, classifier averaging doesn't always result in less generalization error?
- **b)** What are the three basic problems that you can solve using an hmm? Explain each of problem and how it is solved briefly.
- **c)** How does 5x2 cross validation work? Under which circumstances do you use leave one out, 10-fold or 5x2 cross validation?
- **d)** How do you use Parzen windows for density estimation, classification and regression?

QUESTION3) [20points]

	Actual	$g_1(x)$	$g_2(x)$	$g_3(x)$	$g_4(x)$	$g_5(x)$	$f_1(x)$	$f_2(x)$	$f_3(x)$	$f_4(x)$	$f_5(x)$
Train	0	0	0	0	0	0	0	0	0	1	1
	1	0	1	0	0	1	1	1	1	0	0
	0	0	0	1	1	0	0	1	1	0	0
	1	0	1	0	1	0	1	0	1	0	1
Test	?	1	1	1	1	1	1	0	1	0	1
	?	0	0	0	0	0	0	1	0	1	0
	?	0	1	0	1	0	1	0	1	0	1
	?	1	1	1	1	1	0	1	0	1	0

You are given a dataset which contains 4 training inputs and output, 4 test inputs. You have two different types of classifiers f() and g(), you train and test 5 different configurations of f() and g() (for example you might choose different parameters of the classifiers). The training and test outputs that are obtained for each input are shown in the table. Would you choose a classifier of kind f() or g() for this task? Clearly explain the reason behind your choice?
[**Hint:** if M samples (data sets) $X_i=\{x^t_i, r^t_i\}$, $i=1,...,M$ are used to fit $g_i(x)$, $i=1,...,M$

$$\text{Bias}^2(g) = \frac{1}{N}\sum_t \left[\bar{g}(x^t) - f(x^t)\right]^2$$

$$\text{Variance}(g) = \frac{1}{NM}\sum_t \sum_i \left[g_i(x^t) - \bar{g}(x^t)\right]^2$$

$$\bar{g}(x) = \frac{1}{M}\sum_i g_i(x)$$

]

QUESTION4) [10points]

Compute the change in v and w (i.e. Δv and Δw) if you modified the error function for a neural network as follows.

$$E(\mathbf{W}, \mathbf{v}|\mathcal{X}) = \frac{1}{2}\sum_t \left(r^t - y^t\right)^2 + \| v \|^2$$

Hint: Without the regularization term, you would have:

$$y_i = \mathbf{v}_i^T \mathbf{z} = \sum_{h=1}^{H} v_{ih} z_h + v_{i0} \quad z_h = \text{sigmoid}\left(\mathbf{w}_h^T \mathbf{x}\right)$$

$$\Delta v_h = \sum_t \left(r^t - y^t\right) z_h^t$$

$$\begin{aligned}\Delta w_{hj} &= -\eta \frac{\partial E}{\partial w_{hj}} = -\eta \sum_t \frac{\partial E}{\partial y^t}\frac{\partial y^t}{\partial z_h^t}\frac{\partial z_h^t}{\partial w_{hj}} \\ &= -\eta \sum_t -\left(r^t - y^t\right) v_h z_h^t \left(1 - z_h^t\right) x_j^t \\ &= \eta \sum_t \left(r^t - y^t\right) v_h z_h^t \left(1 - z_h^t\right) x_j^t\end{aligned}$$